Continuum Isomap for manifold learnings[☆]

Hongyuan Zha[a,1], Zhenyue Zhang[b,*,2]

[a]*College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[b]*Department of Mathematics, Zhejiang University, Yuquan Campus, Hangzhou 310027, PR China*

Available online 13 December 2006

Abstract

Recently, the Isomap algorithm has been proposed for learning a parameterized manifold from a set of unorganized samples from the manifold. It is based on extending the classical multidimensional scaling method for dimension reduction, replacing pairwise Euclidean distances by the geodesic distances on the manifold. A continuous version of Isomap called continuum Isomap is proposed. Manifold learning in the continuous framework is then reduced to an eigenvalue problem of an integral operator. It is shown that the continuum Isomap can perfectly recover the underlying parameterization if the mapping associated with the parameterized manifold is an isometry and its domain is convex. The continuum Isomap also provides a natural way to compute low-dimensional embeddings for out-of-sample data points. Some error bounds are given for the case when the isometry condition is violated. Several illustrative numerical examples are also provided.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Nonlinear dimension reduction; Manifold learning; Continuum Isomap; Isometric embedding; Perturbation analysis

1. Introduction

The continuous increase in computing power and storage technology makes it possible for us to collect and analyze ever larger amounts of data. In many real-world applications, the data are large and of high dimension; those applications include computational genomics, image analysis and computer vision, document analysis in information retrieval and text mining. Fortunately, in many of those applications, all of the components of those high-dimensional data vectors are not independent of each other and the data points can be considered as lying on or close to a low-dimensional nonlinear manifold embedded in a high-dimensional space. Learning the nonlinear low-dimensional structures hidden in a set of unorganized high-dimensional data points, known as *manifold learning*, represents a very useful and challenging unsupervised learning problem (Roweis and Saul, 2000; Tenenbaum et al., 2000).

Traditional dimension reduction techniques such as principal component analysis and factor analysis usually work well when the data points lie close to a *linear* (affine) subspace of the high-dimensional data space (Hastie et al., 2001). They cannot, in general, discover nonlinear structures embedded in the set of data points. Recently, two novel methods

for manifold learning, the locally linear embedding method (LLE) in Roweis and Saul (2000) and the Isomap method in Tenenbaum et al. (2000), have drawn great interests. Unlike other nonlinear dimension reduction methods, both LLE and Isomap methods emphasize simple algorithmic implementation and avoid nonlinear optimization formulations that are prone to local minima. The focus of this paper is on analyzing the Isomap method, which extends the classical multidimensional scaling (MDS) method by exploiting the use of geodesic distances of the underlying nonlinear manifold (details of Isomap will be presented in Section 2). Since a nonlinear manifold can be parameterized in infinitely many different ways, it is not apparent what parametrization is actually being discovered by a nonlinear dimension reduction method such as Isomap, so one can pose a fundamental question of both theoretical and practical interests as follows:

> What is the low-dimensional nonlinear structure that Isomap tries to discover and for what type of nonlinear manifolds, can Isomap perfectly recover the low-dimensional nonlinear structure?

The general question of when Isomap performs well is first addressed by Bernstien et al. (2000) and Tenenbaum et al. (2000) where asymptotic convergence results are derived for Isomap, highlighting the importance of geodesic convexity of the underlying manifold and isometry for the success of Isomap at recovering the low-dimensional nonlinear structure. Extensions to conformal mappings are also discussed by de Silva and Tenenbaum (2002). Some of the aspects of the question have further been analyzed by Donoho and Grimes (2002) under the framework of continuum Isomap emphasizing nonlinear manifolds constructed from collections of images. In Donoho and Grimes (2002) it is *defined* that continuum Isomap obtains a perfect recovery of the natural parameter space of the nonlinear manifold in question if the geodesic distance on the nonlinear manifold is proportional to the Euclidean distance in the parameter space. Unfortunately, no continuous version of Isomap is presented in Donoho and Grimes (2002) and in all the work we have just mentioned the reason why (continuum) Isomap should work perfectly is explained using the discrete framework of the classical MDS.

The purpose of this paper is to fill those gaps by presenting a continuous version of Isomap using *integral* operators. In particular, we show that for a nonlinear manifold that can be isometrically embedded onto an open and convex subset of an Euclidean space, the continuum Isomap computes a set of eigenfunctions which forms the canonical coordinates (i.e., coordinates with respect to the canonical basis) of the Euclidean space up to a rigid motion. For non-flat manifolds, we argue that certain information will be lost if the Isomap only makes use of a finite number of the eigenfunctions. More importantly, we emphasize that isometry is a more fundamental property than geodesic distance with regard to manifold learning. Local manifold learning methods such as LLE (Roweis and Saul, 2000) and LTSA (Zhang and Zha, 2004) are called for when global methods such as Isomap fail. Besides its theoretical interest, continuum Isomap also provides a more disciplined approach for computing low-dimensional embedding for out-of-sample data points. It also provides a convenient framework for deriving perturbation bounds to deal with the non-isometric case.

The rest of the paper is organized as follows: in Section 2, we review both the classical MDS and its generalization Isomap proposed by Tenenbaum et al. (2000). Several basic concepts from differential geometry such as parameterized manifolds, isometric embedding and geodesic distances will be recalled in Section 3, as a preparation for the discussion of the continuum Isomap derived in the next section. We will show, in Section 4, that the continuum Isomap can perfectly recover the parameter space of the parameterized manifold in question if the geodesic distance on the manifold is proportional to the Euclidean distance in the parameter space. We also illustrate the role played by an Isometry. In Section 5, we derive some perturbation bounds to deal with the case when the isometry assumption is violated. We pay special attention to the case when the mapping associated with the parameterized manifold is bi-Lipschitz. The relation between the continuous and discrete versions of Isomap will be discussed in Section 6. We also show how to compute the low-dimensional embedding for an arbitrary out-of-sample data point on the manifold. Section 7 contains several concluding remarks.

2. Classical multidimensional scaling and Isomap

Suppose for a set of N points $\{x_i\}_{i=1}^N$, $x_i \in \mathscr{R}^m$ with $N > m$, we are given the set of pairwise Euclidean distances

$$d(x_i, x_j) = \|x_i - x_j\|_2,$$

and we are asked to reconstruct the $\{x_i\}$'s from the above set of pairwise distances. We can proceed as follows: without loss of generality, we can assume that the $\{x_i\}$'s are centered, i.e., $\sum_{i=1}^N x_i = 0$. Notice that the squared pairwise distance

reads as

$$d^2(x_i, x_j) = \|x_i - x_j\|_2^2 = x_i^T x_i - 2x_i^T x_j + x_j^T x_j.$$

Denoting by $\psi = [x_1^T x_1, \ldots, x_N^T x_N]^T$ and $X = [x_1, \ldots, x_N]$, the squared-distance matrix $D = [d^2(x_i, x_j)]_{i,j=1}^N$ can be written as

$$D = \psi e^T - 2X^T X + e\psi^T,$$

where e is the N-dimensional vector of all 1's. Since e is a null vector of $J = I - (1/N)ee^T$, ψ can be eliminated by multiplying J on the two sides of D, yielding

$$B \equiv -\frac{1}{2} JDJ = X^T X. \tag{1}$$

Thus, X can be recovered, up to an orthogonal transformation, by the eigendecomposition of B,

$$B = U \operatorname{diag}(\lambda_1, \ldots, \lambda_m)U^T,$$

with $\lambda_1 \geqslant \cdots \geqslant \lambda_m \geqslant 0$ ($\lambda_m = 0$ since X has been centered), and $U \in \mathscr{R}^{N \times m}$ orthonormal, and

$$X = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_m^{1/2})U^T.$$

The preceding embedding result is essentially due to Schoenberg (1935) and Young and Householder (1938) (see also Gower, 1966; Torgerson, 1952 for some related development). The method is generally known as the *classical* MDS, and we will emphasize its role as a dimensional reduction technique. The field of *general* MDS has substantially extended the above idea with the pairwise Euclidean distances replaced by various dissimilarities. Usually visualization of those dissimilarities in a low-dimensional Euclidean space is emphasized with the computation done using nonlinear optimization methods (Cox and Cox, 2001; Williams, 2002).

To motivate the introduction of Isomap (Tenenbaum et al., 2000), we notice that when the set of data points x_i's lie on or close to a low-dimensional nonlinear manifold embedded in a high-dimensional space and the nonlinear structure cannot be adequately represented by a linear approximation, classical MDS tends to fail to recover the low-dimensional structure of the nonlinear manifold. We illustrate this issue using a set of examples. On the top row of Fig. 1, we plot two sets of sample points from 1D curves generated as $x_i = f(\tau_i^*) + \varepsilon_i, i = 1, \ldots, N$, N is the total number of points plotted, the τ_i^*'s are chosen uniformly from a finite interval and ε_i are Gaussian noise. The first curve is a straight line and the second and the third curves are the same parabola. Each figure in the second row plots the τ_i^*'s against the computed τ_i of the 1D embeddings by the classical MDS and the Isomap method, respectively. Notice that for classical MDS, D is first computed using the pairwise distances between the 2D points x_i's, and the scaled first eigenvector of U is used to produce the computed τ_i's. If the τ_i^*'s are perfectly recovered, i.e., $\tau_i = s\tau_i^*, i = 1, \ldots, N$ with $s = \pm 1$, we should see a straight line in the figures of the second row. We see that for the straight line example in the left panel of Fig. 1, the classical MDS can recover the underlying 1D parameterization, but it fails for the nonlinear curve in the middle panel. However, Isomap can still recover the 1D parameterization for the same nonlinear curve as shown in the right panel of Fig. 1.

Isomap was proposed as a general technique for nonlinear dimension reduction (i.e., uncovering the natural parameter space of a nonlinear manifold): the pairwise Euclidean distance $d(x_i, x_j)$ in classical MDS is replaced by the geodesic distance between x_i and x_j on the manifold (defined as the length of the shortest path between two points on the manifold) (Tenenbaum et al., 2000). In a sense, it can be considered as a special case of the general MDS where dissimilarities are taken to be the geodesic distances. Isomap consists of the following steps: (1) a so-called neighborhood graph G of the data points x_i's is constructed with an edge connecting points x_i and x_j if x_j is one of the k nearest neighbors of x_i. Notice that the number of nearest neighbors k is a parameter of the algorithm that needs to be pre-set. One can also choose an ε-neighborhood, i.e., considering x_i and x_j as connected if $\|x_i - x_j\|_2 \leqslant \varepsilon$ (see Tenenbaum et al., 2000 for details). The edge of a connected pair (x_i, x_j) is then assigned a weight $\|x_i - x_j\|_2$, the Euclidean distance between x_i and x_j; (2) the geodesic distance between any two points x_i and x_j is then approximated by the shortest path within the weighted graph G, call it $\hat{d}(x_i, x_j)$; (3) the classical MDS is then applied to the squared-geodesic-distance matrix $\hat{D} = [\hat{d}^2(x_i, x_j)]_{i,j=1}^N$. As we can see that the key difference between the classical MDS and Isomap is that in
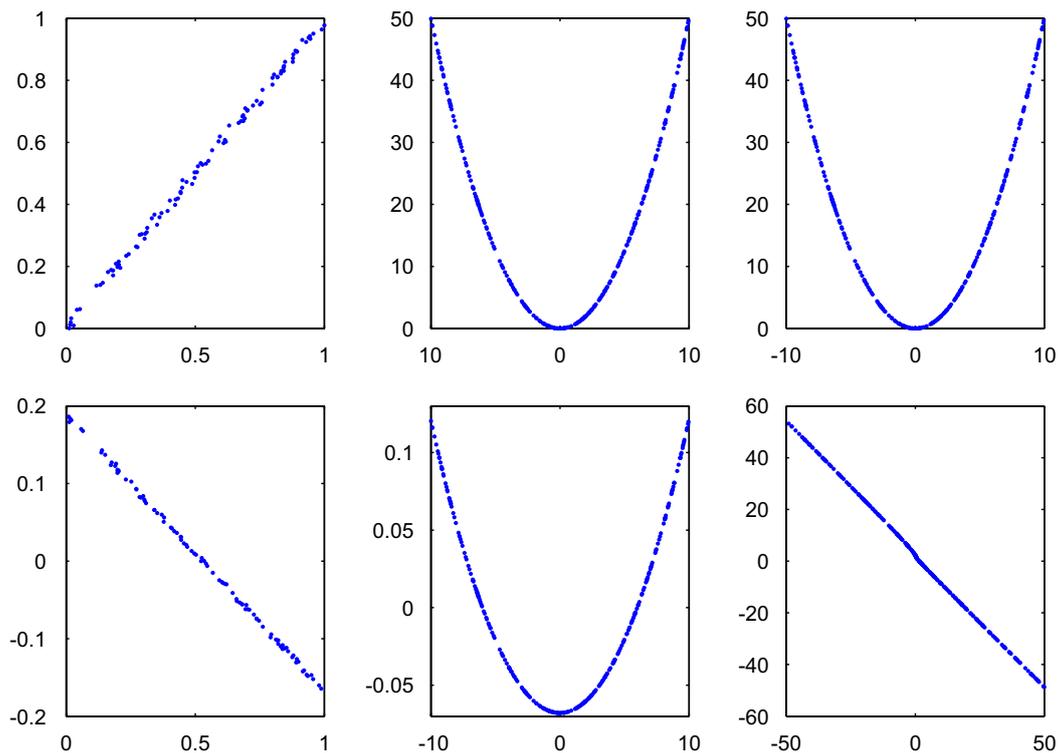
Fig. 1. 2D point sets and their 1D embeddings. Left: Points on a straight line and embeddings using MDS. Middle: Points on a parabola and embeddings using MDS. Right: Points on a parabola and embeddings using Isomap.

the classical MDS pairwise Euclidean distance is used while inIsomap it is the pairwise geodesic distance. Empirical success of Isomap for discovering nonlinear structures of high-dimensional data sets has been demonstrated by Donoho and Grimes (2002) and Tenenbaum et al. (2000).

From both a practical as well as a theoretical viewpoint, one is naturally led to the following questions: What is the low-dimensional nonlinear structure that Isomap tries to discover? And for what type of nonlinear manifolds, can Isomap perfectly recover the low-dimensional nonlinear structure? With issues such as discretization errors, sampling density of the data points and errors due to noise, it is not easy to answer those questions in a clear way using the current framework of Isomap based on a discrete set of data points. It is generally agreed, however, that reasoning in a continuous framework can sometimes crystalize the real issue of the problem and provide intuition for further development. This is the viewpoint taken in Donoho and Grimes (2002) and the same we will follow in this paper as well. Before we discuss Isomap in a continuous framework, we need to first introduce some basic notions from differential geometry.

3. Isometric embedding of manifolds

In this section, we recall several basic facts of differential geometry (do Carmo, 1976; O'Neill, 1997; Spivak, 1965, 1979). The general theory of manifold learning can be cast in the framework of Riemannian geometry, but to avoid unnecessary abstraction, we consider the special case of parameterized manifolds represented as hypersurfaces in Euclidean spaces (Munkres, 1990).

Definition. Let $d \leqslant n$, and Ω open in $\mathscr{R}^d$. Let $f : \Omega \to \mathscr{R}^n$. The set $\mathscr{M} \equiv f(\Omega)$ together with the mapping f is called a *parameterized manifold* of dimension d.

Therefore, $\mathscr{M}$ is characterized by n functions of d variables. We now introduce the concept of a *tangent map* on a manifold.

Definition. Let $f : \Omega \to \mathcal{M}$ be a mapping between two manifolds. The tangent map f_* of f assigns to each tangent vector v of Ω the tangent vector $f_*(v)$ of $\mathcal{M}$ such that if v is the initial velocity of a curve α in Ω, then $f_*(v)$ is the initial velocity of the image curve $f(\alpha)$ in $\mathcal{M}$.

When Ω is an open set of the d-dimensional Euclidean space $\mathcal{R}^d$ and $\mathcal{M}$ is embedded in $\mathcal{R}^m$, and assume that we can write f as

$$f(\tau) = \begin{bmatrix} f_1(\tau) \\ \vdots \\ f_m(\tau) \end{bmatrix} \quad \text{then } J_f(\tau) = \begin{bmatrix} \partial f_1/\partial \tau_1 & \cdots & \partial f_1/\partial \tau_d \\ \vdots & \vdots & \vdots \\ \partial f_m/\partial \tau_1 & \cdots & \partial f_m/\partial \tau_d \end{bmatrix}$$

gives the Jacobian matrix of f, and the tangent map $f_*(v)$ is simply $f_*(v) = J_f v$, here $\tau = [\tau_1, \ldots, \tau_d]^{\mathrm{T}}$.

Definition. The mapping $f : \Omega \to \mathcal{M}$ is an isometry if f is one-to-one and onto and f preserves inner products in the tangent spaces, i.e., for the tangent map f_*,

$$f_*(v)^{\mathrm{T}} f_*(w) = v^{\mathrm{T}} w$$

for any two vectors v and w that are tangent to Ω.

In the case that Ω is an open set of $\mathcal{R}^d$, it is easy to see that f is an isometry if and only if f is one-to-one and onto, and the Jacobian matrix J_f is orthonormal, i.e., $J_f^{\mathrm{T}} J_f = I_d$. In general, $J_f^{\mathrm{T}} J_f$ is the so-called *first fundamental form* of f, also referred to as the *metric tensor* of f, it measures the infinitesimal distortion of distances and angles under f. The larger the deviation of $J_f^{\mathrm{T}} J_f$ from I_d, the more the metric quantities are distorted.

The geodesic distance between two points on a manifold is defined as the length of the shortest path between the two points in question. For an isometry f defined on an open *convex* set of $\mathcal{R}^d$, it is easy to show that the geodesic distance between two points $f(\tau_1)$ and $f(\tau_2)$ on $\mathcal{M}$ is given by

$$d(f(\tau_1), f(\tau_2)) = \|\tau_1 - \tau_2\|_2. \tag{2}$$

We next give two examples to illustrate the above concepts.

Example. First, we consider the set of scaled 2-by-2 rotations of the form

$$R(\theta) = \frac{1}{\sqrt{2}} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad \theta \in (0, 2\pi].$$

We embed $\{R(\theta)\}$ into R^4 by

$$R(\theta) \to f(\theta) = \frac{1}{\sqrt{2}} [\cos\theta, \sin\theta, -\sin\theta, \cos\theta]^{\mathrm{T}}.$$

It is easy to check that $\|J_f(\theta)\|_2 = 1$, and the geodesic distance between $R(\theta_1)$ and $R(\theta_2)$ is $|\theta_1 - \theta_2|$.

Remark. If $d = 1$ and $\mathcal{M}$ represents a *regular* curve, i.e., $f'(\tau) \neq 0$ for all $\tau \in \Omega$, we can always re-parameterize $\mathcal{M}$ by its arc-length s to obtain $g : s \to \mathcal{M}$ and $\|g'(s)\|_2 = 1$, i.e., g is an isometry.

Example. Next, we consider the 2D swiss-roll surface in 3D Euclidean space parameterized as for $u > 0$ (a different swiss-roll surface, however, was used in Roweis and Saul (2000) and Tenenbaum et al. (2000) and will be discussed in the next section)

$$f(u, v) = \left[\frac{1}{\sqrt{2}} u \cos(\log u), v, \frac{1}{\sqrt{2}} u \sin(\log u) \right]^{\mathrm{T}}. \tag{3}$$

It can be verified that $J_f(u, v)$ is orthonormal, i.e., $(J_f(u, v))^{\mathrm{T}} J_f(u, v) = I_2$, and the swiss-roll surface is isometric to $\{(u, v) \mid u > 0\}$.

4. Continuum Isomap

With the above preparation, we are now ready to present a continuous version of Isomap. Let $d(x, y)$ define the geodesic distance between two points x and y on the manifold $\mathcal{M}$. We first need the concept of integration with respect to volume over a parameterized Manifold.

Definition. Let $d \leqslant n$, and Ω open in $\mathcal{R}^d$. Let $f : \Omega \to \mathcal{R}^n$, and $\mathcal{M} \equiv f(\Omega)$. For a real function F defined on $\mathcal{M}$, the integral of F over $\mathcal{M}$ with respect to volume is defined as

$$\int_{\mathcal{M}} F(x)\, dx = \int_{\Omega} F(f(\tau)) h(\tau)\, d\tau \quad \text{where } h(\tau) = \sqrt{\det(J_f^{\mathrm{T}} J_f)}.$$

We now define a continuous version of the matrix B defined in (1) in the form of a continuous kernel $K(x, y)$ as follows (this can be considered as the case when the sample points are uniformly concentrated on $\mathcal{M}$, in fact, $K(x, y)$ actually corresponds to the form of B with X not centered in (1))

$$K(x, y) = \frac{1}{2 \int_{\mathcal{M}} dz} \int_{\mathcal{M}} (d^2(x, z) + d^2(z, y) - d^2(x, y))\, dz$$
$$- \frac{1}{2(\int_{\mathcal{M}} dz)^2} \int_{\mathcal{M} \times \mathcal{M}} d^2(u, v)\, du\, dv.$$

We will restrict ourselves to the case: $f : \Omega \to \mathcal{M}$, and $\Omega \subset R^d$ is an open *convex* subset. Consequences of non-convexity of Ω have been discussed in Bernstien et al. (2000) and Donoho and Grimes (2002) and will also be mentioned at the end of this section.

Define for short and with an abuse of notation,

$$d_f(s, t) \equiv d(f(s), f(t)), \quad K_f(s, t) \equiv K(f(s), f(t)),$$

then the kernel is represented as

$$K_f(s, t) = \frac{1}{2 \int_{\Omega} h(\tau)\, d\tau} \int_{\Omega} (d_f^2(s, \tau) + d_f^2(\tau, t) - d_f^2(s, t)) h(\tau)\, d\tau$$
$$- \frac{1}{2(\int_{\Omega} h(\tau)\, d\tau)^2} \int_{\Omega \times \Omega} d_f^2(\tau, \hat{\tau}) h(\tau) h(\hat{\tau})\, d\tau\, d\hat{\tau}.$$

More generally, we can also consider data points sampled from an arbitrary density function ρ concentrated on Ω to obtain

$$K_f(s, t) = \frac{1}{2 \int_{\Omega} H(\tau)\, d\tau} \int_{\Omega} (d_f^2(s, \tau) + d_f^2(\tau, t) - d_f^2(s, t)) H(\tau)\, d\tau$$
$$- \frac{1}{2(\int_{\Omega} H(\tau)\, d\tau)^2} \int_{\Omega \times \Omega} d_f^2(\tau, \hat{\tau}) H(\tau) H(\hat{\tau})\, d\tau\, d\hat{\tau},$$

where $H(\tau) = \rho(\tau) h(\tau)$.

Parallel to the development in the classical MDS, we consider the eigenvalue problem of the integral operator with kernel K. Let ϕ be an eigenfunction of the kernel K, i.e.,

$$\int_{\mathcal{M}} K(x, y) \rho(y) \phi(y)\, dy = \lambda \phi(x), \quad x \in \mathcal{M}, \tag{4}$$

or equivalently on Ω,

$$\int_{\Omega} K_f(s, t) H(t) \phi(f(t))\, dt = \lambda \phi(f(s)), \quad s \in \Omega. \tag{5}$$

It is not difficult to verify that $\phi(x)$ has zero mean, i.e.,

$$\int_{\mathscr{M}} \rho(x)\phi(x)\,\mathrm{d}x = \int_{\Omega} H(\tau)\phi(f(\tau))\,\mathrm{d}\tau = 0.$$

We now show that if f is an isometry, then the first d largest eigenfunctions form the canonical coordinates of Ω up to a rigid motion.

Theorem 1. *Let* $f : \Omega \subset \mathscr{R}^d \to \mathscr{M}$ *be an isometry, i.e.,* $d_f(s,t) = \|s - t\|_2$ *with* Ω *open and convex. Let* $c = \int_{\Omega} \tau H(\tau)\,\mathrm{d}\tau / \int_{\Omega} H(\tau)\,\mathrm{d}\tau$ *be the mean vector of* Ω. *Assume that* $\phi_1(x), \ldots, \phi_d(x)$ *are the d orthogonal eigenfunctions of the kernel* $K(x, y)$ *corresponding to the d largest eigenvalues* $\lambda_j, \ j = 1, \ldots, d$,

$$\int_{\mathscr{M}} \rho(x)\phi_i(x)\phi_j(x)\,\mathrm{d}x = \lambda_i \delta_{ij} = \begin{cases} \lambda_i, & i = j, \\ 0, & i \neq j. \end{cases} \tag{6}$$

Then the vector function $\phi(\tau) \equiv [\phi_1, \ldots, \phi_d]^{\mathrm{T}}$ *equals to* $\tau - c$ *up to an orthogonal transformation, i.e., there is a constant orthogonal matrix* $P \in \mathscr{R}^{d \times d}$ *such that* $\phi(\tau) = P(\tau - c)$.

Furthermore, $\lambda_j, \ j = 1, \ldots, d$ *are the eigenvalues and P is the eigenvector matrix of the d-by-d symmetric positive definite matrix*

$$A \equiv \int_{\Omega} (\tau - c)H(\tau)(\tau - c)^{\mathrm{T}}\,\mathrm{d}\tau.$$

Proof. With the assumption $d_f(s,t) = \|s - t\|_2$, we have

$$\begin{aligned}
K_f(s,t) &= \frac{1}{2\int_{\Omega} H(\tau)\,\mathrm{d}\tau} \int_{\Omega} (\|s - \tau\|^2 + \|\tau - t\|^2 - \|s - t\|^2)H(\tau)\,\mathrm{d}\tau \\
&\quad - \frac{1}{2(\int_{\Omega} H(\tau)\,\mathrm{d}\tau)^2} \int_{\Omega \times \Omega} \|\tau - \hat{\tau}\|^2 H(\tau)H(\hat{\tau})\,\mathrm{d}\tau\,\mathrm{d}\hat{\tau} \\
&= (s - c)^{\mathrm{T}}(t - c) - \frac{1}{\int_{\Omega} H(\tau)\,\mathrm{d}\tau}(s + t - 2c)^{\mathrm{T}} \int_{\Omega} (\tau - c)H(\tau)\,\mathrm{d}\tau \\
&\quad + \left\| \frac{1}{\int_{\Omega} H(\tau)\,\mathrm{d}\tau} \int_{\Omega} (\tau - c)H(\tau)\,\mathrm{d}\tau \right\|^2.
\end{aligned}$$

By the definition of c, we have $\int_{\Omega}(\tau - c)H(\tau)\,\mathrm{d}\tau = 0$. Therefore,

$$K_f(s,t) = (s - c)^{\mathrm{T}}(t - c).$$

Let $\phi_j, \ j = 1, \ldots, d$, be the d eigenfunctions corresponding to the largest d eigenvalues λ_j of the kernel K. Then by the definition (5) for ϕ_j,

$$\int_{\Omega} (s - c)^{\mathrm{T}}(t - c)H(t)\phi_j(f(t))\,\mathrm{d}t = \lambda_j \phi_j(f(s)).$$

Defining

$$p_j = \frac{1}{\lambda_j} \int_{\Omega} (\tau - c)H(\tau)\phi_j(f(\tau))\,\mathrm{d}\tau, \tag{7}$$

we have

$$\phi_j(x) = \phi_j(f(\tau)) = (\tau - c)^{\mathrm{T}} p_j. \tag{8}$$

Therefore

$$[\phi_1(x), \ldots, \phi_d(x)]^{\mathrm{T}} = [p_1, \ldots, p_d]^{\mathrm{T}}(\tau - c) \equiv P(\tau - c).$$

Substituting (8) into the normalization conditions (6) and using (7), we obtain that

$$\lambda_i \delta_{ij} = p_j^{\mathrm{T}} \int_{\Omega} (\tau - c) H(\tau) \phi_i(f(\tau)) \, \mathrm{d}\tau = \lambda_i p_j^{\mathrm{T}} p_i.$$

It clearly shows that P is orthogonal.

Finally, let us denote $A = \int_{\Omega} (\tau - c) H(\tau) (\tau - c)^{\mathrm{T}} \, \mathrm{d}\tau$. Substituting (8) into (7) gives

$$\lambda_j p_j = \int_{\Omega} (\tau - c) H(\tau) (\tau - c)^{\mathrm{T}} \, \mathrm{d}\tau \cdot p_j \equiv A p_j.$$

Therefore p_j is the eigenvector of A corresponding to the eigenvalue λ_j. $\quad\square$

If $f(\tau)$ and $\hat{f}(\hat{\tau})$ are two different parameterizations of the same manifold $\mathcal{M}$ and both $f(\tau)$ and $\hat{f}(\hat{\tau})$ are isometries, then clearly τ and $\hat{\tau}$ only differ by a rigid motion. Now suppose η is a different parameterization of $\mathcal{M}$ and is related to τ by $\tau = \tau(\eta)$. What is the ϕ function computed by the continuum Isomap in terms of η? The following corollary answers this question.

Corollary 2. *Let $f : \tau \in \Omega \to \mathcal{M}$ be a parameterization expression of $\mathcal{M}$ and not necessarily isometric. If there exists one-to-one mapping $\tau = \tau(\eta)$ such that $f \circ \tau : \eta \in \Gamma \to \mathcal{M}$ is isometric, where $\eta = \eta(\tau)$ is the inverse mapping of $\tau(\eta)$, then the vector function $\phi \equiv [\phi_1, \dots, \phi_d]^{\mathrm{T}}$ computed by the continuum Isomap is given by*

$$\phi = P(\eta(\tau) - c).$$

As an application, we consider a special case. Assume that f is not isometric for its parameter variable τ and has the following property,

$$J_f(\tau)^{\mathrm{T}} J_f(\tau) = \mathrm{diag}(\rho_1(\tau_1), \dots, \rho_d(\tau_k))$$

with all positive ρ_i's. It is easy to verify that there exist d one-variable functions $\eta_i = \eta_i(\tau_i)$, $i = 1, \dots, d$, such that $\eta_i'(\tau_i) = \rho_i^{1/2}(\tau_i)$. Since the ρ_i's are positive, the η_i are strictly monotonically increasing, and therefore the inverse mapping $\tau = \tau(\eta)$ is one-to-one and $\tau_i'(\eta_i) = \rho_i^{-1/2}(\tau_i(\eta_i))$. This gives that

$$J_\tau(\eta) = [\partial \tau_i / \partial \eta_j]_{i,j=1}^d = \mathrm{diag}(\rho_1^{-1/2}(\tau_1), \dots, \rho_d^{-1/2}(\tau_d)).$$

Now parameterizing the manifold as $f \circ \tau : \eta \to \mathcal{M}$, we have

$$J_{f \circ \tau}(\eta) = J_f(\tau) J_\tau(\eta),$$

and hence $J_{f \circ \tau}(\eta)^{\mathrm{T}} J_{f \circ \tau}(\eta) = I_d$, i.e., the manifold parameterized by η is an isometry. It follows from Corollary 2 that, up to a rigid motion, the ϕ function computed by the continuum Isomap has the form of $\eta(\tau)$. Notice that although the geodesic distances are not preserved in the τ space, the deformation only occurs along the individual τ_i directions.

Now we go back to the swiss-roll surface defined by

$$f(u, v) = [u \cos u, v, u \sin u]^{\mathrm{T}}.$$

It is easy to see that $J_f(u, v)^{\mathrm{T}} J_f(u, v) = \mathrm{diag}(1 + u^2, 1)$. Hence f is not an isometry for the given parameterization. However, with the variable transformation

$$w = \int_0^u \sqrt{1 + t^2} \, \mathrm{d}t = \frac{1}{2}(u\sqrt{1 + u^2} + \mathrm{arcsinh}\,(u)).$$

Denoting by $u = u(w)$ the inverse transformation, the swiss-roll surface can be parameterized as

$$\hat{f}(w, v) = [u(w) \cos u(w), v, u(w) \sin u(w)]^{\mathrm{T}}$$
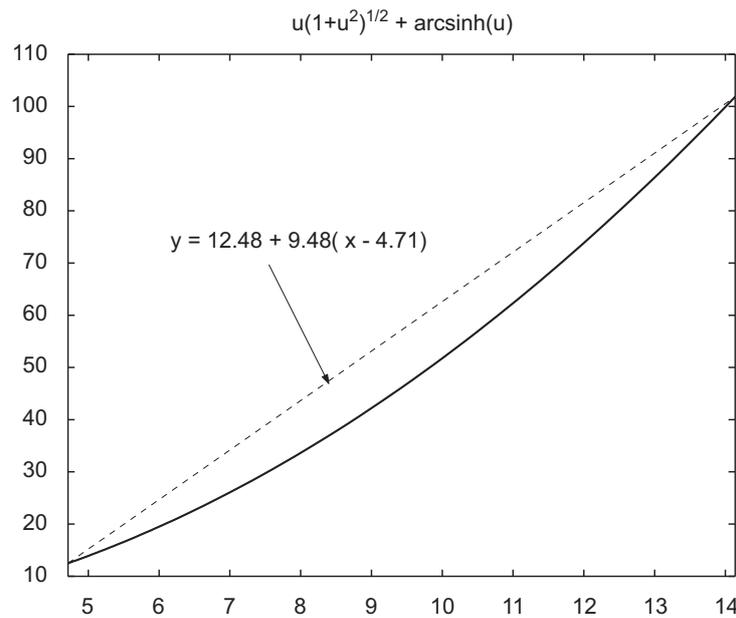
$$u(1+u^2)^{1/2} + \text{arcsinh}(u)$$

Fig. 2. Deformation function for the swiss-roll surface.

and $\hat{f}$ is isometric for $(w, v)^{\mathrm{T}}$, and up to a rigid motion, the ϕ function computed by the continuum Isomap has the form of $(w, v)^{\mathrm{T}}$, i.e.,

$$[\tfrac{1}{2}(u\sqrt{1+u^2} + \text{arcsinh}(u)), v]^{\mathrm{T}}.$$

Hence no deformation (stretching and compressing) occurs in the v direction, but there is certain deformation in the u direction.

However, in Roweis and Saul (2000) and Tenenbaum et al. (2000), a 2D swiss-roll surface embedded in 3D space is parameterized as $f(u, v) = [u\cos u, v, u\sin u]^{\mathrm{T}}$. For a data set sampled uniformly from the interval $[\tfrac{3}{2}\pi, \tfrac{9}{2}\pi]$ along the u-direction, the computed ϕ-coordinates by Isomap seem to be the original sample points (u_i, v_i). How to explain this phenomenon? In fact, the retrieved should be (w_i, v_i) with a rigid motion, where

$$w_i = \tfrac{1}{2}(u_i\sqrt{1+u_i^2} + \text{arcsinh}(u_i)).$$

However, within this interval $[\tfrac{3}{2}\pi, \tfrac{9}{2}\pi]$, the function $\tfrac{1}{2}(u\sqrt{1+u^2} + \text{arcsinh}(u))$ is very close to a straight line as is illustrated in Fig. 2, i.e., $w_i \approx p(u_i + c)$, $i = 1, \ldots, N$, for constants p and c. This explains why the points computed by Isomap seem to be uniformly distributed in a square. In Section 5, we will show a perturbation result for non-isometric mappings.

4.1. Convexity condition

Recall that we have assumed that Ω is a *convex* open set. The convexity is crucial for the continuum Isomap to work correctly, this was clearly pointed out in Bernstien et al. (2000) and Donoho and Grimes (2002). The reason is also quite simple, if there is a hole in the manifold, the geodesic curve needs to move *around* the hole and the relationship $d_f(s, \tau) = \|s - \tau\|_2$ will no longer hold even if $J_f(\tau)^{\mathrm{T}}J_f(\tau) = I_d$ still holds true. This is actually a drawback of methods such as Isomap that depend on *global* pairwise distances. As we have mentioned before, geodesic distance is a global property of a manifold while isometry is defined locally, i.e., property of the tangent spaces at each point. Proportionality of geodesic distances to Euclidean distances in the parameter space is a consequence of isometry. In the non-convex case, however, isometry can still hold but proportionality of geodesic distances to Euclidean distances will fail to be true. A global method such as Isomap can no longer handle this case and a local method is called for. In fact, if you roll a piece of paper into a swiss-roll shape, you can flatten it back without regard whether the shape of
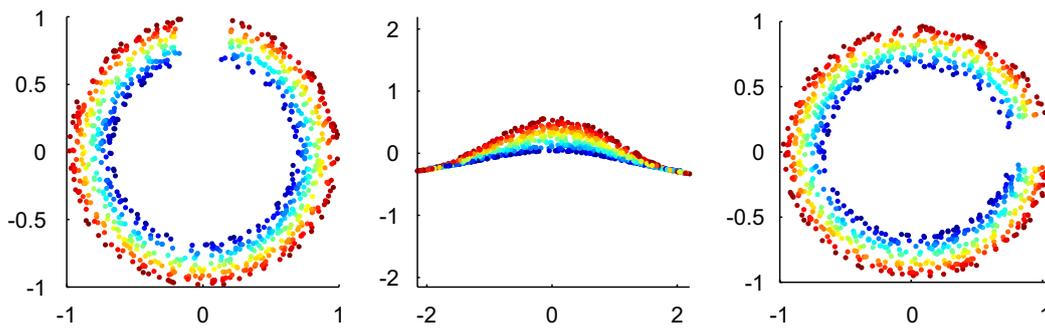
Fig. 3. Broken ring data set: (left) The original data set, (middle) reconstruction using Isomap, (right) reconstruction using orthogonal LTSA.

the piece of paper is convex or not. Local methods such as the local tangent space alignment (LTSA) method proposed in Zhang and Zha (2004) can still perfectly recover the low-dimensional structure as is illustrated in Fig. 3, where the original data form a broken ring which is clearly non-convex, Isomap fails to recover the original coordinates while LTSA does very well.

Remark. We note that if $\mathcal{M}$ is not isometric to a flat space, then the number of nonzero eigenvalues of the integral operator with kernel $K(x, y)$ defined at the beginning of the section will be *infinite*. If we select a finite number of the eigenfunctions, we cannot expect them to fully represent the low-dimensional structure of the given nonlinear manifold, certain information has been lost going from infinite to finite.

4.2. Connection with discrete Isomap

In the definition of the kernel $K(x, y)$ (4) if we assume that the density is concentrated on $\mathcal{M}$, i.e., $\int_{\mathcal{M}} \rho(\tau) \, d\tau = 1$, we can write

$$K(x, y) = \frac{1}{2} \int_{\mathcal{M}} (d^2(x, z) + d^2(z, y) - d^2(x, y)) \rho(z) \, dz$$
$$- \frac{1}{2} \int_{\mathcal{M} \times \mathcal{M}} d^2(u, v) \rho(u) \rho(v) \, du \, dv. \tag{9}$$

Now for a given set of sample points $x_1, \ldots, x_N$, let ρ be the corresponding empirical density

$$\rho(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x; x_i), \quad \delta(x; x_i) = \begin{cases} 1, & x = x_i, \\ 0, & x \neq x_i. \end{cases}$$

Here $\delta(x; x_i)$ is a Kronecker function, giving

$$\int_{\mathcal{M}} f(x) \delta(x; x_i) \, dx = \begin{cases} f(x_i), & x_i \in \mathcal{M}, \\ 0, & x_i \notin \mathcal{M}. \end{cases}$$

Plug in the above ρ into the expression for $K(x, y)$, we obtain

$$2K(x, y) = \frac{1}{N} \sum_{i=1}^{N} d^2(x, x_i) + \frac{1}{N} \sum_{i=1}^{N} d^2(x_i, y) - d^2(x, y) + \frac{1}{N^2} \sum_{i,j=1}^{N} d^2(x_i, x_j).$$

Now let $D = [d^2(x_i, x_j)]_{i,j=1}^{N}$ and $K = [K(x_i, x_j)]_{i,j=1}^{N}$, it can be readily checked that with $J = I - (1/N)ee^{\mathrm{T}}$,

$$K = -\frac{1}{2} J D J.$$

Therefore, with the empirical density the continuum Isomap reduces to the original discrete Isomap.

5. Perturbation analysis

In some applications, the parameterized manifold we are interested in is not necessarily associated with an isometry. We want to analyze the eigenfunctions computed by the continuum Isomap in this non-isometry case. To this end, let $\phi_j(x)$, $j = 1, \ldots, d$, be the d eigenfunctions corresponding to the largest d eigenvalues λ_j of the kernel $K(x, y)$, i.e.,

$$\int_{\mathscr{M}} K(x, y)\rho(y)\phi_j(y)\,\mathrm{d}y = \lambda_j \phi_j(x).$$

Recalling that $\int_{\mathscr{M}} \rho(x)\phi_j(x)\,\mathrm{d}x = 0$, we have

$$\lambda\phi_j(x) = c_j - \frac{1}{2}\int_{\mathscr{M}} d^2(x, y)\phi_j(y)\rho(y)\,\mathrm{d}y, \tag{10}$$

where c_j is a constant defined by

$$c_j = \frac{1}{2\int_{\mathscr{M}} \rho(x)\,\mathrm{d}x}\int_{\mathscr{M}}\int_{\mathscr{M}} d^2(x, y)\rho(x)\rho(y)\phi_j(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Theorem 3. *Let $f : \Omega \subset \mathscr{R}^d \to \mathscr{M} \subset \mathscr{R}^n$ be a continuous map from an open and convex set Ω to $\mathscr{M}$. Assume that the geodesic distance $d(x, y)$ can be written as*

$$d_f^2(\tau, \hat{\tau}) = \alpha\|\tau - \hat{\tau}\|_2^2 + \eta(\tau, \hat{\tau}), \tag{11}$$

where α is a constant and $\eta(\tau, \hat{\tau})$ defines deviation from isometry. Then there are constant vectors $c = \int_{\Omega}\tau H(\tau)\,\mathrm{d}\tau / \int_{\Omega} H(\tau)\,\mathrm{d}\tau$ and p_j such that

$$\phi_j(x) = p_j^{\mathrm{T}}(\tau - c) + e_j(\tau),$$

where $e_j(\tau) = \varepsilon_j^{(0)} - \varepsilon_j(\tau)$ has zero mean, $\int_{\Omega} H(\tau)e_j(\tau)\,\mathrm{d}\tau = 0$, with

$$\varepsilon_j(\tau) = \frac{1}{2\lambda_j}\int_{\Omega}\eta(\tau, \hat{\tau})H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau}, \quad \varepsilon_j^{(0)} = \frac{1}{\int_{\Omega} H(\tau)\,\mathrm{d}\tau}\int_{\Omega}\varepsilon_j(\tau)H(\tau)\,\mathrm{d}\tau.$$

Proof. Substituting (11) into (10), we obtain

$$\lambda\phi_j(x) = c_j - \frac{\alpha}{2}\int_{\Omega}\|\tau - \hat{\tau}\|_2^2 H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau} - \frac{1}{2}\int_{\Omega}\eta(\tau, \hat{\tau})H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau}.$$

Because $\int_{\Omega} H(\tau)\phi_j(f(\tau))\,\mathrm{d}\tau = \int_{\mathscr{M}}\rho(x)\phi_j(x)\,\mathrm{d}x = 0$, we have

$$\int_{\Omega}\|\tau - \hat{\tau}\|_2^2 H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau} = \int_{\Omega}(\|\hat{\tau} - c\|_2^2 - 2(\tau - c)^{\mathrm{T}}(\hat{\tau} - c))H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau}$$

$$= \frac{2}{\alpha}(\tilde{c}_j - \lambda p_j^{\mathrm{T}}(\tau - c)),$$

where

$$\tilde{c}_j = \frac{\alpha}{2}\int_{\Omega}\|\tau - c\|_2^2 H(\tau)\phi_j(f(\tau))\,\mathrm{d}\tau, \quad p_j = \frac{\alpha}{\lambda}\int_{\Omega}\tau H(\tau)\phi_j(f(\tau))\,\mathrm{d}\tau.$$

Therefore

$$\lambda\phi_j(x) = c_j - \tilde{c}_j + \lambda p_j^{\mathrm{T}}(\tau - c) - \frac{1}{2}\int_{\Omega}\eta(\tau, \hat{\tau})H(\hat{\tau})\phi_j(f(\hat{\tau}))\,\mathrm{d}\hat{\tau}. \tag{12}$$

On the other hand, it is not difficult to verify that

$$c_j - \tilde{c}_j = \frac{1}{2 \int_\Omega H(\tau) \, d\tau} \int_{\Omega \times \Omega} \eta(\tau, \hat{\tau}) H(\tau) H(\hat{\tau}) \phi_j(f(\hat{\tau})) \, d\tau \, d\hat{\tau}. \tag{13}$$

The result required follows immediately by substituting (13) into (12). $\quad \square$

Sometimes, it will be convenient to write the deviation from isometry in terms of a bi-Lipschitz condition. The following corollary of the above theorem handles this case.

Corollary 4. *If there are constants* $\alpha_2 \geqslant \alpha_1 > 0$ *such that the geodesic distance function* $d(x, y)$ *satisfies*

$$\alpha_1 \|\tau - \hat{\tau}\|_2^2 \leqslant d^2(f(\tau), f(\hat{\tau})) \leqslant \alpha_2 \|\tau - \hat{\tau}\|_2^2,$$

then $\phi_j(x) = p_j^T(\tau - c) + \varepsilon_j^{(0)} - \varepsilon_j(\tau)$ *and*

$$|\varepsilon_j(\tau)| \leqslant \frac{\alpha_2 - \alpha_1}{4\lambda_j} \int_\Omega \|\tau - \hat{\tau}\|_2^2 H(\hat{\tau}) |\phi(f(\hat{\tau}))| \, d\hat{\tau}.$$

Proof. Denote $\alpha = (\alpha_1 + \alpha_2)/2$, then

$$d_f^2(\tau, \hat{\tau}) = \alpha \|\tau - \hat{\tau}\|_2^2 + \eta(\tau, \hat{\tau}), \tag{14}$$

with $|\eta(\tau, \hat{\tau})| \leqslant ((\alpha_2 - \alpha_1)/2) \|\tau - \hat{\tau}\|_2^2$. By Theorem 3 and the upper bound of $|\eta(\tau, \hat{\tau})|$ above, we obtain that

$$|\varepsilon_j(\tau)| = \left| \frac{1}{2\lambda_j} \int_\Omega \eta(\tau, \hat{\tau}) H(\hat{\tau}) \phi(f(\hat{\tau})) \, d\hat{\tau} \right| \leqslant \frac{\alpha_2 - \alpha_1}{4\lambda_j} \int_\Omega \|\tau - \hat{\tau}\|_2^2 H(\hat{\tau}) |\phi(f(\hat{\tau}))| \, d\hat{\tau},$$

completing the proof. $\quad \square$

Example. We now use a concrete example to illustrate the above perturbation analysis. In particular, we consider the semi-sphere $\mathcal{M}$,

$$f(s, t) = \begin{pmatrix} s \\ t \\ \sqrt{1 - s^2 - t^2} \end{pmatrix}, \quad s^2 + t^2 \leqslant r^2 < 1.$$

For $f(s, t), f(\tilde{s}, \tilde{t}) \in \mathcal{M}$, it is easy to verify that

$$d = d(f(s, t), f(\tilde{s}, \tilde{t})) = \arccos(f(s, t)^T f(\tilde{s}, \tilde{t}))$$

and

$$\left\| \begin{pmatrix} s \\ t \end{pmatrix} - \begin{pmatrix} \tilde{s} \\ \tilde{t} \end{pmatrix} \right\|^2 = 2(1 - \cos(d)) - (z - \tilde{z})^2 = d^2 - 2\left(\frac{d^4}{4!} - \frac{d^6}{6!} + \cdots\right) - (z - \tilde{z})^2,$$

where $z = \sqrt{1 - s^2 - t^2}$, $\tilde{z} = \sqrt{1 - \tilde{s}^2 - \tilde{t}^2}$. Therefore,

$$d^2(f(s, t), f(\tilde{s}, \tilde{t})) = \left\| \begin{pmatrix} s \\ t \end{pmatrix} - \begin{pmatrix} \tilde{s} \\ \tilde{t} \end{pmatrix} \right\|^2 + 2\left(\frac{d^4}{4!} - \frac{d^6}{6!} + \cdots\right) + (z - \tilde{z})^2. \tag{15}$$

On the other hand,

$$(z - \tilde{z})^2 \leqslant \frac{(s + \tilde{s})^2 + (t + \tilde{t})^2}{(z + \tilde{z})^2} \left\| \begin{pmatrix} s \\ t \end{pmatrix} - \begin{pmatrix} \tilde{s} \\ \tilde{t} \end{pmatrix} \right\|^2 \leqslant \frac{r^2}{1 - r^2} \left\| \begin{pmatrix} s \\ t \end{pmatrix} - \begin{pmatrix} \tilde{s} \\ \tilde{t} \end{pmatrix} \right\|^2.$$
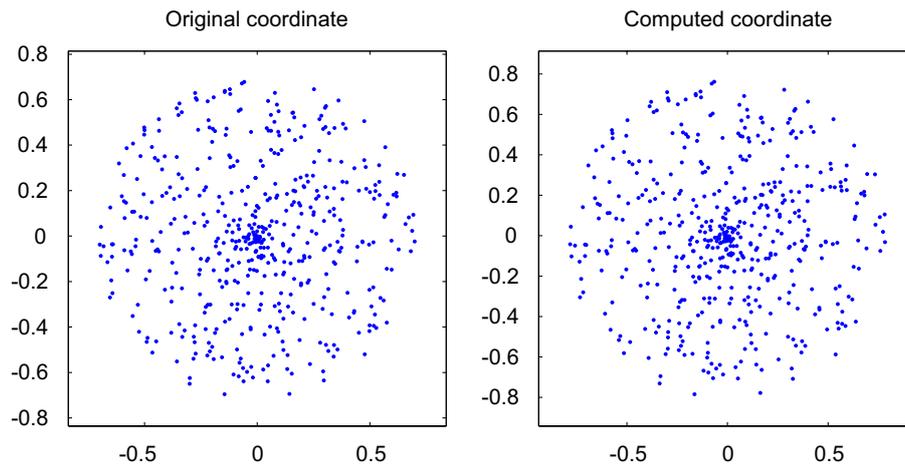
Fig. 4. 2D point sets (left) and the computed coordinates by Isomap (right).

Substituting it into (15) yields

$$d^2 \leqslant \frac{1}{1-r^2} \left\| \binom{s}{t} - \binom{\tilde{s}}{\tilde{t}} \right\|^2 + \frac{d^4}{12}.$$

With the constraint $d < \pi$, $s^2 + t^2 \leqslant r^2$ and $\tilde{s}^2 + \tilde{t}^2 \leqslant r^2$, we obtain that

$$\left\| \binom{s}{t} - \binom{\tilde{s}}{\tilde{t}} \right\|^2 \leqslant d^2 \leqslant \alpha \left\| \binom{s}{t} - \binom{\tilde{s}}{\tilde{t}} \right\|^2,$$

where $\alpha = 2/(1-r^2)(1 + \sqrt{1 - 2/(3(1-r^2))r^2})$. For small r, $\alpha \approx 1$.

In Fig. 4, we plot the original 2D coordinates $\tau_i^* = (s_i, t_i)^{\mathrm{T}}$ of $n = 500$ 3D points x_i sampled from the semi-sphere as following (using the notation of MATLAB):

```
phi   = (pi/4) * rand(1,n);
theta = (2*pi) * rand(1:n);
s     = sin(phi).*cos(theta);   s = s-mean(s);
t     = sin(phi).*sin(theta);   t = t-mean(t);
z     = cos(phi);
```

For this example, $\alpha = 1.2679$. Let $\{\tau_i\}$ be the computed coordinates using Isomap. Under a rigid motion ($\{\tau_i^*\}$ and $\{\tau_i\}$ have been centered) $\hat{\tau}_i = Q\tau_i$, where the orthogonal matrix Q minimizes $\|[\alpha^*, \tau_n^*] - Q[\alpha, \tau_n]\|_2$, we compute the errors $\varepsilon_i = \|\tau_i^* - \hat{\tau}_i\|_2$. Fig. 5 plots the absolute errors $\{\varepsilon_i\}$ and the relative errors $\{\varepsilon_i/\|\tau_i^*\|_2\}$.

6. Low-dimensional embedding for out-of-sample data points

For a given set of sample points $x_1, \ldots, x_N$, the discrete Isomap provides the low-dimensional embedding for each x_i, $i = 1, \ldots, N$, in terms of the largest d eigenvectors of the N-by-N matrix K. In many applications, a problem of great practical interest is how to map an out-of-sample data point, i.e., for a point x sampled from the manifold but not necessarily one of the x_i's, how to find its low-dimensional embedding?

In the definition of the kernel $K(x, y)$ (4), if we assume that the density is concentrated on $\mathcal{M}$, i.e., $\int_{\mathcal{M}} \rho(x) \, \mathrm{d}x = 1$, we can write (10) for $x \in \mathcal{M}$ as

$$\lambda \phi_j(x) = \frac{1}{2} \int_{\mathcal{M}} \left(\int_{\mathcal{M}} d^2(z, y)\rho(z) \, \mathrm{d}z - d^2(x, y) \right) \rho(y)\phi_j(y) \, \mathrm{d}y. \tag{16}$$
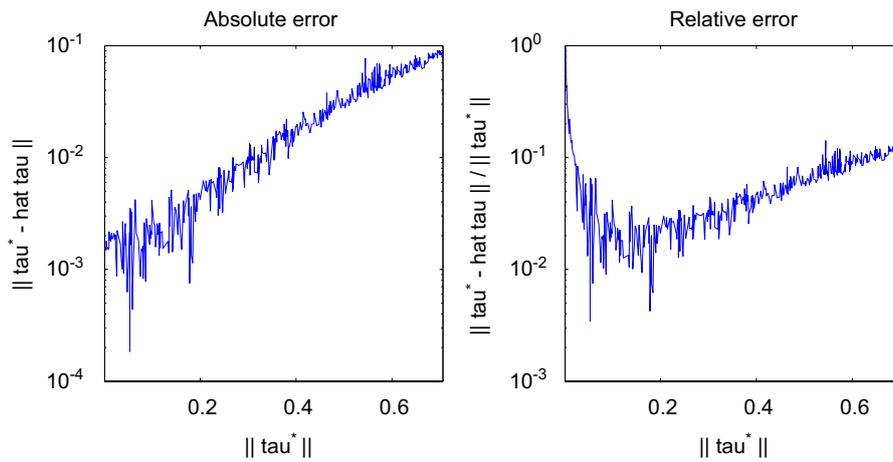
Fig. 5. Error curves with respect to $\|\tau^*\|_2$.

Now for a given set of sample points $x_1, \ldots, x_N$, let ρ be the corresponding empirical density $\rho(x) = (1/N)\sum_{i=1}^{N} \delta(x - x_i)$ as before. Plugging in the above ρ into expression (16), we obtain that for $x \in \mathcal{M}$

$$\phi_j(x) = \frac{1}{2\lambda_j N} \sum_{i=1}^{N} \left(\frac{1}{N} \sum_{k=1}^{N} d^2(x_k, x_i) - d^2(x, x_i) \right) \phi_j(x_i). \tag{17}$$

As we have shown, if f is isometry from Ω to $\mathcal{M}$, then low-dimensional embedding for the given data point x is given by $\phi \equiv [\phi_1(x), \ldots, \phi_d(x)]^{\mathrm{T}}$ up to a rigid motion.

Remark. The above process is actually based on a very simple linear algebra result: let $A = U\Sigma V^{\mathrm{T}} \in \mathcal{R}^{m \times n}$ with $m \geqslant n$ be the singular value decomposition of A with

$$U = [u_1, \ldots, u_m], \quad V = [v_1, \ldots, v_n], \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n),$$

then $Av_i = \sigma_i u_i$, $A^{\mathrm{T}} u_i = \sigma_i v_i$. So if we know v_i, then u_i can be recovered as $u_i = Av_i/\sigma_i$. Similarly, $v_i = A^{\mathrm{T}} u_i/\sigma_i$. Eq. (17) can be obtained by extending one-side of the N-by-N matrix K to infinity, and use the above trick to recover the infinite-dimensional singular vectors from the finite dimensional ones.

So far we considered an out-of-sample point that lies on the manifold $\mathcal{M}$, we show how we can extend the definition in a continuous fashion to points lying close to $\mathcal{M}$. To this end, let $x \notin \mathcal{M}$, lying close to $\mathcal{M}$. Denote by $\hat{x} \in \mathcal{M}$ a point such that $\hat{x} = \mathrm{argmin}_{y \in \mathcal{M}} \|x - y\|_2$. If we define the quasi-manifold distance between x and a point $y \in \mathcal{M}$ by

$$d(x, y) = \|x - \hat{x}\|_2 + d(\hat{x}, y),$$

then we can extend the definition of ϕ_j in (17) for $x \notin \mathcal{M}$ using the distance defined above. It is easy to verify that

$$\phi_j(x) = \phi_j(\hat{x}) - \frac{1}{2\lambda_j N} \sum_{i=1}^{N} (\|x - \hat{x}\|_2^2 + 2\|x - \hat{x}\|_2 d(\hat{x}, x_i))\phi_j(x_i)$$

$$= \phi_j(\hat{x}) - \frac{\|x - \hat{x}\|_2}{\lambda_j N} \sum_{i=1}^{N} d(\hat{x}, x_i)\phi_j(x_i), \tag{18}$$

since $\sum_{t=1}^{N} \phi_j(x_t) = 0$. It can be readily checked that the ϕ_j thus defined is continuous at each $x \in \mathcal{M}$, but may not be smooth.

However, for x that has two or more nearest points $x_{i_\ell} \in \mathcal{M}$ such that $\|x - x_{i_1}\|_2 = \|x - x_{i_2}\|_2$, for example, the value of $\phi_j(x)$ defined by (18) is not uniquely determined. A better way is to define $\phi_j(x)$ using x's nearest neighbors located on the manifold. Let $x_{i_\ell} \in \mathcal{M}, \ell = 1, \ldots, k$ be the k-nearest-neighbors of x and U forms an orthogonal basis
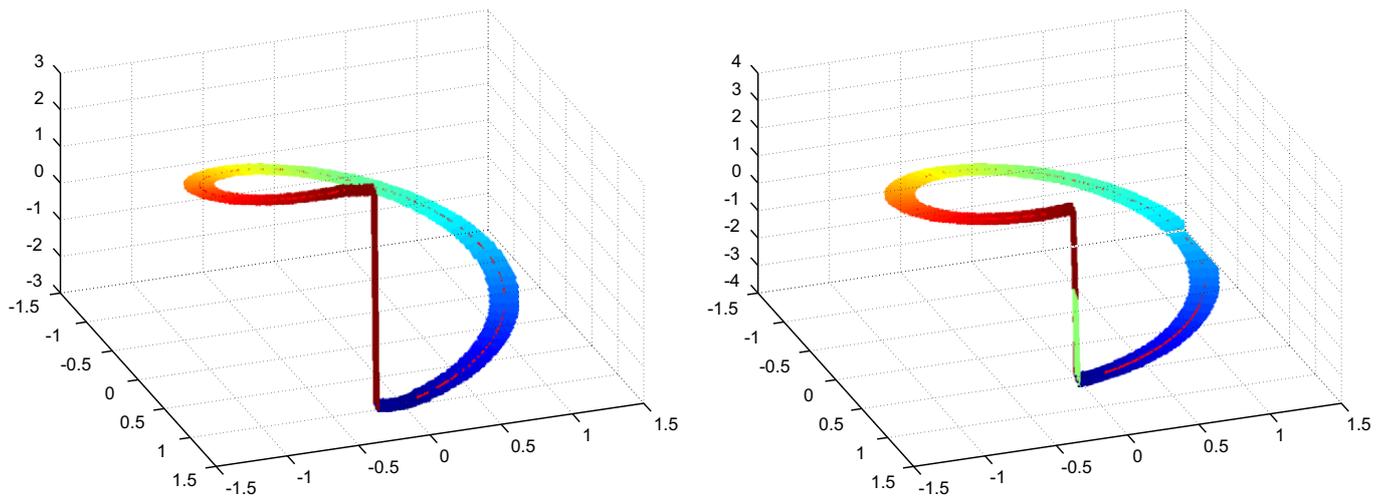
Fig. 6. The curves of ϕ's defined by (18) (left) and (20) (right).

for the linear subspace spanned by $x_{i_1}, \ldots, x_{i_k}$. Assume that t and t_ℓ are, respectively, the local coordinates of x and x_{i_ℓ} with respect to span$\{U\}$,

$$x = \bar{x} + Ut + v, \quad x_{i_\ell} = \bar{x} + Ut_\ell + \xi_\ell,$$

where $\bar{x} = (1/k)\sum_\ell x_{i_\ell}$ is the mean of the k neighbors, v is orthogonal to span$\{U\}$ and $\xi_\ell, \ell = 1, \ldots, k$, are the residual errors. Let $\alpha_1, \ldots, \alpha_k$ minimize $\|t - \sum_\ell \alpha_\ell t_\ell\|$ subject to the constraint $\alpha_1 + \cdots + \alpha_k = 1$. We can write $t = \sum_\ell \alpha_\ell t_\ell + \xi = t_\alpha + \xi$. It is easy to verify that

$$x = \sum_{\ell=1}^{k} \alpha_\ell x_{i_\ell} + v + U\xi + \sum_{\ell=1}^{k} \alpha_\ell \xi_\ell. \tag{19}$$

Then we can re-define $\phi_j(x)$ as follows:

$$\phi_j(x) = \sum_{\ell=1}^{k} \alpha_\ell \phi_j(x_{i_\ell}) - \frac{\|x - \sum_{\ell=1}^{k} \alpha_\ell x_{i_\ell}\|_2}{\lambda_j N} \sum_{i=1}^{N} \left(\sum_{\ell=1}^{k} \alpha_\ell d(x_{i_\ell}, x_i) \right) \phi_j(x_i). \tag{20}$$

We summarize the above in the following theorem.

Theorem 5. *Assume that $f : \Omega \subset \mathcal{R}^d \to \mathcal{M}$ is an isometry. Let $d(x, \mathcal{M})$ be the distance of x to the manifold $\mathcal{M}$. Denote $\tau_\alpha = \sum_{\ell=1}^{k} \alpha_\ell \tau_{i_\ell}$, $x_\alpha = f(\tau_\alpha)$, and $\eta = \|\xi\| + \sum_{\ell=1}^{k} \|\xi_\ell\|$. Then*

$$|\phi_j(x) - \phi_j(x_\alpha)| \leqslant \frac{d(x, \mathcal{M}) + \eta}{\lambda_j N} \left| \sum_{i=1}^{N} \left(\sum_{\ell=1}^{k} \alpha_\ell d(x_{i_\ell}, x_i) \right) \phi_j(x_i) \right|.$$

Proof. The proof is simple. In fact, by Theorem 1, $\phi_j(x_{i_\ell}) = p_j(\tau_{i_\ell} - c)$. We have that

$$\sum_{\ell=1}^{k} \alpha_\ell \phi_j(x_{i_\ell}) = p_j(\tau_\alpha - c) = \phi_j(x_\alpha).$$

Note that $\|v\| = d(x, \mathcal{M})$. Therefore, the result of the theorem follows from (20) immediately. □

In Fig. 6, we plot the ϕ's defined by (18) and (20), $k = 2$, with x close to the 2D manifold parameterized as $f(\tau) = [\cos(\tau), \sin(\tau)]^T$ with $\tau \in [0, \frac{15}{16}\pi]$. Note that for x close to two ends of the manifold corresponding to $\tau = 0$
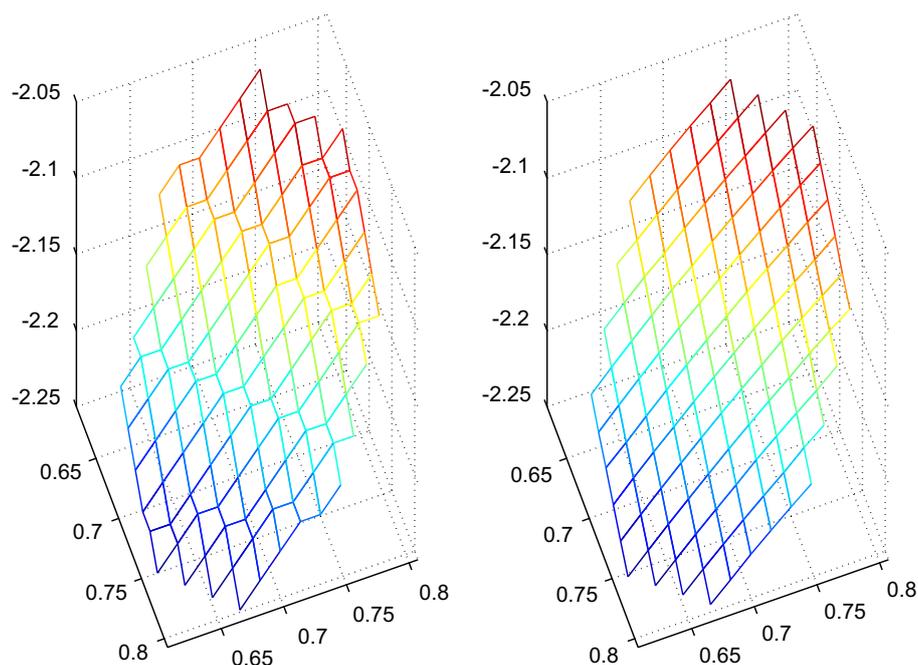
Fig. 7. The Zoom of ϕ's nearby $(\sqrt{2}/2, \sqrt{2}/2)^{\mathrm{T}}$: the left for (18) and the right for (20).

or $\tau = \frac{15}{16}\pi$, the ϕ defined by (18) has much distortions, while the ϕ defined by (20) seems to have smooth stretch. Furthermore, the latter seems to be much smoother. Two zoomed parts for x nearby $(\sqrt{2}/2, \sqrt{2}/2)^{\mathrm{T}}$ for both ϕ's are plotted in Fig. 7.

7. Conclusions

Isomap is a generalization of the classical multi-dimension scaling method for nonlinear dimension reduction. We proposed a continuous version of the Isomap method and showed that for a nonlinear manifold that can be isometrically embedded onto an Euclidean space, the continuum Isomap computes a set of eigenfunctions that forms the canonical coordinates of the Euclidean space up to a rigid motion. This answers the questions of what the low-dimensional nonlinear structure is that Isomap tries to discover and when it can perfectly discover it. We further show that the continuum Isomap also provides a natural way to compute low-dimensional embedding for out-of-sample data points, and derive some bounds for the case when the isometry condition is violated. One line of future research follows our perturbation analysis, in order to better understand Isomap in the non-isometric case, we need to investigate the issue of *optimal embedding*, i.e., dimensionality reduction will result in certain amount of geometric distortion, but we seek to minimize this distortion under certain criterion.

Acknowledgments

References

Bernstien, M., de Silva, V., Langford, J., Tenenbaum, J., 2000. Graph approximations to geodesics on embedded manifolds. Technical Report, Department of Psychology, Stanford University.
Cox, T., Cox, M., 2001. Multidimensional Scaling. Chapman & Hall, London.
do Carmo, M., 1976. Differential Geometry of Curves and Surfaces. Prentice-Hall, Englewood Cliffs, NJ.

Donoho, D., Grimes, C., 2002. When does Isomap recover the natural parametrization of families of articulated images? Technical Report 2002-27, Department of Statistics, Stanford University.

Gower, J.C., 1966. Some distance properties of latent root and vector methods in multivariate analysis. Biometrika 53, 315–328.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.

Munkres, J., 1990. Analysis on Manifold. Addison-Wesley, Redwood City, CA.

O'Neill, B., 1997. Elementary Differential Geometry. second ed. Academic Press, San Diego.

Roweis, S., Saul, L., 2000. Nonlinear dimension reduction by locally linear embedding. Science 290, 2323–2326.

Schoenberg, I.J., 1935. Remarks to Maurice Fréchet's article Sur la dénition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert. Ann. Math. 38, 724–732.

de Silva, V., Tenenbaum, J.B., 2002. Unsupervised learning of curved manifolds. In: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, New York.

Spivak, M., 1965. Calculus on Manifolds. Addison-Wesley, Redwood City.

Spivak, M., 1979. A Comprehensive Introduction to Differential Geometry. second ed. Publish or Perish, Boston.

Tenenbaum, J., De Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimension reduction. Science 290, 2319–2323.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419.

Williams, C., 2002. On a connection between kernel PCA and metric multidimensional scaling. Mach. Learn. 46, 11–19.

Young, G., Householder, A.S., 1938. Discussion of a set of points in terms of their mutual distances. Psychometrika 3, 19–22.

Zha, H., Zhang, Z., 2003. Isometric embedding and continuum Isomap. In: Proceedings of the 20th International Conference on Machine Learning (ICML-2003), pp. 864–871.

Zhang, Z., Zha, H., 2004. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J. Sci. Comput. 26, 313–338.