Multilinear Maximum Distance Embedding via L_1-norm Optimization

Yang Liu, Yan Liu and Keith C.C. Chan
Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
{csygliu, csyliu, cskcchan}@comp.polyu.edu.hk

Abstract

Dimensionality reduction plays an important role in many machine learning and pattern recognition tasks. In this paper, we present a novel dimensionality reduction algorithm called multilinear maximum distance embedding (M^{2}DE), which includes three key components. To preserve the local geometry and discriminant information in the embedded space, M^{2}DE utilizes a new objective function, which aims to maximize the distances between some particular pairs of data points, such as the distances between nearby points and the distances between data points from different classes. To make the mapping of new data points straightforward, and more importantly, to keep the natural tensor structure of high-order data, M^{2}DE integrates multilinear techniques to learn the transformation matrices sequentially. To provide reasonable and stable embedding results, M^{2}DE employs the L_1-norm, which is more robust to outliers, to measure the dissimilarity between data points. Experiments on various datasets demonstrate that M^{2}DE achieves good embedding results of high-order data for classification tasks.

Introduction

Dimensionality reduction (DR) is one of the vital problems in machine learning and pattern recognition. Traditional DR techniques, such as principal component analysis (PCA) (Hotelling 1933) and linear discriminant analysis (LDA) (Fisher 1936), seek the linear transformation matrix to map high-dimensional data into low-dimensional feature space. However, if the original data hold the nonlinear structure, linear methods may ignore the subtleties of the data distribution. Manifold learning, a kind of nonlinear DR techniques based on the assumption that the high-dimensional input data lie on or close to an intrinsically smooth low-dimensional manifold, received more and more attention recently. The representative manifold learning algorithms include isometric feature mapping (Isomap) (Tenenbaum, de Silva, & Langford 2000), locally linear embedding (LLE) (Roweis & Saul 2000), and Laplacian eigenmaps (LE) (Belkin & Niyogi 2001). Isomap is a global manifold learning method that aims to preserve the geometry at all scales by mapping nearby points on the manifold to nearby points in low-dimensional space, and faraway points to faraway points. In contrast with Isomap, LLE and LE assume that the global nonlinear structure can be uncovered by keeping all local structures of the dataset, and thus only attempt to preserve the local geometry. Following above algorithms, many manifold learning techniques have been developed, such as stochastic neighbor embedding (SNE) (Hinton & Roweis 2002), locally linear coordination (LLC) (Teh & Roweis 2002), semidefinite embedding (SDE) (Weinberger & Saul 2004), and maximum variance unfolding (MVU) (Weinberger & Saul 2006).

In this paper, we propose a novel manifold learning algorithm called multilinear maximum distance embedding (M^{2}DE). Unlike most of the manifold learning techniques that attempt to *preserve* the distances or relationships between data points, M^{2}DE uses a new objective function to *maximize* the distances between some particular pairs of data points. By maximizing the distances between nearby data points, the local nonlinear structure of the dataset can be flattened in the embedded space. By maximizing the distance between data points from different classes, the separability is well preserved after embedding. Unlike traditional methods that first unfold the input data to vectors even though the data are high-order tensors, M^{2}DE directly works on tensor space and learns a series of transformation matrices using an iterative strategy. With the explicit function, the mapping of new data point becomes straightforward. Unlike existing manifold learning algorithms which measure the dissimilarity between data points using Frobenius norm (F-norm, also known as L_2-norm in the vector form operation), M^{2}DE is formulated by L_1-norm based optimization. As known, F-norm is more sensitive to outliers than L_1-norm because the large squared errors dominate the sum. Some recent work on DR also demonstrated that L_1-norm based PCA can achieve better embedding results than the conventional F-norm based PCA (Huang & Ding 2008; Kwak 2008; Pang, Li, & Yuan 2010). In summary, the proposed algorithm has the following attractive characters:

1) By introducing a new objective function, M^{2}DE not only keeps the nonlinear structure of the dataset but also maximizes the separability for classification task.

2) By integrating the multilinear techniques, M^{2}DE overcomes the out-of-sample problem (Bengio *et al.* 2003). More importantly, if the data are high-order tensors, the intrinsic structure of data can be well preserved.

3) By utilizing the L_1-norm to measure the dissimilarity between data points, M^{2}DE is robust to outliers, and hence, shows more reasonable and stable embedding results.

Maximum Distance Embedding

DR discovers the compact representation of original high-dimensional observations. Mathematically, DR can be stated as follows: Given n data points $\mathbf{x}_1, ..., \mathbf{x}_n$ in the high-dimensional space $\mathbb{R}^D$, find their low-dimensional representations $\mathbf{y}_1, ..., \mathbf{y}_n \in \mathbb{R}^d$ with $d \le D$, such that the essentials in original data can be captured according to some criteria.

The method proposed in this paper intends to capture both the manifold structure of the dataset and the discriminant information for classification task by maximizing the distances between nearby data points and the distances between data points from different classes simultaneously.

Figure 1 illustrates the idea behind the proposed maximum distance embedding (MDE). Figure 1(a) is the original 2-D data from three classes. The data points within the same class are equally distributed on the manifold. Figure 1(b) shows the 1-D embedding that only preserves the local geometry. Although the manifold structure within each class is successfully described, some data points from class 1 and class 2 are inseparable because the discriminant information is ignored in the embedding process. Figure 1(c) shows the 1-D embedding that only maximizes the discriminant information. Obviously, the local geometry of dataset is seriously distorted, i.e., the embedded data points within the same class are not equally distributed any more. Figure 1(d) is the 1-D result of MDE. By maximizing the distances between nearby data points, the local geometry is preserved after embedding. Moreover, by maximizing the distances between data points from different classes, the discriminant information is well kept in the subspace.

Based on above consideration, we define the objective function for the proposed algorithm as follows:

$$\max J(\mathbf{y}_1,...,\mathbf{y}_n) = \sum_{i,j} (w_{ij}^l + w_{ij}^d) d(\mathbf{y}_i, \mathbf{y}_j) \quad (1)$$

where $d(\mathbf{y}_i,\mathbf{y}_j)$ is the distance metric to measure the dissimilarity between embedded data points $\mathbf{y}_i$ and $\mathbf{y}_j$. w_{ij}^l and w_{ij}^d are two weighting parameters. To emphasize the local details between data points $\mathbf{x}_i$ and $\mathbf{x}_j$, we define w_{ij}^l as follows:

$$w_{ij}^l = \begin{cases} \exp(-d(\mathbf{x}_i,\mathbf{x}_j)^2/\sigma_1) & \text{if } \mathbf{x}_j \in O(\mathbf{x}_i;k) \text{ or } \mathbf{x}_i \in O(\mathbf{x}_j;k) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $O(\mathbf{x}_i;k)$ denotes the set of k nearest neighbors of $\mathbf{x}_i$ and σ_1 is a positive parameter. Clearly, by maximizing the distances between nearby points, the local nonlinear structure of the dataset can be flattened to the greatest extent and well displayed in the embedded low-dimensional space. Inheriting the assumption of local manifold learning techniques, MDE can uncover the global nonlinear structure of the dataset by keeping all local geometries. Furthermore, we define w_{ij}^d to describe the discriminant information:

$$w_{ij}^d = \begin{cases} \sigma_2 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to different classes} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where σ_2 is a positive parameter. By maximizing the distance between data points from different classes, the separability is well preserved in the embedded space.

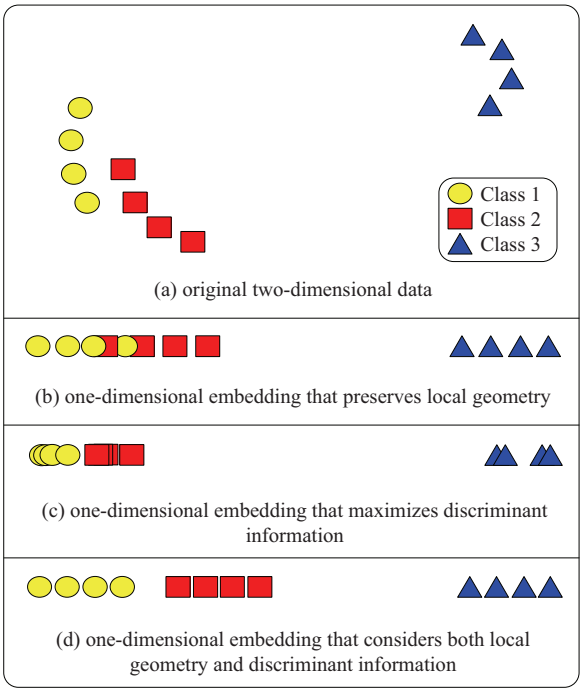


Figure 1: Schematic illustration of the main idea behind MDE. (a) original 2-D data. (b) 1-D embedding that preserves the local geometry. (c) 1-D embedding that maximizes the discriminant information. (d) 1-D embedding by MDE, which considers both local geometry and discriminant information.

Multilinear Maximum Distance Embedding

In this section, we present the multilinear formulation of proposed method. By integrating multilinear algebra into MDE, the out-of-sample problem (Bengio *et al.* 2003) and vectorization problem (Vasilescu & Terzopoulos 2003) can be effectively addressed. As known, the out-of-sample problem exists in most of the manifold learning algorithms, i.e., it is not possible to embed new data points without reconstructing the whole low-dimensional space. Furthermore, traditional manifold learning algorithms usually unfold input data to vectors before embedding, even though the data are naturally high-order tensors. This kind of vectorization increases the computational cost of data analysis and destroys the intrinsic structure of high-order data.

To tackle both out-of-sample and vectorization problems, multilinear algebra (Lathauwer 1997; Vasilescu & Terzopoulos 2003) has been introduced into DR, and then some multilinear based manifold learning techniques have been proposed (He, Cai, & Niyogi 2005; Dai & Yeung 2006; Liu, Liu, & Chan 2009). Inspired by previous work, we propose the multilinear maximum distance embedding (M^2DE) algorithm. First we give the following definition from multilinear algebra.

Definition 1: (mode-k product). The mode-k product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_k \times I_k}$, denoted

by $\mathcal{X} \times_k \mathbf{U}$, is an $(I_1 \times ... \times I_{k-1} \times J_k \times I_{k+1} \times ... \times I_N)$-tensor of which the entries are given by

$$(\mathcal{X} \times_k \mathbf{U})_{i_1 i_2 \cdots i_{k-1} j_k i_{k+1} \cdots i_N} = \sum_{i_k=1}^{I_k} \mathcal{X}_{i_1 i_2 \cdots i_{k-1} i_k i_{k+1} \cdots i_N} \mathbf{U}_{j_k i_k}, j_k = 1, ..., J_k .$$

In general, the goal of multilinear DR can be described as follows. Given n data points $\mathcal{X}_1$, ..., $\mathcal{X}_n$ in the tensor space $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$. Without unfolding the input data points to $I_1 \times I_2 \times ... \times I_N$-dimensional vectors, multilinear embedding methods seek to find N transformation matrices $\mathbf{V}_k = [\mathbf{v}_k^1, ..., \mathbf{v}_k^{I_k'}] \in \mathbb{R}^{I_k \times I_k'}$ ($I_k' \ll I_k$, $k = 1, ..., N$) such that n low-dimensional data points $\mathcal{Y}_1$, ..., $\mathcal{Y}_n$ in the subspace $\mathbb{R}^{I_1' \times I_2' \times \cdots \times I_N'}$ can be calculated by the multilinear transformation $\mathcal{Y}_j = \mathcal{X}_j \times_1 \mathbf{V}_1^T \times_2 \cdots \times_N \mathbf{V}_N^T$ ($j = 1, ..., n$).

Based on above definitions from multilinear algebra, we can formulate the objective function of M^{2}DE as follows:

$$\arg\max_{\mathbf{V}_k | _{k=1}^N} J(\mathbf{V}_1, ..., \mathbf{V}_N) = \sum_{i,j} (w_{ij}^l + w_{ij}^d) d(\mathcal{Y}_i, \mathcal{Y}_j) \quad (4)$$

where $\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{V}_1^T \times_2 \cdots \times_N \mathbf{V}_N^T$ ($i = 1, ..., n$).

L_1-norm Optimization

Generally, $d(\bullet,\bullet)$ in Eq. (1) and (4) can be any distance metric. Most of the manifold learning algorithms try to optimize the objective functions based on different least-squares formulations, which are expressed by the F-norm. However, it is known that the F-norm is sensitive to outliers since the large squared errors dominate the sum (Huang & Ding 2008; Kwak 2008). In this paper, we utilize L_1-norm in the objective function. Compared with F-norm, L_1-norm is more robust to outliers. Some recent work on DR has already demonstrated that L_1-norm based methods can effectively reduce the negative influence of outliers and hence, achieve better embedding results (Huang & Ding 2008; Kwak 2008; Pang, Li, & Yuan 2010).

By embedding original data to the low-dimensional tensor subspace, we expect to obtain a meaningful representation of original data with less sensitivity to the outliers. By employing the L_1-norm, we can rewrite Eq. (4) as follows:

$$\arg\max_{\mathbf{V}_k | _{k=1}^N} J(\mathbf{V}_1, ..., \mathbf{V}_N) = \sum_{i,j} (w_{ij}^l + w_{ij}^d) \| (\mathcal{X}_i - \mathcal{X}_j) \times_1 \mathbf{V}_1^T \cdots \times_N \mathbf{V}_N^T \|_1 \quad (5)$$
$$s.t.\ \mathbf{V}_k^T \mathbf{V}_k = \mathbf{I}_{J_k'},\ k = 1, ..., N$$

The constraints in Eq. (5) are to ensure the orthonormality of the transformation matrices.

When $N \geq 2$, it is difficult to find a global solution for such a high-order optimization problem. Instead, we use an iterative strategy to obtain a local solution. To introduce the iterative strategy, we will make use of the following definition and properties.

Definition 2: (mode-k unfolding). The mode-k unfolding of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ ($N \geq 3$) into a matrix $\mathbf{X}^k \in \mathbb{R}^{I_k \times \prod_{j \neq k} I_j}$, i.e., $\mathbf{X}^k \Leftarrow_k \mathcal{X}$, is defined as: $\mathbf{X}^k_{i_k, j} = \mathcal{X}_{i_1 i_2 ... i_N}$, $j = \sum_{m=2}^{N-1} (i_{p(m)} - 1) \prod_{o=m+1}^{N} I_{p(o)} + i_{p(N)}$, where $p(m)$ is the m^{th} element of the sequence $\{k, k+1, \ldots, N-1, N, 1, 2, \ldots, k-1\}$.

Property 1: Given a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and the matrices $\mathbf{U} \in \mathbb{R}^{J_k \times I_k}$, $\mathbf{V} \in \mathbb{R}^{J_l \times I_l}$ ($k \neq l$), then

$$(\mathcal{X} \times_k \mathbf{U}) \times_l \mathbf{V} = (\mathcal{X} \times_l \mathbf{V}) \times_k \mathbf{U} = \mathcal{X} \times_k \mathbf{U} \times_l \mathbf{V}.$$

Property 2: If $\mathbf{X}^k \Leftarrow_k \mathcal{X}$, then $\| \mathcal{X} \times_k \mathbf{U}_k^T \|_F = \| \mathbf{U}_k^T \mathbf{X}^k \|_F$.

Assume that $\mathbf{V}_1$, …, $\mathbf{V}_{k-1}$, $\mathbf{V}_{k+1}$, …, $\mathbf{V}_N$ are fixed, we can obtain $\mathbf{V}_k$ by a greedy algorithm. First we compute $\mathbf{v}_k^1$, i.e., the first column of matrix $\mathbf{V}_k$. Eq. (5) now becomes:

$$\begin{aligned} \arg\max_{\mathbf{v}_k^1} J(\mathbf{v}_k^1) &= \sum_{i,j} (w_{ij}^l + w_{ij}^d) \| (\mathbf{v}_k^1)^T \mathbf{X}_{ij}^k \|_1 \\ &= \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} | (w_{ij}^l + w_{ij}^d) (\mathbf{v}_k^1)^T (\mathbf{x}_{ij}^k)^m | \quad (6) \\ s.t.\ (\mathbf{v}_k^1)^T \mathbf{v}_k^1 &= 1 \end{aligned}$$

where $\mathbf{X}_{ij}^k = [(\mathbf{x}_{ij}^k)^1, ..., (\mathbf{x}_{ij}^k)^{\prod_{o \neq k} I_o}] \in \mathbb{R}^{I_k \times \prod_{o \neq k} I_o}$ is the mode-k unfolding of the tensor $\mathcal{X}_{ij}^k$, i.e., $\mathbf{X}_{ij}^k \Leftarrow_k \mathcal{X}_{ij}^k$, and $\mathcal{X}_{ij}^k = (\mathcal{X}_i - \mathcal{X}_j) \times_1 \mathbf{V}_1^T \cdots \times_{k-1} \mathbf{V}_{k-1}^T \times_{k+1} \mathbf{V}_{k+1}^T \cdots \times_N \mathbf{V}_N^T$. Here $(w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^1)^T (\mathbf{x}_{ij}^k)^m$ is a scalar, and $|\bullet|$ denotes the absolute value operation. The second equality holds since $w_{ij}^l, w_{ij}^d \geq 0$ for any i and j.

We use $\mathbf{v}_k^1(t)$ to denote the value of $\mathbf{v}_k^1$ after the t^{th} iteration. Then $\mathbf{v}_k^1(t+1)$ can be computed as follows:

$$\mathbf{v}_k^1(t+1) = \frac{\sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{x}_{ij}^k)^m}{\| \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{x}_{ij}^k)^m \|} \quad (7)$$

where $\|\bullet\|$ denotes the F-norm, and $p_{ij}^m(t)$ is the polarity function (Kwak 2008; Pang, Li, & Yuan 2010) defined as:

$$p_{ij}^m(t) = \begin{cases} 1 & \text{if } (w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^1(t))^T (\mathbf{x}_{ij}^k)^m \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad (8)$$

To prove the convergence of above iteration procedure, we only need to prove $J(\mathbf{v}_k^1(t+1)) \geq J(\mathbf{v}_k^1(t))$. First, we have:

$$\begin{aligned} J(\mathbf{v}_k^1(t+1)) &= \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t+1) (w_{ij}^l + w_{ij}^d) (\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m \\ &\geq \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m \end{aligned}$$

The inequality results from the fact that $p_{ij}^m(t+1)$ is the optimal polarity corresponding to $(w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m$, i.e., $p_{ij}^m(t+1)(w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m \geq 0$ for any i, j, and m. But for $p_{ij}^m(t)$, $p_{ij}^m(t)(w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m < 0$ may happen.

Moreover, let $\mathbf{q}(t) = \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{x}_{ij}^k)^m$, then:

$$\begin{aligned} J(\mathbf{v}_k^1(t+1)) &\geq \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{v}_k^1(t+1))^T (\mathbf{x}_{ij}^k)^m \\ &= (\mathbf{v}_k^1(t+1))^T \mathbf{q}(t) = \left[\frac{\mathbf{q}(t)}{\| \mathbf{q}(t) \|} \right]^T \mathbf{q}(t) = \| \mathbf{q}(t) \| \\ &\geq \| \mathbf{q}(t) \| \frac{[\mathbf{q}(t)]^T \mathbf{q}(t-1)}{\| \mathbf{q}(t) \| \times \| \mathbf{q}(t-1) \|} = [\mathbf{q}(t)]^T \frac{\mathbf{q}(t-1)}{\| \mathbf{q}(t-1) \|} \\ &= \left[\sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} p_{ij}^m(t) (w_{ij}^l + w_{ij}^d) (\mathbf{x}_{ij}^k)^m \right]^T (\mathbf{v}_k^1(t)) \\ &= \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{\prod_{o \neq k} I_o} | (w_{ij}^l + w_{ij}^d) (\mathbf{v}_k^1(t))^T (\mathbf{x}_{ij}^k)^m | \\ &= J(\mathbf{v}_k^1(t)) \end{aligned}$$

The second inequality results from the fact that $[\mathbf{q}(t)]^T\mathbf{q}(t\text{-}1) \le \|\mathbf{q}(t)\|\times\|\mathbf{q}(t\text{-}1)\|$, which is known as the Cauchy-Schwarz inequality. Therefore, the iteration procedure will finally converge and thus we can obtain a local optimal solution of $\mathbf{v}_k^1$ by updating it using Eq. (7) until $\mathbf{v}_k^1(t+1) = \mathbf{v}_k^1(t)$.

Based on the obtained $\mathbf{v}_k^1$, we can compute the remaining vectors $\mathbf{v}_k^2, ..., \mathbf{v}_k^{I'_k}$ of matrix $\mathbf{V}_k$ by a greedy method. First, we initialize the data matrix $(\mathbf{X}_{ij}^k)^1 = \mathbf{X}_{ij}^k$ for $i, j = 1, ..., n$. Then we update it as follows:

$$(\mathbf{X}_{ij}^k)^{r+1} = (\mathbf{X}_{ij}^k)^r - \mathbf{v}_k^r((\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r) \quad r = 1,...,I'_k - 1 \tag{9}$$

Finally, we iteratively calculate $\mathbf{v}_k^{r+1}$ by the following Eq. (10) and Eq. (11) until the result converges.

$$\mathbf{v}_k^{r+1}(t+1) = \frac{\sum_{i=1}^n\sum_{j=1}^n\sum_{m=1}^{\prod_{o\neq k}I_o} p_{ij}^m(t)(w_{ij}^l + w_{ij}^d)((\mathbf{x}_{ij}^k)^{r+1})^m}{\|\sum_{i=1}^n\sum_{j=1}^n\sum_{m=1}^{\prod_{o\neq k}I_o} p_{ij}^m(t)(w_{ij}^l + w_{ij}^d)((\mathbf{x}_{ij}^k)^{r+1})^m\|} \tag{10}$$

$$p_{ij}^m(t) = \begin{cases} 1 & \text{if } (w_{ij}^l + w_{ij}^d)(\mathbf{v}_k^{r+1}(t))^T((\mathbf{x}_{ij}^k)^{r+1})^m \ge 0 \\ -1 & \text{otherwise} \end{cases} \tag{11}$$

By employing above procedure, the orthonormality of $\mathbf{V}_k$ is guaranteed: From Eq. (10), we know that $\mathbf{v}_k^{r+1}$ is a linear combination of $((\mathbf{x}_{ij}^k)^{r+1})^m$, i.e., a linear combination of the columns from $(\mathbf{X}_{ij}^k)^{r+1}$. To prove that $\mathbf{v}_k^{r+1}$ and $\mathbf{v}_k^r$ are perpendicular, i.e., $(\mathbf{v}_k^r)^T\mathbf{v}_k^{r+1} = 0$, we only need to show that $(\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^{r+1} = \mathbf{0}^T$, where $\mathbf{0}^T$ is the zero vector with the length $\prod_{o=1,o\neq k}^N I_o$. Consider Eq. (9), we have the following:

$$\begin{aligned}(\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^{r+1} &= (\mathbf{v}_k^r)^T((\mathbf{X}_{ij}^k)^r - \mathbf{v}_k^r((\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r)) \\ &= (\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r - (\mathbf{v}_k^r)^T\mathbf{v}_k^r((\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r) \\ &= (\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r - (\mathbf{v}_k^r)^T(\mathbf{X}_{ij}^k)^r = \mathbf{0}^T\end{aligned}$$

The third equality results from the property that $(\mathbf{v}_k^r)^T\mathbf{v}_k^r = 1$, i.e., $\mathbf{v}_k^r$ is a unit vector, which can be observed from Eq. (10). Actually, Eq. (9) can be viewed as a Gram-Schmidt process, which is used to eliminate the relevance between different columns of the transformation matrix $\mathbf{V}_k$.

Till now, we have already shown how to obtain the transformation matrix $\mathbf{V}_k$ when $\mathbf{V}_1$, …, $\mathbf{V}_{k-1}$, $\mathbf{V}_{k+1}$, …, $\mathbf{V}_N$ are fixed. The iterative strategy can then be presented. First we fix $\mathbf{V}_2$, …, $\mathbf{V}_N$, and obtain $\mathbf{V}_1$. Then we fix $\mathbf{V}_1$, $\mathbf{V}_3$, …, $\mathbf{V}_N$, and obtain $\mathbf{V}_2$. The rest can be deduced by analogy. At last we fix $\mathbf{V}_1$, $\mathbf{V}_2$, …, $\mathbf{V}_{N-1}$, and obtain $\mathbf{V}_N$. Repeat above steps until the whole algorithm converges. Algorithm 1 describes the detailed procedure of M^{2}DE.

To analyze the computational cost of M^{2}DE, we simply assume that the sample tensors and embedded tensors are of uniform size in each order, respectively, i.e., $I_1 = \cdots = I_N = I$ and $I'_1 = \cdots = I'_N = I'$. In the training process, the time cost of M^{2}DE is $O(n^2NI^NI') \times T_{max1} \times T_{max2}$. Generally, the algorithm will converge within a few iterations. To embed a new data point $\mathcal{X}$, we use the transformation $\mathcal{Y} = \mathcal{X}\times_1\mathbf{V}_1^T\cdots\times_N\mathbf{V}_N^T$. So the test time cost is $O(\sum_{i=1}^N (I')^iI^{N+1-i})$. The space needed to store transformation matrices is $(NI'I)$.

Algorithm 1 Multilinear Maximum Distance Embedding

Input: Training data $\mathcal{X}_1, ..., \mathcal{X}_n \in \mathbb{R}^{I_1\times\cdots\times I_N}$;
Embedded low dimensions $I'_1, \cdots, I'_N$;
Parameters σ_1, σ_2; Iteration numbers T_{max1}, T_{max2}.
Output: Transformation matrices $\mathbf{V}_k = \mathbf{V}_k^{T_1} \in \mathbb{R}^{I_k\times I'_k} (k = 1,...,N)$
initialize $\mathbf{V}_k^0$ as arbitrary columnly orthogonal matrices;
for t_1 = 1, ..., T_{max1} **do**
  for k = 1, ..., N **do**
    $\mathcal{X}_{ij}^k = (\mathcal{X}_i - \mathcal{X}_j)\times_1\mathbf{V}_1^T\cdots\times_{k-1}\mathbf{V}_{k-1}^T\times_{k+1}\mathbf{V}_{k+1}^T\cdots\times_N\mathbf{V}_N^T$;
    $\mathbf{X}_{ij}^k \Leftarrow_k \mathcal{X}_{ij}^k$;
    for t_2 = 1, ..., T_{max2} **do**
      compute $\mathbf{v}_k^1(t_2)$ using Eq. (7) and Eq. (8);
    end for
    initialize $(\mathbf{X}_{ij}^k)^1 = \mathbf{X}_{ij}^k$;
    for r = 1, ..., $I'_k - 1$ **do**
      compute $(\mathbf{X}_{ij}^k)^{r+1}$ using Eq. (9);
      for t = 1, ..., T_{max2} **do**
        compute $\mathbf{v}_k^{r+1}(t_2)$ using Eq. (10) and Eq. (11);
      end for
    end for
  end for
end for

Experiments

In this section, we evaluate the proposed method using pattern recognition tasks on USPS digit database (Hull 1994) and Honda/UCSD video database (Lee *et al.* 2005). Images in USPS database are second-order tensors, and videos in Honda/UCSD database are third-order tensors.

The recognition process composes of three steps. First, the subspace is calculated from the training dataset. Second, for the image database, the test images are embedded into d-dimensional subspace (vector-based methods) or ($d\times d$)-dimensional tensor subspace (tensor based methods); for video database, the test data are embedded into ($d_1\times d_2\times d_3$)-dimensional tensor subspace. Finally, the k nearest neighbor algorithm is applied to low-dimensional subspace for classification. In all experiments, we empirically set T_{max1} = 10, T_{max2} = 5, and $\sigma_1 = \sigma_2 = 5$. For F-norm based multilinear DR algorithms, we set iteration number T_{max} = 10.

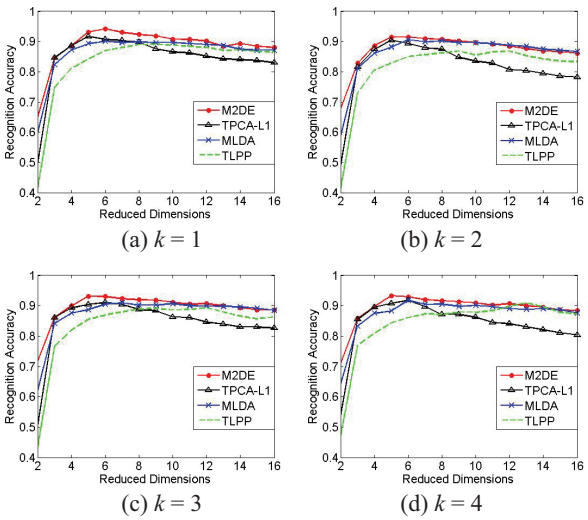
USPS Digit Database

The United State Postal Service (USPS) database (Hull 1994) of hand written digital characters contains 11000 normalized grayscale images of size 16×16, with 1100 images for each of the ten classes: from 0 to 9.

In this database, we conduct three experiments. First we compare M^{2}DE with other twelve typical DR algorithms: PCA, multilinear PCA (MPCA) (Lu, Plataniotis, & Venetsanopoulos 2008), L_1-norm PCA (PCA-L1) (Kwak 2008),

Table 1: Comparison of recognition accuracy (%) as well as corresponding optimal reduced dimensions on USPS database

Methods	M^{2}DE	PCA-L1	TPCA-L1	PCA	MPCA	LDA	MLDA	LPP	TLPP	NPE	TNPE	IsoPro	MIE
Recog.	**93.3**	90.5	91.8	82.9	87.4	89.1	91.8	85.2	91.0	87.6	91.2	88.3	91.5
Dims	5^2	54	6^2	29	12^2	20	6^2	38	13^2	22	6^2	27	8^2



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

Figure 2: Recognition accuracy of M^{2}DE, TPCA-L1, MLDA, and TLPP on USPS database with different neighborhood sizes

L_1-norm tensor PCA (TPCA-L1) (Pang, Li, & Yuan 2010), LDA, multilinear LDA (MLDA) (Yan *et al.* 2007), locality preserving projections (LPP) (He & Niyogi 2003), tensor LPP (TLPP) (He, Cai, & Niyogi 2005), neighborhood preserving embedding (NPE) (He *et al.* 2005), tensor NPE (TNPE) (Dai & Yeung 2006), IsoProjection (Cai, He, & Han 2007), and multilinear isometric embedding (MIE) (Liu, Liu, & Chan 2009). Here LPP and TLPP; NPE and TNPE; IsoProjection and MIE are linear and multilinear versions of three representative manifold learning algorithms LE, LLE, and Isomap, respectively. We fix the neighborhood size $k = 4$. For each digit, 100 images are randomly selected for training and the remaining 1000 images are used for test. We repeat the experiment 10 times on different randomly selected training sets and calculate the average recognition accuracy.

Table 1 lists the best recognition results and corresponding optimal reduced dimensions of all algorithms. For the same embedding strategy and objective function, L_1-norm based algorithms such as PCA-L1 and TPCA-L1, get much better performance than F-norm based algorithm such as PCA and MPCA. For the same distance metric and objective function, multilinear algorithms such as TLPP and MIE, performs better than the linear algorithms such as LPP and IsoProjection on the second-order image data. For the same distance metric, algorithms that consider discriminative information such as LDA and MLDA, or consider manifold structure such as NPE and TNPE, achieve higher recognition accuracy than the algorithms that only consider global linear structure such as PCA and MPCA. By integrating mulilinear representation, discriminant information, manifold structure, and L_1-norm optimization in a unified framework, M^{2}DE outperforms all the other algorithms.

Table 2: Comparison of recognition accuracy (%) as well as corresponding optimal reduced dimensions on USPS database with random noise

Methods	M^{2}DE	TPCA-L1	MLDA	TLPP
Recog.	**92.1**	90.8	87.6	86.7
Dims	6^2	6^2	7^2	9^2

In the second experiment, we choose three multilinear DR algorithms that have comparatively better performance from the above twelve algorithms to compare with M^{2}DE in detail. TPCA-L1 is an L_1-norm based algorithm; MLDA is a discriminant algorithm; and TLPP is a manifold learning algorithm. We vary the neighborhood size k from 1 to 4 and observe the performance of these algorithms in different reduced dimensions (from 2×2 to 16×16). The results are given in Figure 2. M^{2}DE shows stable and better performance than TPCA-L1, MLDA, and TLPP in most of the reduced dimensions under various values of k.

To further demonstrate that the proposed algorithm is robust to the outliers, we conduct the following experiment. Among 1000 training images, 20 percent are selected and occluded with a square noise consisting of random black and white dots whose size is 4×4, located at a random position. Similarly, 20 percent of 10000 test images are also occluded using the same way. We compare the classification accuracy of M^{2}DE, TPCA-L1, MLDA, and TLPP on the whole dataset with occluded images. The other settings are similar as those in the first experiment. The best average recognition accuracy and the corresponding optimal reduced dimensions of these four algorithms are shown in Table 2. Compared with the results in Table 1, the performance of MLDA and TLPP degrades seriously since the large squared errors dominate the sum when the occluded images appear in the learning procedure. However, the performance of M^{2}DE and TPCA-L1 are relatively robust because the L_1-norm is less sensitive to the outliers.

Honda/UCSD Video Database

In this subsection, we use the first dataset of Honda/UCSD video database (Lee *et al.* 2005) to test the performance of proposed algorithm. This dataset contains 75 videos from 20 human subjects. Each video sequence is recorded in an indoor environment at 15 frames per second, and each lasted for at least 15 seconds. The resolution of each video sequence is 640×480. In our experiment, the original vid-

Table 3: Comparison of recognition accuracy (%) as well as corresponding optimal reduced dimensions on Honda/UCSD video database

Methods	M^2DE	TPCA-L1	MPCA	MLDA	TLPP	TNPE	MIE
Recog.	**95.7**	92.3	89.2	92.6	91.5	90.8	92.3
Dims	3×5×1	5×4×1	6×6×2	3×3×2	10×5×1	3×6×1	2×4×1

eos are downsampled into 64×48 pixels. In order to collect more training and test data, we further cut each original video to several shorter ones of uniform length: 3 seconds, i.e., 45 frames. Therefore, the input data are third-order tensors of size 64×48×45.

We compare M^2DE with the L_1-norm based multilinear algorithm TPCA-L1 as well as five F-norm based multilinear algorithms: MPCA, MLDA, TLPP, TNPE, and MIE. For each individual, we randomly select 10 videos, 5 for training and 5 for test. We fix the neighborhood size $k = 4$. Like previous experiments, we repeat the experiment 10 times and calculate the average recognition accuracy. The recognition accuracy and the corresponding optimal reduced dimensions ($d_1 \times d_2 \times d_3$) of these seven algorithms are reported in Table 3. By integrating L_1-norm based optimization strategy into multilinear maximum distance embedding procedure, M^2DE gives good results on the naturally third-order video data.

Conclusion

This paper proposes a new DR algorithm called multilinear maximum distance embedding (M^2DE). By maximizing the distances between nearby data points and the distances between data points from different classes, the nonlinear structure of the dataset is flattened and the discriminant information is well preserved. By taking the data in the high-order form as the input and explicitly learning the transformation matrices, the tensor structure of data is well kept and the embedding of new data points is straightforward. By employing the L_1-norm to measure the dissimilarity between data points, M^2DE shows more stable embedding results. Experiments on both image and video databases demonstrate that M^2DE outperforms most representative DR techniques on classification tasks.

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS 14*, 585-591.

Bengio, Y.; Paiement, J.; Vincent, P.; Delalleau, O.; Roux, N.L.; and Ouimet, M. 2003. Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. In *NIPS 16*.

Cai, D.; He, X.; and Han, J. 2007. Isometric projection. In *Proc. the 22nd AAAI*, 528-533.

Dai, G., and Yeung, D. Y. 2006. Tensor embedding methods. In *Proc. the 21st AAAI*, 330-335.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7: 179-188.

He, X.; Cai, D.; and Niyogi, P. 2005. Tensor subspace analysis. In *NIPS 18*.

He, X.; Cai, D.; Yan, S.; and Zhang, H. J. 2005. Neighborhood preserving embedding. In *Proc. ICCV*, 1208-1213.

He, X., and Niyogi, P. 2003. Locality preserving projections. In *NIPS 16*.

Hinton, G., and Roweis, S. 2002. Stochastic neighbor embedding. In *NIPS 15*, 833-840.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Edu. Psychol.* 24: 417-441, 498-520.

Huang, H., and Ding, C. 2008. Robust tensor factorization using R1 norm. In *Proc. CVPR*, 1-8.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE TPAMI* 16(5): 550-554.

Kwak, N. 2008. Principal component analysis based on L1-norm maximization. *IEEE TPAMI* 30(9): 1672-1680.

Lathauwer, L. 1997. *Signal processing based on multilinear algebra*. Doctoral Dissertation, E.E. Dept.-ESAT, K.U. Leuven, Belgium.

Lee, K. C.; Ho, J.; Yang, M.; and Kriegman, D. 2005. Visual tracking and recognition using probabilistic appearance manifolds. *Comput. Vis. Image Underst.* 99(3): 303-331.

Liu, Y.; Liu, Y.; and Chan, K. C. C. 2009. Multilinear isometric embedding for visual pattern analysis. In *Proc. ICCV Workshop on Subspace Methods*, 212-218.

Lu, H.; Plataniotis, K. N.; and Venetsanopoulos, A. N. 2008. MPCA: multilinear principal component analysis of tensor objects. *IEEE Trans. Neural Netw.* 19(1): 18-39.

Pang, Y.; Li, X.; and Yuan, Y. 2010. Robust tensor analysis with L1-norm. *IEEE Trans. CSVT* 20(2): 172-178.

Roweis, S. T., and Saul, L.K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290: 2323-2326.

Teh, Y. W., and Roweis, S. T. 2002. Automatic alignment of hidden representations. In *NIPS 15*, 841-848.

Tenenbaum, J.; de Silva, V.; and Langford, J. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290: 2319-2323.

Vasilescu, M. A. O., and Terzopoulos, D. 2003. Multilinear subspace analysis of image ensembles. In *Proc. CVPR*, 93-99.

Weinberger, K. Q., and Saul, L. K. 2004. Unsupervised learning of image manifolds by semidefinite programming. In *Proc. CVPR*, 988-995.

Weinberger, K. Q., and Saul, L. K. 2006. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Proc. the 21st AAAI*, 1683-1686.

Yan, S.; Xu, D.; Yang, Q.; Zhang, L.; Tang, X.; and Zhang, H. J. 2007. Multilinear discriminant analysis for face recognition. *IEEE Trans. Image Process.* 16(1): 212-220.