PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition 37 (2004) 781–788

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Locality pursuit embedding

Wanli Min[a,*], Ke Lu[b], Xiaofei He[c]

[a] *Department of Statistics, The University of Chicago, 5734 S University Ave., Chicago, IL 60637, USA*
[b] *School of Computer Science and Engineering, University of Electronic Science & Technology of China Chengdu, Sichuan 610054, China*
[c] *Computer Science Department, The University of Chicago, 1100 E 58 Street, Chicago, IL 60637, USA*

Received 26 March 2003; accepted 29 September 2003

Abstract

Dimensionality reduction techniques are widespread in pattern recognition research. Principal component analysis, as one of the most popular methods used, is optimal when the data points reside on a linear subspace. Nevertheless, it may fail to preserve the local structure if the data reside on some nonlinear manifold, which is indisputably important in many real applications, especially when nearest-neighbor search is involved. In this paper, we propose *locality pursuit embedding*, a linear algorithm that arises by solving a variational problem. It produces a linear embedding that respects the local geometrical structure described by the Euclidean distances. Some illustrative examples are presented along with applications to real data sets.

Keywords: Locality preserving; Manifold learning; Principal component analysis; Tangent space; Dimension reduction

1. Introduction

Real data of natural and social sciences is often very high dimensional. However, the underlying structure can in many cases be characterized by a small number of parameters. Reducing the dimensionality of such data is beneficial for visualizing the intrinsic structure and it is also an important preprocessing step in many statistical pattern recognition problems.

Recently, there has been extensive interest in developing low-dimensional representations when the data arise from sampling a probability distribution on a manifold [1–5]. Classical techniques for manifold learning, such as PCA [6], MDS [7], are designed to operate when the submanifold is embedded linearly or almost linearly in the observation space. PCA finds a d-dimensional subspace of $\mathbf{R}^n$ which captures as much of the variation in the data set as possible. Specifically, given data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ with zero mean, it finds $y_i = \mathbf{w}^t\mathbf{x}_i$ maximizing

$$\sum_{i=1}^{m} \|y_i - \bar{y}\|^2,$$

where $\mathbf{w}$ is the transformation vector, and $\bar{y} = \sum_i y_i/m$ is the mean. Thus PCA builds a global linear model of the data (a d-dimensional hyperplane). For linearly embedded manifolds, PCA is guaranteed to discover the dimensionality of the manifold and produce a compact representation in the form of an orthonormal basis. However, for the data on a nonlinear submanifold embedded in the feature space, PCA has two problems. First, PCA has difficulty in discovering the underlying structure. For example, the covariance matrix of data sampled from a helix in $\mathbf{R}^3$ has full-rank and thus three principal components. The helix is actually a one-dimensional manifold and can be parameterized with a single parameter. Second, embedding given by PCA preserves only the global structure while local structure is emphasized in many real applications, especially when nearest-neighbor search is involved.

Classical MDS finds an embedding that preserves pairwise distances between data points. It is equivalent to PCA

[*] Corresponding author. Tel.: +1-773-702-0959; fax: +1-773-702-8330.
E-mail address: wmin@galton.uchicago.edu (W. Min).

0031-3203/$30.00 © 2003 Published by Elsevier Ltd on behalf of Pattern Recognition Society.
doi:10.1016/j.patcog.2003.09.005

when those distances are Euclidean. Recently several nonlinear techniques have been proposed to discover the nonlinear structure of the manifold. LLE [2] and Laplacian Eigenmap [1] are local approaches. They essentially seek to map nearby points on a manifold to nearby points in a low-dimensional space. They can approximate a broader range of manifold whose local structure is close to Euclidean but whose global geometry is not. Isomap [3] is a global approach. It builds on classical MDS but seeks to preserve the intrinsic geometry of the data, as captured in the geodesic manifold distances between all pairs of data points.

These nonlinear methods do yield impressive results on some benchmark artificial data sets besides some real applications. However, their nonlinear property makes them computationally expensive. Moreover, they yield mappings that are defined only on the *training* data points and how to evaluate the map on novel *test* points remains unclear.

In this paper, we propose a new *linear* dimensionality reduction algorithm, called locality pursuit embedding (LPE). Heuristically, a nonlinear manifold embedded in $\mathbf{R}^n$ can be characterized by its linear tangent space on each patch. We prove that performing a PCA on each local patch will reveal the tangent space information and thus the projection to the tangent space will preserve the local structure. The new algorithm is distinct from several perspectives.

1. The maps are designed to maximize a different objective function which intends to preserve the local structure, rather than the global structure, as PCA and MDS do. In many real world applications, e.g. image retrieval, one will ultimately need to do a nearest-neighbor search in the low-dimensional space. Since LPE is designed for preserving local structure, it is likely that a nearest-neighbor search in the low-dimensional space will yield similar results to that in the high-dimensional space. This algorithm can be applied to a high-dimensional indexing scheme that would allow quick retrieval.

2. LPE is linear. This makes it fast and suitable for practical applications. While a number of nonlinear techniques (such as Laplacian Eigenmap [1], LLE [2], Isomap [3]) have property (1) above, they are computationally intensive and thus hard to be applied to real problem.

3. LPE can be performed either supervised or unsupervised. In fact, when the class information is available, it can be easily utilized to find a better projection.

The rest of this paper is organized as follows. Section 2 describes the principal component analysis. The LPE algorithm is proposed in Section 3 followed by a justification in Section 4. Experimental results are shown in Section 5. Concluding remarks and future work are in Section 6.

2. Principal component analysis

In the PCA transformation, the sample vector $\mathbf{x}$ is first subtracted by the sample average:

$$\mathbf{x} \leftarrow \mathbf{x} - \overline{\mathbf{x}}. \tag{1}$$

Denote by $\mathbf{X} \in R^{m\times n}$ the matrix whose rows are the centered sample vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbf{R}^n$. The sample covariance matrix $\Sigma = \mathbf{X}^{\mathrm{t}}\mathbf{X}/m$ has the decomposition

$$\Sigma = V\Lambda V^{\mathrm{t}},$$

where $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ are the eigenvalues in descending order and V is an orthogonal matrix whose column vectors are the corresponding eigenvectors of Σ. The optimal projection of $\mathbf{x}_i$ to a d-dimensional $(d < n)$ space is to the space spanned by d leading eigenvectors. In fact, these eigenvectors turn out to be an orthogonal basis of the local tangent space of the intrinsic nonlinear manifold (see Theorem 1). For sample vectors with nonzero mean, the sample covariance is

$$\Sigma = \frac{1}{m}\mathbf{X}^{\mathrm{t}}L\mathbf{X},$$

$$L = I - \frac{1}{m}\mathbf{e}\mathbf{e}^{\mathrm{t}},$$

where $\mathbf{e}$ is a m-dimensional vector taking 1 at each entry. We adopt the same notation throughout this paper. In the sequel $\|\cdot\|$ denotes the L^2 norm of vectors.

3. Locality pursuit embedding

To develop locality-based algorithms for dimensionality reduction, we face two fundamental questions.

1. What is an appropriate representation of local structure?

2. How to preserve the local structure in the space of reduced dimension?

LLE [2] asserts the matrix (w_{ij}), which minimizes $\sum_{i=1}^{m} \|\mathbf{x}_i - \sum_{j=1}^{m} w_{ij}\mathbf{x}_j\|^2$ subject to $\sum_{j=1}^{m} w_{ij} = 1$, as the local structure. Isomap [3] treats the dissimilarity matrix based on estimated geodesic distances as a representation of the local structure. We propose that the local tangent space of the intrinsic manifold as a representation of the local structure, and each data point in the neighborhood can be represented by its local coordinates, i.e. its projection to the nearby tangent space. If the tangent space dimension is much less than the dimension of ambient space, a projection to the local tangent space will achieve the two goals simultaneously: dimensionality reduction and locality preserving.

In this section, we introduce a new linear dimensionality reduction algorithm—LPE. The primary goal of LPE is to preserve local structure.

3.1. The linear dimensionality reduction problem

The generic problem of linear dimensionality reduction can be stated as follows. Given m points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbf{R}^n$ denoted as column vectors, find a transformation matrix $W \in \mathbf{R}^{n\times d}$ that maps these m points to a set of points $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m \in \mathbf{R}^d$ $(d \ll n)$, such that $\mathbf{y}_i = W^{\mathrm{t}}\mathbf{x}_i$ “represents” $\mathbf{x}_i$. Our method is of particular applicability in the special case where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathscr{M}$ and $\mathscr{M}$ is a smooth nonlinear manifold embedded in $\mathbf{R}^n$.

3.2. The algorithm

Given m points $\mathbf{x}_1, \ldots \mathbf{x}_m$ in $\mathbf{R}^n$, for each point $\mathbf{x}_i$, denote by $ne(i)$ the set of its neighborhood. It could be K-nearest neighbor (KNN) or δ-neighborhood, to name a few. To preserve the local structure of $\{\mathbf{x}_k : k \in ne(i)\}$ in $\{\mathbf{y}_k : k \in ne(i)\}$, the data points should be projected to the local tangent space. This can be achieved by maximizing the generalized variance of $\{\mathbf{y}_k : k \in ne(i)\}$, i.e. $\sum_{k \in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2$, which is equivalent to PCA's objective function. Here $\bar{\mathbf{y}}^i$ is the mean vector of $\{\mathbf{y}_k : k \in ne(i)\}$. In principle, it is possible to discover the local geometrical structure by maximizing the objective function in each neighborhood respectively at a price of computation time. As a compromise, we consider the linear projection $\mathbf{y} = W^{\mathrm{t}}\mathbf{x}$ that maximizes a global objective function constructed locally, i.e. the sum of all local objective functions, $\sum_{i=1}^{n} \sum_{k \in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2$. The maximization problem reduces to an eigenvalue problem. The projective directions are the leading eigenvectors. The algorithm procedure is formally stated below.

1. Initialize the $m \times m$ matrix $L = \mathbf{0}$.

2. For each $i = 1, 2, \ldots, m$, denote by $ne(i)$ the KNN of $\mathbf{x}_i$. $k \in ne(i)$ if and only if $\mathbf{x}_k$ is among the KNN of $\mathbf{x}_i$. Define diagonal matrix D_i:

$$D_i = \begin{pmatrix} I_{1 \in ne(i)} & 0 & \cdots \\ 0 & I_{2 \in ne(i)} & 0 \\ \cdots & \cdots & \cdots \\ \cdots & 0 & I_{m \in ne(i)} \end{pmatrix}, \tag{2}$$

where I is an indicator function defined by

$$I_{k \in ne(i)} = \begin{cases} 1, & \mathbf{x}_k \in \text{KNN of } \mathbf{x}_i \\ 0, & \text{otherwise.} \end{cases}$$

with $k = 1, \ldots, m$. Recursively update the L matrix

$$L \leftarrow L + D_i - \frac{1}{K} D_i \mathbf{e}\mathbf{e}^{\mathrm{t}} D_i, \qquad i = 1, 2, \ldots, m.$$

3. Solve the following eigenvector problem:

$$\mathbf{X}^{\mathrm{t}} L\, \mathbf{X}\mathbf{w} = \lambda \mathbf{w}. \tag{3}$$

Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be the solutions of Eq. (3), ordered according to their eigenvalues, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$. The matrix L is symmetric positive semi-definite, so is $\mathbf{X}^{\mathrm{t}} L \mathbf{X}$. Therefore, the eigenvectors are orthogonal. The embedding is as follows:

$$W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_d),$$

$$\mathbf{x}_i \rightarrow \mathbf{y}_i = W^{\mathrm{t}} \mathbf{x}_i.$$

4. Justification

In this section, we give a theoretical justification of our algorithm.

4.1. Local tangent space

In many examples, the observed data points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ can be considered lying on a submanifold $\mathscr{M}$ that is linearly or almost linearly embedded in $\mathbf{R}^n$ of higher dimension. PCA, as a dimension-reduction tool, chooses the projections that best represent the whole data in the sense of smallest global reconstruction error, but the directions do not necessarily best preserve the locality.

In the case that sample vectors reside on a smooth *nonlinear* manifold, it is necessary to consider its local geometry (local tangent space) rather than global structure. To this aim, based on the Euclidean distances ($\|\mathbf{x}_i - \mathbf{x}_j\|$) metric, we can perform a PCA on the neighborhood. The next theorem shows that under some regularity conditions, local tangent space can be constructed based on the eigenvectors of the local sample covariance matrix and the local principal components give a representation of data sets in this tangent space. To state the theorem, let $\mathscr{M}$ be a manifold embedded in $\mathbf{R}^n$, we begin with two assumptions

1. *Local smoothness*: The manifold has a sufficiently smooth (at least locally) generating function $\mathbf{x} = \mathbf{f}(\mathbf{z}) \in \mathbf{R}^n, \mathbf{z} \in \mathbf{R}^d$ where $d \ll n$.
2. *Dense sampling*: The observed data points $\mathbf{x}_i = \mathbf{f}(\mathbf{z}_i), i = 1, 2, \ldots$ is a simple random sample with $\mathbf{z}_i$s are independent and identically distributed (iid) with mean $\mu_{\mathbf{z}}$, covariance matrix $C = E(\mathbf{z} - \mu_{\mathbf{z}})\ (\mathbf{z} - \mu_{\mathbf{z}})^{\mathrm{t}}$ of full rank. Denote the observed sample vectors in the neighborhood of $\mathbf{f}(\mu_{\mathbf{z}})$ by $\mathbf{x}_i = \mathbf{f}(\mathbf{z}_i), i = 1, \ldots, m$. Assume $m \to \infty$ as total sample size increases.

The next theorem shows that local tangent space can be constructed based on the eigenvectors of the local sample covariance matrix.

Theorem 1. *Under assumptions* 1, 2, *let* $\Sigma = \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{t}}/m$ *be the sample covariance matrix of* $\mathbf{x} = \mathbf{f}(\mathbf{z})$, *then the eigenvectors of* Σ *form a basis of the tangent space of* $\mathscr{M}$ *at the* $\mathbf{f}(\mu_{\mathbf{z}})$.

Proof. Since $\mathbf{f}(\mathbf{z})$ is sufficiently smooth, it is differentiable near $\mu_{\mathbf{z}}$ and its Jacobian $J_{\mathbf{f}}(\mu_{\mathbf{z}}) = (\partial \mathbf{f}/\partial \mathbf{z}_1, \ldots, \partial \mathbf{f}/\partial \mathbf{z}_d)|_{\mu_{\mathbf{z}}} \in R^{n \times d}$, its tangent space at $\mathbf{f}(\mu_{\mathbf{z}})$ is spanned by the columns of $J_{\mathbf{f}}(\mu_{\mathbf{z}})$. By first-order Taylor expansion:

$$\mathbf{f}(\mathbf{z}) = \mathbf{f}(\mu_{\mathbf{z}}) + J_{\mathbf{f}}(\mu_{\mathbf{z}})(\mathbf{z} - \mu_{\mathbf{z}}) + \mathbf{O}(\|\mathbf{z} - \mu_{\mathbf{z}}\|^2). \tag{4}$$

The central limit theorem implies $\bar{\mathbf{z}} - \mu_{\mathbf{z}} \sim \mathbf{N}_d(\mathbf{0}, W/m)$, it follows that, up to order $\mathbf{O}(\|\mathbf{z} - \mu_{\mathbf{z}}\|)$,

$$\bar{\mathbf{x}} = \mathbf{f}(\mu_{\mathbf{z}}) + J_{\mathbf{f}}(\bar{\mathbf{z}} - \mu_{\mathbf{z}}) \approx \mathbf{f}(\mu_{\mathbf{z}}) + \mathbf{O}_{\mathbf{p}}(m^{-1/2}). \tag{5}$$

We say $x \sim \mathbf{O}_{\mathbf{p}}(m^{-1/2})$ if for any $\varepsilon > 0$, exists $A_\varepsilon \in (0, \infty)$ such that $P(|\sqrt{m}x| > A_\varepsilon) < \varepsilon$. Referring to Eqs. (4), (5), we rewrite the sample covariance matrix $\mathbf{\Sigma}$ as follows:

$$\mathbf{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{t}}$$

Table 1
Eigenvalues and eigenvectors of local variance matrix of (x, e^x) near $x = 0$.

$x \in (-1.0, 1.0)$		$x \in (-0.5, 0.5)$		$x \in (-0.25, 0.25)$	
0.013	0.812	0.001	0.171	0.000	0.043
$\begin{pmatrix} -0.766 \\ 0.642 \end{pmatrix}$	$\begin{pmatrix} 0.642 \\ 0.766y \end{pmatrix}$	$\begin{pmatrix} -0.717 \\ 0.697 \end{pmatrix}$	$\begin{pmatrix} 0.697 \\ 0.717 \end{pmatrix}$	$\begin{pmatrix} -0.704 \\ 0.710 \end{pmatrix}$	$\begin{pmatrix} 0.710 \\ 0.704 \end{pmatrix}$

$$= \frac{1}{m}\sum_{i=1}^{m} J_{\mathbf{f}}(\mu_{\mathbf{z}})(\mathbf{z}_i - \mu_{\mathbf{z}})(\mathbf{z}_i - \mu_{\mathbf{z}})^{\mathrm{t}} J_{\mathbf{f}}^{\mathrm{t}}(\mu_{\mathbf{z}}) + \mathbf{O}_{\mathbf{p}}(m^{-1})$$

$$\approx J_{\mathbf{f}}(\mu_{\mathbf{z}})\; C J_{\mathbf{f}}^{\mathrm{t}}(\mu_{\mathbf{z}}).$$

Let $\mathbf{v}$ be the eigenvector of Σ with corresponding eigenvalue λ, i.e.

$$\Sigma \mathbf{v} = J_{\mathbf{f}}(\mu_{\mathbf{z}})\; C J_{\mathbf{f}}^{\mathrm{t}}(\mu_{\mathbf{z}})\mathbf{v} = \lambda \mathbf{v}. \tag{6}$$

Notice $J_{\mathbf{f}}(\mu_{\mathbf{z}})\; C J_{\mathbf{f}}^{\mathrm{t}}(\mu_{\mathbf{z}})$ has the same rank (denoted by r) as $J_{\mathbf{f}}(\mu_{\mathbf{z}})$ since C is symmetric and positive definite. Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$ be the r leading positive eigenvalues in Eq. (6) with corresponding normalized eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$, it follows that

$$\mathbf{v}_i = \frac{1}{\lambda_i} J_{\mathbf{f}}(\mu_{\mathbf{z}})\; W J_{\mathbf{f}}^{\mathrm{t}}(\mu_{\mathbf{z}})\mathbf{v}_i. \tag{7}$$

Since Eq. (7) holds for every i, clearly each of $\mathbf{v}_1, \ldots, \mathbf{v}_r$ is in the (tangent) space spanned by the columns of $J_{\mathbf{f}}(\mu_{\mathbf{z}})$ (of dimension r). The orthogonality of $\mathbf{v}_1, \ldots, \mathbf{v}_r$ shows that they form a basis of the local tangent space of $\mathscr{M}$ near $\mathbf{f}(\mu_{\mathbf{z}})$. □

Remark 1. The two assumptions arise naturally in view of the bias-variance trade-off phenomena in many nonparametric regression problems. As a matter of fact, there are two competing constraints, choosing a neighborhood small enough such that Taylor expansion Eq. (4) can be justified and keeping sufficiently many data points in the neighborhood such that the local covariance matrix can be well estimated.

Remark 2. Rather than treating the data points in the same neighborhood equally, we could introduce a weight function (Heat kernel, Epanechnikov kernel, etc.) and compute the weighted covariance matrix. It can improve the robustness of the algorithm against outliers.

Remark 3. Denote $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in R^{n \times r}$. The projection of $\mathbf{x}_i$ onto the local tangent space is $\mathbf{y}_i = V^{\mathrm{t}}(\mathbf{x}_i - \bar{\mathbf{x}})$. This is the same as principal components found by PCA since PCA solves the same equation as Eq. (6). So a geometrical interpretation of PCA is that the data points are embedded to the local tangent space of lower dimension by orthogonal projections.

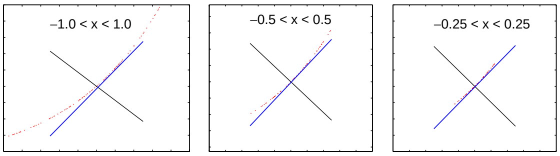

Fig. 1. The eigenvectors of local covariance matrix in the neighborhood $|x| < 1$, 0.5, 0.25 (from left to right). The leading direction is in blue.

As an illustration of Theorem 1, we give a geometrically intuitive example.

Example 1. We sample 100 data points according to the relationship $y = \exp(x)$:

$$y_i = \exp(x_i) \qquad i = 1, \ldots, 100$$

$x_i \sim \mathrm{Unif}(-1, 1)$. In Table 1, we summarize the empirical eigenvectors and corresponding eigenvalues of the local variance matrix in the neighborhood of $|x| < 0.25, |x| < 0.5$ and $|x| < 1.0$. Clearly the eigenvectors approach the tangent direction $(1/\sqrt{2}, 1/\sqrt{2})^{\mathrm{t}}$ and normal direction $(-1/\sqrt{2}, 1/\sqrt{2})^{\mathrm{t}}$ of the curve $y = \exp(x)$ at the point $(x = 0, y = 1)$ (See Fig. 1) as the chosen neighborhood is more concentrated around the $(x = 0, y = 1)$.

There are two issues worth pointing out. First, if the neighborhood is too small so that it contains few data points or too wide so that the Taylor expansion, Eq. (4), incurs large error, then the local covariance matrix's eigenvectors will deviate from the true tangent (normal) direction.

Second, if the data points contain an independent noise term, i.e. $y_i = \exp(x_i) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$, then the result is worse but the eigenvectors still approach the local tangent (normal) direction provided that σ^2 is small (say 0.01 in this example).

4.2. Optimal linear embedding

As seen from the above result, pursuit of the maximum variance of projection within a small neighborhood leads to a discovery of the local tangent space, which preserves the structure faithfully.

It is well known that classical manifold learning technique (PCA, MDS) will fail if the nonlinear structure of submanifold cannot be regarded as a small perturbation from a linear approximation. However, we can view the data points as a union of small patches on a nonlinear manifold, and each patch is homeomorphic to a Euclidean space given it is small. By Theorem 1, performing a PCA on each small patch can preserve locality.

For each $\mathbf{x}_i$, denote by $ne(i)$ the indices of its neighborhood, including $\mathbf{x}_i$ itself. In this paper we set $ne(i)$ to the KNN. The $ne(i)$ is a small patch around $\mathbf{x}_i$ on the intrinsic manifold. Suppose this neighborhood is small enough so that the submanifold could be well approximated linearly on this patch. To this point, we look for an embedding that attains locally maximal variance under appropriate constraint, in analogy to PCA. So the objective function on this patch is $\sum_{k\in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2$ where $\bar{\mathbf{y}}^i$ is the mean vector of $\{\mathbf{y}_k : k \in ne(i)\}$.

An ideal embedding should seek to maximize $\sum_{k\in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2$ on each patch $ne(i)$, which is unrealistic in most cases. To this end, a reasonable criterion is to maximize a global objective function, which is the sum of the objective functions on each patch:

$$\sum_{i=1}^{m} \sum_{k\in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2. \tag{8}$$

To simplify the expression, let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)^{\mathrm{t}}$, $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_m)^{\mathrm{t}}$ be the data matrix of the m observations $\mathbf{x}_k \in \mathbf{R}^n$ and their projections $\mathbf{y}_k \in \mathbf{R}^d$, then we have

$$\begin{aligned}
&\sum_{k\in ne(i)} (\mathbf{y}_k - \bar{\mathbf{y}}^i)(\mathbf{y}_k - \bar{\mathbf{y}}^i)^{\mathrm{t}} \\
&= \frac{1}{2K} \sum_{k,l\in ne(i)} (\mathbf{y}_k - \mathbf{y}_l)(\mathbf{y}_k - \mathbf{y}_l)^{\mathrm{t}} \\
&= \frac{1}{2K} \sum_{k,l\in ne(i)} (\mathbf{y}_k\mathbf{y}_k^{\mathrm{t}} + \mathbf{y}_l\mathbf{y}_l^{\mathrm{t}} - \mathbf{y}_k\mathbf{y}_l^{\mathrm{t}} - \mathbf{y}_l\mathbf{y}_k^{\mathrm{t}}) \\
&= \sum_{k\in ne(i)} \mathbf{y}_k\mathbf{y}_k^{\mathrm{t}} - \frac{1}{K}\left(\sum_{k\in ne(i)} \mathbf{y}_k\right)\left(\sum_{k\in ne(i)} \mathbf{y}_k\right)^{\mathrm{t}} \\
&= (D_i\mathbf{Y})^{\mathrm{t}}(D_i\mathbf{Y}) - \frac{1}{K}(D_i\mathbf{Y})^{\mathrm{t}}\mathbf{e}\,\mathbf{e}^{\mathrm{t}}(D_i\mathbf{Y}) \\
&= \mathbf{Y}^{\mathrm{t}}(D_i - \frac{1}{K} D_i\mathbf{e}\,\mathbf{e}^{\mathrm{t}}D_i)\mathbf{Y}.
\end{aligned}$$

The last two "=" in the preceding display hold since $D_i\mathbf{Y} = (\mathbf{y}_1 I_{1\in ne(i)}, \dots, \mathbf{y}_m I_{m\in ne(i)})^{\mathrm{t}}$, i.e. the projection of $\{\mathbf{x}_k : k \notin ne(i)\}$ is set to zero and the others remain unchanged. Observe $\|\mathbf{y}\|^2 = Tr(\mathbf{y}\mathbf{y}^{\mathrm{t}})$ for any vector $\mathbf{y} \in \mathbf{R}^d$ which entails the following:

$$\begin{aligned}
&\sum_{i=1}^{m} \sum_{k\in ne(i)} \|\mathbf{y}_k - \bar{\mathbf{y}}^i\|^2 \\
&= \sum_{i=1}^{m} \sum_{k\in ne(i)} Tr[(\mathbf{y}_k - \bar{\mathbf{y}}^i)(\mathbf{y}_k - \bar{\mathbf{y}}^i)^{\mathrm{t}}] \\
&= Tr\left[\mathbf{Y}^{\mathrm{t}} \sum_{i=1}^{m} \left(D_i - \frac{1}{K} D_i\mathbf{e}\,\mathbf{e}^{\mathrm{t}}D_i\right)\mathbf{Y}\right] \\
&= Tr[\mathbf{Y}^{\mathrm{t}}L\mathbf{Y}],
\end{aligned}$$

where L is a symmetric positive semi-definite matrix of $m \times m$ dimension.

$$L = \begin{pmatrix} d_{11}(1-\frac{1}{K}) & -\frac{1}{K}d_{12} & \cdots & -\frac{1}{K}d_{1m} \\ -\frac{1}{K}d_{12} & d_{22}(1-\frac{1}{K}) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{1}{K}d_{1m} & \cdots & \cdots & d_{mm}(1-\frac{1}{K}) \end{pmatrix}. \tag{9}$$

Here $d_{kl} = \sum_{i=1}^{m} I_{k,l\in ne(i)}$ is the number of neighborhoods into which both $\mathbf{x}_k$, $\mathbf{x}_l$ fall.

If we consider only linear transformations: $\mathbf{y}_i = W^{\mathrm{t}}\mathbf{x}_i$, where $W = (\mathbf{w}_1, \dots, \mathbf{w}_d)$ subject to the constraint $W^{\mathrm{t}}W = I_d$, the objective function reduces to a quadratic form: $Tr[W^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}L\mathbf{X}W] = \sum_{i=1}^{d} \mathbf{w}_i^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}L\mathbf{X}\mathbf{w}_i$. And the optimization problem:

$$\max_{\|\mathbf{w}\|=1} \sum_{i=1}^{d} \mathbf{w}^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}L\mathbf{X}\mathbf{w} \tag{10}$$

is solved by taking $\mathbf{w}_1, \dots, \mathbf{w}_d$ as the d leading eigenvectors of $\mathbf{X}^{\mathrm{t}}\ L\mathbf{X}$. This leads to step 3 in our algorithm.

Interestingly, it can be shown that L given by Eq. (9) is the Laplacian matrix of a weighted graph with the weight matrix $W^{\mathrm{t}}W/K$, where $W = (w_{ij})_{m\times m}$ with elements $w_{ij} = I_{j\in ne(i)}$. For details, see [8]. This observation leads to a simplified coding.

5. Experiment results

Some illustrative examples are discussed in this section. We also apply LPE to image retrieval, in which the local structure is particularly important.

5.1. Simply synthetic example

We illustrate the difference between PCA and LPE by a simple synthetic example.

Example 2. Consider the surface $Z = 4\exp(-X^2 - Y^2/4)$ in $\mathbf{R}^3$. The sample points are drawn from this surface with

Gaussian noise.

$$(X, Y, Z)^t$$
$$=(R\cos(\theta) + \varepsilon_1, 2R\sin(\theta) + \varepsilon_2, 4\exp(-R^2) + \varepsilon_3)^t,$$

where $R=0.5, 1, 2, \theta=k/180\pi, k=1,\ldots,360, \varepsilon_i \sim N(0, 0.01)$, iid. Although the data points are in $\mathbf{R}^3$, they are defined by two parameters x, y, therefore the projection to the X, Y plane preserves the structure of the data points faithfully.

The leading two projections by PCA are in the directions of $(0,1,0)^t$, $(0,0,1)^t$ which are the Y, Z axes. In Fig. 2 the three original rings $R = 0.5, 1.0, 2.0$ are green, red and blue, which are the same as their projections. The projection consists of three segments corresponding to the three rings. Clearly the mapping from a ring to a line segment collapses the structure, in other words, it is not isomorphic. The phenomenon arises since PCA projection is trying to align the three rings and therefore the "between" structure among the three rings makes it unable to preserve the "within" structure.

LPE, on the other hand, returns the leading two projections along $(1,0,0)^t$, $(0,1,0)^t$, which is the X, Y plane. Each ring in the original data is still projected to a ring. Clearly it successfully preserves the structure not only locally, but also globally.

5.2. Handwritten digital images

We applied our algorithm to collections of images of handwritten digits. The data set is from the MNIST database. It consists of 974 examples of the digit "8". The size of each image is 28 × 28 pixels, with 256 gray levels per pixel. Thus each image is represented by a vector of 784 dimension. We apply LPE to construct a two-dimensional representation of these 784-dimensional vectors (see Fig. 3). Each point in the two-dimensional space corresponds to an image of digit "8". We select several points in the first dimension (horizontal) and the second dimension (vertical). These selected points are connected to give a guidance of the direction. Their corresponding images of digit "8" are shown along respective direction. The first dimension appears to describe the slant of each digit and the second dimension appears to describe the fatness of each digit. More specifically, along the horizontal direction starting from left, the selected digits changes from right-slanted to no slant; going upward along the vertical direction, the selected digits changes from fat to thin. As can be seen, though our method is still a linear algorithm, it is somehow capable of discovering the nonlinear structure of the data manifold.

5.3. Image retrieval

Due to the rapid growth of the number of digital images, there is an increasing demand for effective image management tools. Content-based image retrieval [9–11] use low-level features (color, texture, shape, etc.) automatically extracted from the images themselves to search for images relevant to a user's query. Typically, the dimensions of image feature vector range from tens to hundreds. For

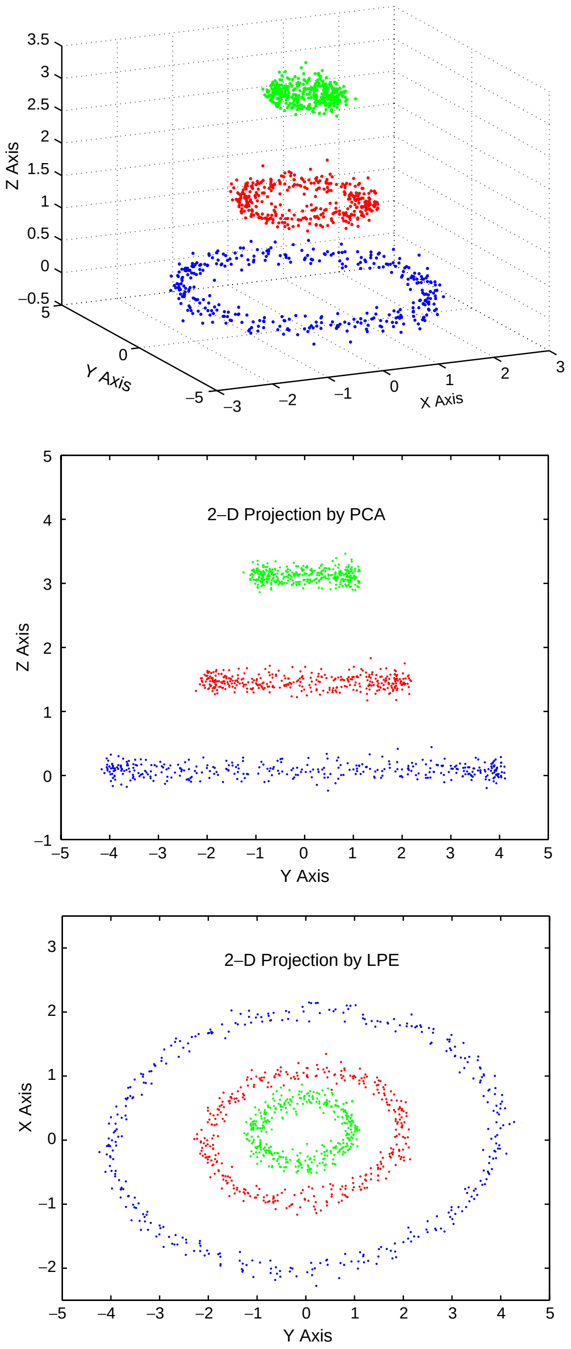

Fig. 2. 3-D plot of sample(left); 2-D projection by PCA(middle); 2-D projection by LPE (right).

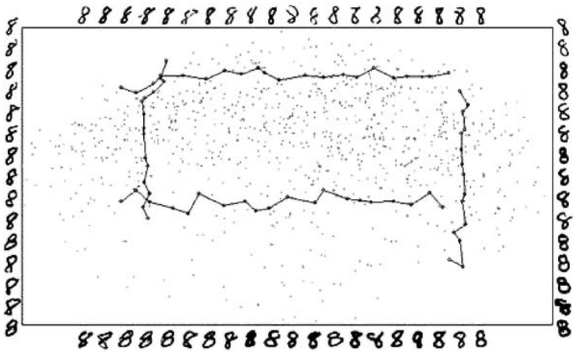

Fig. 3. 2-D visualization of handwritten digital images. The fatness of each digit varies across horizontal direction and the slant of each digit changes along vertical direction.

example, a color histogram may contain 256 bins. High dimensionality creates several problems for image retrieval. First, learning from examples is computationally infeasible if it has to rely on high-dimensional representations, which is known as “curse of dimensionality”. Learnability thus necessitates dimensionality reduction. Second, in large multimedia databases, high-dimensional representation is computationally intensive and most users do not wait around to provide a great deal of feedbacks. Hence for storage and speed concern, dimensionality reduction is needed.

PCA is one of the most frequently used linear algorithms for high-dimensional indexing. It is optimal in the global sense. However, in image space, the local structure is more important than the global structure in most cases. In fact, if the distance between two images is large enough, then the absolute distance makes little sense. For example, it is meaningless to say that a tiger is more similar to a dog than to a horse. Therefore, the locality preserving property is especially important for image retrieval.

In this section, we performed several experiments to evaluate the effectiveness of the proposed approach over a large image database. The database we use consists of 3000 images of 30 semantic categories from the Corel dataset. It is a large and heterogeneous image set. A retrieved image is considered correct if it belongs to the same category of the query image. Three types of color features and three types of texture features are used in our system, which are listed in Table 2. Each image is represented by a 435-dimensional vector in the image space.

Table 2
Image features used for retrieval

Color-1	Color histogram in HSV space with quantization 256
Color-2	First and second moments in Lab space
Color-3	Color coherence vector in LUV space with quantization 64
Color-4	Tamura coarseness histogram
Color-5	Tamura dictionary
Color-6	Pyramid wavelet texture feature

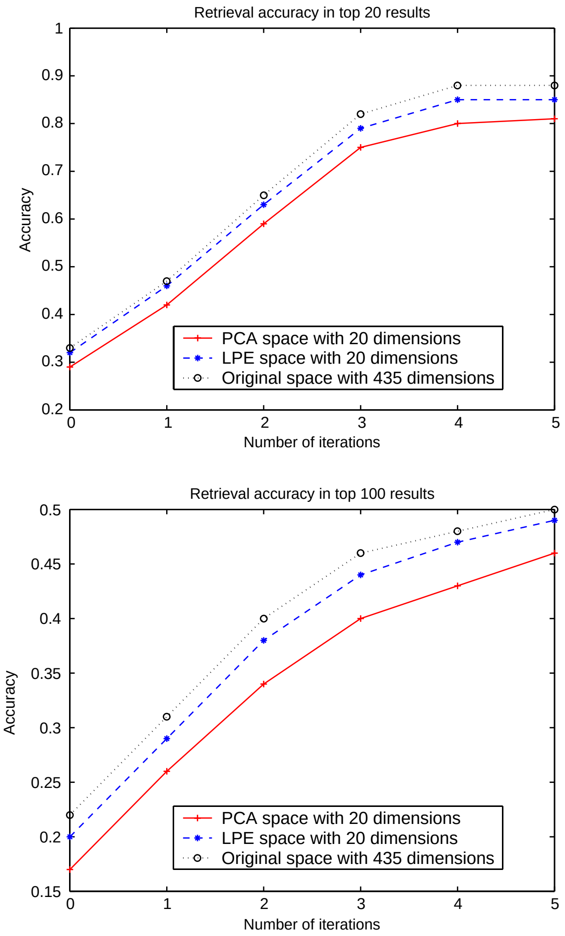

Fig. 4. Comparison of image retrieval performance in the original space, PCA space and LPE space.

We designed an automatic feedback scheme to model the retrieval process. At each iteration, the system marks the first three incorrect images from the top 50 matches as irrelevant examples and also selects at most 3 correct images as relevant examples (relevant examples in the previous iterations are excluded from the selection). These automatically generated feedbacks are added into the query example set to refine the retrieval. To evaluate the performance of our algorithms, we define the retrieval accuracy as follows:

$$\mathit{Accuracy} = \frac{\text{relevant images retrieved in top } N \text{ returns}}{N}.$$

We compare the retrieval performances in the original space, PCA space with 20 dimensions and LPE space with 20 dimensions. Fig. 4 shows the fraction of relevant images among the top $N(=20,100)$ images returned in each space, as a function of the number of iterations of user feedback. Rui's relevance feedback scheme [9] is used in these experiments.

As can be seen, image retrieval performs better in LPE space than in PCA space. Furthermore, though the dimensionality of the original space is very high, its intrinsic dimensionality is very low, hence we can reduce it to a low-dimensional subspace without sacrificing much performance.

6. Conclusion and future work

A new linear dimensionality reduction algorithm, called locality pursuit embedding (LPE) is introduced in this paper. Different from PCA which preserves the global structure by maximizing the variance of the whole data set, LPE preserves the local structure by maximizing the variance of each local patch. This is of particular interest in the real applications which emphasize local structure, especially when the nearest neighbor kind of classifiers are used. This observation leads to a new criterion for choosing appropriate projective maps. Experimental results show the effectiveness of our algorithm.

In this paper, we use LPE to seek a projection which preserves local structure. It works in an unsupervised learning manner. When the training samples are labelled, a possible extension of our method is to incorporate such prior information into the adjacency graph to make our approach have more discriminatory power. Furthermore, with neighborhood preserving character, the LPE algorithm has a natural relationship with clustering. We are currently exploring this relationship in theory and practice.

Acknowledgements

The authors are very grateful to the anonymous referee for their suggestions and comments, which improved the presentation of this paper.

References

[1] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Advances in Neural Information Processing Systems, Vol. 15, Vancouver, British Columbia, Canada, 2001.

[2] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[3] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[4] M. Brand, Charting a Manifold, NIPS, Vancouver, Canada, 2002.

[5] X. He, P. Niyogi, Locality Preserving Projections, NIPS, Vancouver, Canada, 2003.

[6] I.T. Jolliffe, Principal Component Analysis, Springer, NY, 1989.

[7] T. Cox, M. Cox, Multidimensional Scalling, Chapman & Hall, London, 1994.

[8] Fan R.K. Chung, Spectral Graph Theory, Regional Conference Series in Mathematics, No. 92, 1997.

[9] Y. Rui, T.S. Huang, S. Mehrotra, M. Ortega, A relevance feedback architecture for content-based multimedia information retrieval systems, in: Proceedings of the IEEE Workshop Content-Based Access of Image and Video Libraries, Puerto Rico, 1997, pp. 82–89.

[10] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dorn, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the QBIC system, IEEE Comput. 28 (1995) 23–32.

[11] W.Y. Ma, B.S. Manjunath, Netra: a toolbox for navigating large image databases, Multimedia Sys. 7 (1999) 184–198.

About the Author—WANLI MIN received the B.S. degree in Physics from University of Science and Technology of China in 1997. He is currently working toward the Ph.D. degree in the Department of Statistics, the University of Chicago. His research interests include time series analysis, probability theory and statistical learning theory.

About the Author—KE LU received the B.S. degree in thermal power engineering from Chongqing University, China in 1996, and the M.S. degree in Computer Engineering in 2003 from the University of electronic science and technology of China where he is currently a lecturer. His research interests include pattern recognition and multimedia information retrieval.

About the Author—XIAOFEI HE received the B.S. degree in computer science from Zhejiang University, Zhejiang, China, in 2000. He is currently working toward the Ph.D. degree in the Department of Computer Science, University of Chicago, Chicago, IL. His research interests are manifold learning, statistical learning theory, face recognition and image retrieval.