Journal of Machine Learning Research 14 (2013) 2945-2977 Submitted 8/11; Revised 4/12; Published 10/13

Parallel Vector Field Embedding

Binbin Lin BINBINLIN@ZJU.EDU.CN
Xiaofei He XIAOFEIHE@CAD.ZJU.EDU.CN
Chiyuan Zhang PLUSKID@GMAIL.COM
State Key Lab of CAD&CG
College of Computer Science
Zhejiang University
Hangzhou, 310058, China

Ming Ji MINGJI1@ILLINOIS.EDU
Department of Computer Science
University of Illinois at Urbana Champaign
Urbana, IL 61801, USA

Editor: Mikhail Belkin

Abstract

We propose a novel local isometry based dimensionality reduction method from the perspective of vector fields, which is called parallel vector field embedding (PFE). We first give a discussion on local isometry and global isometry to show the intrinsic connection between parallel vector fields and isometry. The problem of finding an isometry turns out to be equivalent to finding orthonormal parallel vector fields on the data manifold. Therefore, we first find orthonormal parallel vector fields by solving a variational problem on the manifold. Then each embedding function can be obtained by requiring its gradient field to be as close to the corresponding parallel vector field as possible. Theoretical results show that our method can precisely recover the manifold if it is isometric to a connected open subset of Euclidean space. Both synthetic and real data examples demonstrate the effectiveness of our method even if there is heavy noise and high curvature.

Keywords: manifold learning, isometry, vector field, covariant derivative, out-of-sample extension

1. Introduction

In many data analysis tasks, one is often confronted with very high dimensional data. There is a strong intuition that the data may have a lower dimensional intrinsic representation. Various researchers have considered the case when the data is sampled from a submanifold embedded in much higher dimensional Euclidean space. Consequently, estimating and extracting the low dimensional manifold structure, or specifically the intrinsic topological and geometrical properties of the data manifold, become a crucial problem. These problems are often referred to as *manifold learning* (Belkin and Niyogi, 2007).

The most natural technique to exact low dimensional manifold structure with given finite samples is dimensionality reduction. The early work for dimensionality reduction includes principal component analysis (PCA, Jolliffe, 1989), multidimensional scaling (MDS, Cox and Cox, 1994) and linear discriminant analysis (LDA, Duda et al., 2000). PCA is probably the most popular dimensionality reduction methods. Given a data set, PCA finds the directions along which the data

has maximum variance. However, these linear methods may fail to recover the intrinsic manifold structure when the data manifold is not a low dimensional subspace or an affine manifold.

There are various works on nonlinear dimensionality reduction in the last decade. The typical work includes isomap (Tenenbaum et al., 2000), locally linear embedding (LLE, Roweis and Saul, 2000), Laplacian eigenmaps (LE, Belkin and Niyogi, 2001), Hessian eigenmaps (HLLE, Donoho and Grimes, 2003) and diffusion maps (Coifman and Lafon, 2006; Lafon and Lee, 2006; Nadler et al., 2006). Isomap generalizes MDS to the nonlinear manifold case which tries to preserve pairwise geodesic distances on the data manifold. Diffusion maps tries to preserve another meaningful distance, that is, diffusion distance on the manifold. Isomap is an instance of global isometry based dimensionality reduction techniques, which tries to preserve the distance function or the metric of the manifold globally. One limitation of Isomap is that it requires the manifold to be geodesically convex. HLLE is based on local isometry criterion, which successfully overcomes this problem. Laplacian operator and Hessian operator are two of the most important differential operators in manifold learning. Intuitively, Laplacian measures the smoothness of the functions, while Hessian measures how a function changes the metric of the manifold. However, the Laplacian based methods like LLE and LE mainly focus on the smoothness of the embedding function, which may not be an isometry. The major difficulty in Hessian based methods is that they have to estimate the second order derivative of embedding functions, and consequently they have strong requirement on data samples.

One natural nonlinear extension of PCA is kernel principal component analysis (kernel PCA, Schölkopf et al., 1998). Interestingly, Ham et al. (2004) showed that Isomap, LLE and LE are all special cases of kernel PCA with specific kernels. Recently, Maximum Variance Unfolding (MVU, Weinberger et al., 2004) is proposed to learn a kernel matrix that preserves pairwise distances on the manifold. MVU can be thought of as an instance of local isometry with additional consideration that the distances between two points that are not neighbors are maximized.

Tangent space based methods have also received considerable interest recently, such as local tangent space alignment (LTSA, Zhang and Zha, 2004), manifold charting (Brand, 2003), Riemannian Manifold Learning (RML, Lin and Zha, 2008) and locally smooth manifold learning (LSML, Dollár et al., 2007). These methods try to find coordinates representation for curved manifolds. LTSA tries to construct a global coordinate via local tangent space alignment. Manifold charting has a similar strategy, which tries to expand the manifold by splicing local charts. RML uses normal coordinate to unfold the manifold, which aims to preserve the metric of the manifold. LSML tries to learn smooth tangent spaces of the manifold by proposing a smoothness regularization term of tangent spaces. Vector diffusion maps (VDM, Singer and Wu, 2011) is a much recent work which considers the tangent spaces structure of the manifold to define and preserve the vector diffusion distance.

In this paper, we propose a novel dimensionality reduction method, called parallel vector field embedding (PFE), from the perspective of vector fields. The theory of vector fields is a basic tool for discovering the geometry and topology of the manifold. We first give a discussion on local isometry and global isometry to show the intrinsic connection between parallel vector fields and isometry. The problem of finding an isometry turns out to be equivalent to finding orthonormal parallel vector fields on the data manifold. Therefore, we first find orthonormal parallel vector fields by minimizing the covariant derivative of a vector field. We then find an embedding function whose gradient field is as close to the parallel field as possible. In this way, the obtained embedding function would vary linearly along the geodesics of the manifold. Naturally, the corresponding embedding consisted

of embedding functions preserves the metric of the manifold. As pointed out by Goldberg et al. (2008), almost all spectral methods including LLE, LE, LTSA and HLLE use global normalization for embedding, which sacrifices isometry. PFE overcomes this problem by normalizing vector fields locally. Our theoretical study shows that, if the manifold is isometric to a connected open subset of Euclidean space, our method can faithfully recover the metric structure of the manifold.

The organization of the paper is as follows: In the next section, we provide a description of the dimensionality reduction problem from the perspectives of isometry and vector fields. In Section 3, we introduce our proposed Parallel Field Embedding algorithm. The extensive experimental results on both synthetic and real data sets are presented in Section 4. Finally, we provide some concluding remarks and suggestions for future work in Section 5.

2. Dimensionality Reduction from Geometric Perspective

Let $(\mathcal{M}, g)$ be a d-dimensional Riemannian manifold embedded in a much higher dimensional Euclidean space $\mathbb{R}^m$, where g is a Riemannian metric on $\mathcal{M}$. A Riemannian metric is a Euclidean inner product g_p on each of the tangent space $T_p\mathcal{M}$, where p is a point on the manifold $\mathcal{M}$. In addition we assume that g_p varies smoothly (Petersen, 1998). This means that for any two smooth vector fields X, Y the inner product $g_p(X_p, Y_p)$ should be a smooth function of p. The subscript p will be suppressed when it is not needed. Thus we might write $g(X,Y)$ or $g_p(X,Y)$ with the understanding that this is to be evaluated at each p where X and Y are defined. Generally we use the induced metric for $\mathcal{M}$. That is, the inner product defined in the tangent space of $\mathcal{M}$ is the same as that in the ambient space $\mathbb{R}^m$, that is, $g(u,v) = \langle u, v\rangle$ where $\langle\cdot,\cdot\rangle$ denote the canonical inner product in $\mathbb{R}^m$. In the problem of dimensionality reduction, one tries to find a smooth map: $F : \mathcal{M} \to \mathbb{R}^d$, which preserves the topological and geometrical properties of $\mathcal{M}$.

However, for some kinds of manifolds, it is impossible to preserve all the geometrical and topological properties. For example, consider a two-dimensional sphere, there is no such map that maps the sphere to a plane without breaking the topology of the sphere. Thus, there should be some assumptions of the data manifold. In this paper, we consider a relatively general assumption that the manifold $\mathcal{M}$ is diffeomorphic to an open subset of the Euclidean space $\mathbb{R}^d$. In other words, we assume that there exists a topology preserving map from $\mathcal{M}$ to $\mathbb{R}^d$.

Definition 1 (Diffeomorphism, Lee, 2003) *A diffeomorphism between manifolds $\mathcal{M}$ and $\mathcal{N}$ is a smooth map $F : \mathcal{M} \to \mathcal{N}$ that has a smooth inverse. We say $\mathcal{M}$ and $\mathcal{N}$ are diffeomorphic if there exists a diffeomorphism between them.*

For example, there is a diffeomorphism between a semi-sphere and a subset of $\mathbb{R}^2$. However, there is no diffeomorphism between a sphere and any subset of $\mathbb{R}^2$. In this paper, we only consider manifolds that are diffeomorphic to an open connected subset of Euclidean space like semi-sphere, swiss roll, swiss roll with hole, and so on.

2.1 Local Isometry and Global Isometry

With the assumption that the manifold is diffeomorphic to an open subset of $\mathbb{R}^d$, the goal of dimensionality reduction is to preserve the intrinsic geometry of the manifold as much as possible. Ideally, one hopes to preserve the metric of the manifold, or intuitively the pairwise distance between data points. This leads to the concept of isometry. Here we consider two kinds of isometry, that are,

local isometry and global isometry.[1] In the following we give the definitions and properties of local isometry and global isometry.

Definition 2 (Local Isometry, Lee, 2003) *Let $(\mathcal{M},g)$ and $(\mathcal{N},h)$ be two Riemannian manifolds, where g and h are metrics on them. For a map between manifolds $F:\mathcal{M}\to\mathcal{N}$, F is called local isometry if $h(dF_p(v),dF_p(v))=g(v,v)$ for all $p\in\mathcal{M}, v\in T_p\mathcal{M}$. Here dF is the differential of F.*

dF is also known as *pushforward* and denoted as F_* in many texts. For a fixed point $p\in\mathcal{M}$, dF is a linear map between $T_p\mathcal{M}$ and the corresponding tangent space $T_{F(p)}\mathcal{N}$. According to the definition, dF preserves the norm of tangent vectors. Moreover, we have the following theorem:

Theorem 1 (Petersen, 1998) *Let $F:\mathcal{M}\to\mathcal{N}$ be a local isometry, then*

1. *F maps geodesics to geodesics.*
2. *F is distance decreasing.*
3. *if F is also a bijection, then it is distance preserving.*

Intuitively, local isometry preserves the metric of the manifold locally. If the local isometry F is also a diffeomorphism, then it becomes global isometry.

Definition 3 (Global Isometry, Lee, 2003) *A map $F:\mathcal{M}\to\mathcal{N}$ is called global isometry between manifolds if it is a diffeomorphism and also a local isometry.*

If F is a global isometry, then its inverse F^{-1} is also a global isometry. We have the following proposition.

Proposition 1 *A global isometry preserves geodesics. If $F:\mathcal{M}\to\mathcal{N}$ is a global isometry, then for any two points $p,q\in\mathcal{M}$, we have $d(p,q)=d(F(p),F(q))$, where $d(\cdot,\cdot)$ denotes geodesic distance between two points.*

Proof For any two points $p,q\in\mathcal{M}$, we have $d(p,q)\geq d(F(p),F(q))$ according to the third statement of Theorem 1. Since $F^{-1}:\mathcal{N}\to\mathcal{M}$ is also a global isometry, we have $d(p,q)\leq d(F(p),F(q))$. Thus $d(p,q)=d(F(p),F(q))$. ■

Clearly, we hope the map F is a global isometry. This is because that, although local isometry maps geodesics to geodesics, the shortest geodesic between two points on $\mathcal{M}$ may not be the shortest geodesic on $\mathcal{N}$. Please see Figure 1 as an illustrative example. Clearly the map F is a local isometry. However, consider two points p and q on $\mathcal{M}$, we have $d(p,q)>d(F(p),F(q))$. Therefore F is not a global isometry.

It is usually very difficult to find a global isometry. Isomap is designed to find the global isometry. However, it is known that the computational cost is very expensive since pairwise distances have to be estimated. Also, it has been shown that Isomap cannot handle geodesically non-convex manifolds where in that case the geodesic distances cannot be accurately estimated. On the other hand, based on our assumption that the manifold $\mathcal{M}$ is diffeomorphic to an open subset of $\mathbb{R}^d$, it suffices to find a local isometry which is also a diffeomorphism, according to Definition 3.

1. In many differential geometry textbooks, global isometry is often referred to as *isometry* or *Riemannian isometry.*

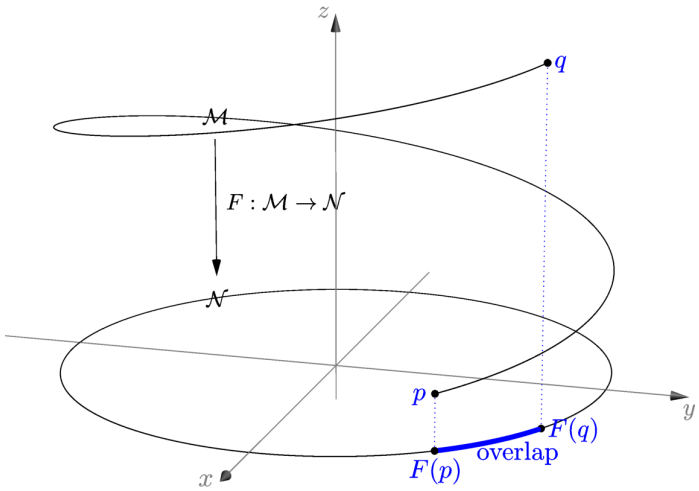


Figure 1: Local isometry but not global isometry. F is a map from $\mathcal{M}$ to $\mathcal{N}$. Clearly F is a local isometry. However, due to the overlap, it is not a global isometry.

2.2 Gradient Fields and Local Isometry

Our analysis has shown that finding a global isometry is equivalent to finding a local isometry which is also a diffeomorphism. Given a map $F = (f_1, \ldots, f_d) : \mathcal{M} \to \mathbb{R}^d$, there is a deep connection between local isometry and the differential $dF = (df_1, \ldots, df_d)$. For a function f on the manifold $f : \mathcal{M} \to \mathbb{R}$, we will not strictly distinguish between its *differential* df and its *gradient field* ∇f in this paper. Actually, they are dual 1-form which is uniquely determined by each other once the metric of the manifold is given. For the relationship between local isometry and differential, we have the following proposition:

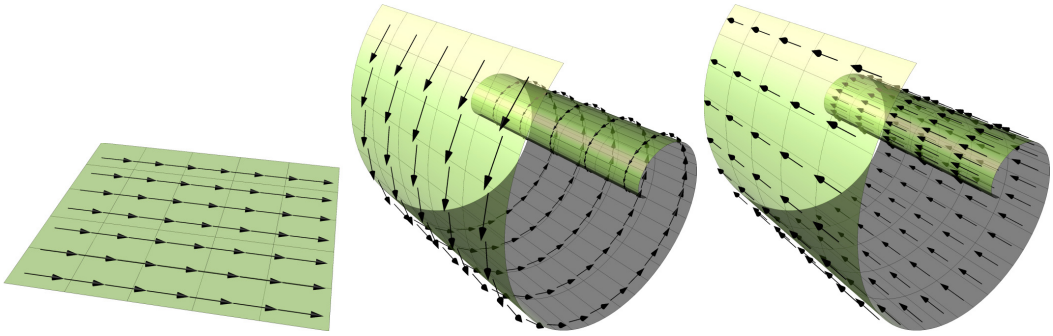
Proposition 2 *Consider a map $F : \mathcal{M} \subset \mathbb{R}^m \to \mathbb{R}^d$. Let $f_i, i = 1, \ldots, d$ denote the component of F which maps the manifold to $\mathbb{R}$, that is, $F = (f_1, \ldots, f_d)$. The following three statements are equivalent:*

1. *F is a local isometry.*
2. *dF_p is an orthogonal transformation for all $p \in \mathcal{M}$.*
3. *$\langle df_i, df_j \rangle_p = \delta_{ij}, i, j = 1, ..., d, \forall p \in \mathcal{M}$.*

Proof $2 \Leftrightarrow 3$ is trivial by the definition of orthogonal transformation. $2 \Rightarrow 1$ is obvious. Next we prove $1 \Rightarrow 2$. Since we use the induced metric for the manifold $\mathcal{M}$, the computation of inner product in tangent space is the same as the standard inner product in Euclidean space. We have $g(u, v) = \langle u, v \rangle, \forall u, v \in T_p\mathcal{M}$. According to Definition 2, we have $\langle u, u \rangle = \langle dF_p(u), dF_p(u) \rangle, \forall p \in \mathcal{M}, u \in T_p\mathcal{M}$. For arbitrary vectors u and $v \in T_p\mathcal{M}$, then we have

$$\langle dF_p(u+v), dF_p(u+v) \rangle = \langle u+v, u+v \rangle.$$

By expanding both side and notice that $\langle dF_p(u), dF_p(u) \rangle = \langle u, u \rangle$ and $\langle dF_p(v), dF_p(v) \rangle = \langle v, v \rangle$, we have $\langle dF_p(u), dF_p(v) \rangle = \langle u, v \rangle$ which implies that dF is an orthogonal transformation. ■



(a) $\mathbb{R}^2$ (b) Swiss Roll (c) Swiss Roll

Figure 2: Examples of parallel fields. The parallel fields on Euclidean space are all constant vector fields.

This proposition indicates that finding a local isometry F is equivalent to finding d orthonormal differentials $df_i, i = 1,...,d$, or gradient fields $\nabla f_i, i = 1,...,d$ since we have $\langle \nabla f_i, \nabla f_j \rangle = \langle df_i, df_j \rangle = \delta_{ij}$.

3. Parallel Field Embedding

In this section, we introduce our parallel field embedding (PFE) algorithm for dimensionality reduction.

Our goal is to find a map $F = (f_1, \ldots, f_d) : \mathcal{M} \subset \mathbb{R}^m \to \mathbb{R}^d$ which preserves the metric of the manifold. According to Proposition 2, its differential (or gradient fields) should be orthonormal. In the next subsection, we show that if such differential exists, then each df_i (or ∇f_i) has to be *parallel*, that is $\nabla df_i = 0$ (or $\nabla\nabla f_i = 0$). Naturally, we propose a vector field based method for solving this problem. We first try to find orthonormal parallel vector fields on the manifold. Then we try to reconstruct the map whose gradient fields can best approximate the parallel fields. In Theorem 2, we show that if the manifold can be isometrically embedded into the Euclidean space, then there exist orthonormal parallel fields and each parallel field is exactly a gradient field. Therefore, in such cases our proposed approach can find the optimal embedding which is a global isometry.

3.1 Parallel Vector Fields

In this subsection, we will show the properties of parallel fields. We will also discuss the relationship among local isometry, global isometry and parallel fields.

Definition 4 (Parallel Field, Petersen, 1998) *A vector field X on manifold $\mathcal{M}$ is a parallel vector field (or parallel field) if*

$$\nabla X \equiv 0,$$

where ∇ is the covariant derivative on the manifold $\mathcal{M}$.

Figure 2 shows some examples of parallel fields on Euclidean space and Swiss Roll. Given a point p on the manifold and a vector v_p on the tangent space $T_p\mathcal{M}$, then $\nabla_{v_p} X$ is a vector at point p which measures how the vector field X changes along the direction v_p at point p. Since p is arbitrary, given

a vector field Y, then $\nabla_Y X$ is a vector field which measures how the vector field X changes along the vector field Y on the manifold. Since $\nabla X : Y \mapsto \nabla_Y X$ is a map which maps a vector field Y to another vector field $\nabla_Y X$, $\nabla X \equiv 0$ also means that for any vector field Y on the manifold, we have $\nabla_Y X = 0$ and vice versa. For parallel fields, we also have the following proposition:

Proposition 3 *Let V and W be parallel fields on $\mathcal{M}$ associated with the metric g. We define a function $h : \mathcal{M} \to \mathbb{R}$ as follows:*

$$h(p) = g_p(V,W),$$

where g_p represents the inner product at p. Then $h(p) =$ constant, $\forall p \in \mathcal{M}$.

Proof Since V and W are parallel fields, we have $\nabla V = \nabla W = 0$ or $\nabla_Y V = \nabla_Y W = 0$ for any vector field Y.

We first show that a vector field is a derivation. For simplicity, let v_p be a tangent vector at point p on Euclidean space $\mathbb{R}^m$. Then v_p defines the directional derivative in the direction v_p at p as follows:

$$v_p f = D_v f(p) = \frac{d}{dt}|_{t=0} f(p+tv).$$

For tangent vectors on a general manifold, we define them as derivations which satisfies the Leibniz's rule, that is, $v_p(fg) = f(p)v_p g + g(p)v_p f$. If p varies, then all these v_ps constitute a vector field. Since for each point p, v_p is a derivation at p. Then the vector field is a derivation on the manifold. Let X be an arbitrary smooth vector field, then we apply it to the function h and we have

$$\begin{aligned} X(h) &= Xg(V,W) \\ &= g(\nabla_X V, W) + g(V, \nabla_X W) \\ &= 0 + 0 = 0. \end{aligned}$$

The second equation is due to the property of the covariant derivative. Since the above equation holds for arbitrary X, we have $h(p) = g_p(V,W) =$ constant. ■

Corollary 1 *Let V and W be parallel fields on $\mathcal{M}$ associated with the metric g, then $\int_{\mathcal{M}} g(V,W)dx = 0$ if and only if $\forall\ p \in \mathcal{M}$, $g_p(V,W) = 0$.*

Proof From Proposition 3, we see that $g(V,W) =$ constant. Thus, $\int_{\mathcal{M}} g(V,W)dx = \text{vol}(\mathcal{M})g(V,W)$. Since $\text{vol}(\mathcal{M}) > 0$, $\int_{\mathcal{M}} g(V,W)dx = 0$ if and only if $g(V,W) = 0$ or $\forall p \in \mathcal{M}$, $g_p(V,W) = 0$. ■

This corollary tells us if we want to check the orthogonality of the parallel fields at every point, it suffices to compute the integral of the inner product of the parallel fields. This is much more convenient for finding orthogonal parallel fields.

Also we have the following corollary:

Corollary 2 *Let V be a parallel vector field on $\mathcal{M}$, then $\forall p \in \mathcal{M}, \|V_p\| =$ constant where V_p represents the vector at p of the vector field V.*

Proof Let $W = V$ in Proposition 3, then $\forall p \in \mathcal{M}$, we have $\|V_p\|^2 = g_p(V,V) =$ constant. ■

Since every tangent vector of a parallel field has a constant length, we can perform normalization

of the parallel field simply as dividing every tangent vector of the parallel field by a same length. According to these results, finding orthonormal parallel fields becomes much easier: we first find orthogonal parallel fields on the manifold one by one by requiring that the integral of the inner product of two parallel fields is zero. We then normalize the vectors of parallel fields to be unit norm.

Before presenting our main result, we still need to introduce some concepts and properties on the relationship between isometry and parallel fields. We begin with the properties of the differential of a map. We have the following lemma.

Lemma 1 (Please see Lemma 3.5 in Lee, 2003) *Let $F : \mathcal{M} \to \mathcal{N}$ and let $p \in \mathcal{M}$.*

1. *$dF : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$ is linear.*
2. *If F is a diffeomorphism, then $dF : T_p\mathcal{M} \to T_{F(p)}\mathcal{N}$ is an isomorphism.*

This lemma shows that locally dF is an isomorphism if it is a diffeomorphism. Next, we show that a parallel field is uniquely determined locally.

Proposition 4 *Let $\mathcal{M}$ be an open connected manifold. For a given $p \in \mathcal{M}$, a parallel field X is uniquely determined by the vector X_p, where X_p denotes the value of the vector field X on the point p.*

Proof The equation $\nabla X \equiv 0$ is linear in X, so the space of parallel fields is a vector space. Therefore, it suffices to show that $X \equiv 0$ provided $X_p = 0$. According to Proposition. 3, a parallel field has constant length. Thus for any point $q \in \mathcal{M}$, we have $\|X_q\|^2 = \|X_p\|^2 = 0$. ■

Next we show that the differential of an isometry preserves covariant derivative.

Lemma 2 (Please see the exercise (2) in Chapter 3 of Petersen, 1998) *If $F : \mathcal{M} \to \mathcal{N}$ is a global isometry, then we have $dF(\nabla_X Y) = \nabla_{dF(X)} dF(Y)$ for all vector fields X and Y.*

More importantly, we show that an isometry preserves parallelism, that is, its differential carries a parallel vector field to another parallel vector field.

Proposition 5 *If $F : \mathcal{M} \to \mathcal{N}$ is a global isometry, then*

1. *dF maps parallel fields to parallel fields.*
2. *dF is an isometric isomorphism on the space of parallel fields.*

Proof Let Y be a parallel field on $\mathcal{M}$, we show that $dF(Y)$ is also a parallel field. It suffices to show that $\nabla_Z dF(Y) = 0$ for arbitrary Z on $\mathcal{N}$. According to Lemma 2, for any $p \in \mathcal{M}$ and any vector field X on $\mathcal{M}$, we have $\nabla_{dF(X)} dF(Y) = dF(\nabla_X Y) = 0$ hold at p. Since X is arbitrary and dF is an isomorphism (Lemma. 1) at p, then $dF(X)_p$ can be an arbitrary vector at p. Since p is also arbitrary, then all these tangent vectors $dF(X)_p$ constitute an arbitrary vector field. Thus $\nabla_Z dF(Y) = 0$ holds for arbitrary vector field Z which proves the first statement.

For the second statement, since parallel fields are determined locally (Proposition 4), we only have to show that dF is an isometrically isomorphism locally. Firstly, F is diffeomorphism. According to Lemma 1, dF is a local isomorphism. Secondly, according to the Definition 2, dF is a

local isometry. Combining these two facts, dF is an isometric isomorphism on the space of parallel fields. ■

Now we show that the gradient fields of a local isometry are also parallel fields.

Proposition 6 *If $F = (f_1, \ldots, f_d) : \mathcal{M} \to \mathbb{R}^d$ is a local isometry, then each df_i is parallel, that is, $\nabla df_i = 0$, $i = 1, \ldots, d$.*

Proof According to the property of local isometry, for very point $p \in \mathcal{M}$, there is a neighborhood $U \subset \mathcal{M}$ of p such that $F|_U : U \to F(U)$ is a global isometry of U onto an open subset $F(U)$ of $\mathbb{R}^d$ (please see Lee, 2003, pg. 187). Since $F(U)$ is an open subset of $\mathbb{R}^d$, we can choose an orthonormal basis $\partial_i, i = 1, \ldots, d$ for $F(U)$. Since $F^{-1}|_U$ is also a global isometry, thus $dF^{-1}(\partial_i)$ is an orthonormal basis of U. It can be seen as $\langle dF^{-1}(\partial_i), dF^{-1}(\partial_j)\rangle = \langle \partial_i, \partial_j \rangle = \delta_{ij}$. The first equation is due to the definition of isometry. Since $F|_U$ is a global isometry, $dF|_U$ is orthonormal with respect to these coordinates. Thus we can rewrite $F|_U$ as $F(x) = Ox + b$, where $x \in U$, O is an orthonormal matrix and $b \in \mathbb{R}^d$. Note that $dF = (df_1, \ldots, df_d)$, thus df_i has constant coefficients and we have $\nabla df_i = 0$ at each local neighborhood. Since we can choose arbitrary p, $\nabla df_i \equiv 0$ holds on the whole manifold. ■

This proposition tells us that the gradient field of a local isometry is also a parallel field. Since it is usually not easy to find a global isometry directly, in this paper, we try to find a set of orthonormal parallel fields first, and then find an embedding function whose gradient field is equal to the parallel field. Our main theorem will show that such an embedding is a global isometry.

3.2 Objective Function

As stated before, we first try to find vector fields which are as parallel as possible on the manifold. Let V be a smooth vector field on $\mathcal{M}$. By definition, the covariant derivative of V should be zero. That is, $\nabla V \equiv 0$. Naturally, we define our objective function as follows:

$$E(V) = \int_{\mathcal{M}} \|\nabla V\|_{\mathrm{HS}}^2 dx, \quad s.t., \int_{\mathcal{M}} \|V\|^2 = 1, \tag{1}$$

where $\|\cdot\|_{\mathrm{HS}}$ denotes Hilbert-Schmidt tensor norm (see Defant and Floret, 1993). The constraint removes arbitrary scale of the vector field. Once we obtain the first parallel vector field V_1, by using orthogonality constraint $\int_{\mathcal{M}} g(V_1, V_2) = 0$, we can find the second vector field V_2, and so on. After finding d orthogonal vector fields $V_1, \ldots, V_d$, we normalize the vector fields at each point:

$$\|V_i|_x\| = 1, \forall x \in \mathcal{M}, \tag{2}$$

where $V_i|_x \in T_x\mathcal{M}$ denote the vector at x of the vector field V_i.

$E(V)$ enforce the vector fields to be parallel. The norm of the tangent vector $\|V_i|_x\|$ represents the scale of the map at x, thus the normalization $\|V_i|_x\|$ removes the scale locally. This is quite different from traditional manifold learning algorithms. Traditional methods remove the scale by normalizing the norm of embedding function which is a global normalization. As pointed out by Goldberg et al. (2008), the embedding functions obtained by these traditional methods are not isometry. As we discussed in Section 2, the gradient fields of the isometry have to be orthonormal parallel fields. However, traditional manifold learning methods may not satisfy this requirement, which will be shown in our experiments.

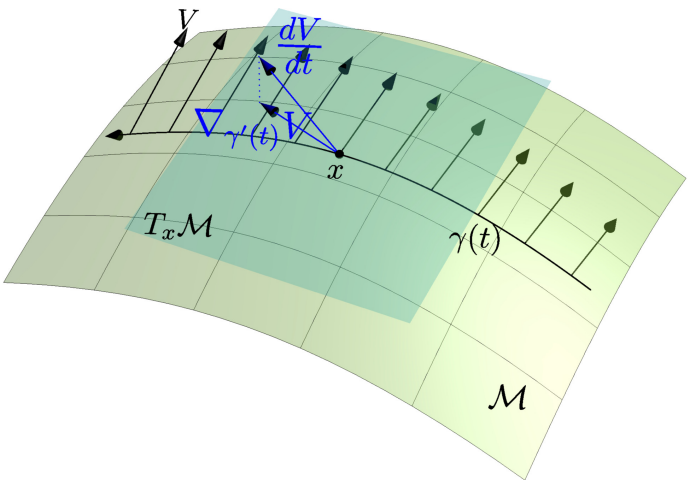


Figure 3: Covariant derivative demonstration. Let V, Y be two vector fields on manifold $\mathcal{M}$. Given a point $x \in \mathcal{M}$, let $\nabla_Y V|_x$ denote the tangent vector at x of the vector field $\nabla_Y V$, we show how to compute the vector $\nabla_Y V|_x$. Let $\gamma(t)$ be a curve on $\mathcal{M}$: $\gamma : I \to \mathcal{M}$ which satisfies $\gamma(0) = x$ and $\gamma'(0) = Y_x$. Then the covariant derivative along the direction $\frac{d\gamma(t)}{dt}|_{t=0}$ can be computed by projecting $\frac{dV}{dt}|_{t=0}$ to the tangent space $T_x\mathcal{M}$ at x. In other words, $\nabla_{\gamma'(0)} V|_x = P_x(\frac{dV}{dt}|_{t=0})$, where $P_x : v \in \mathbb{R}^m \to P_x(v) \in T_x\mathcal{M}$ is the projection matrix. It is not difficult to check that the computation of $\nabla_Y V|_x$ is independent to the choice of the curve γ.

In the following we provide some explanation of our objective function. ∇V is the covariant derivative of V that measures the change of the vector field V. If ∇V vanishes, V is a parallel vector field which we are looking for. Formally, ∇V is a $(1,1)$-tensor which maps a vector field Y to another vector field $\nabla_Y V$ and satisfies $\nabla_{\alpha Y} V = \alpha \nabla_Y V$ for any function α. For a fixed point $x \in \mathcal{M}$, let $\nabla V|_x$ denote the tensor value at x of the tensor field ∇V. It is a linear map on the tangent space $T_x\mathcal{M}$. We show what $\nabla V|_x$ is when given an orthonormal basis. Let $\partial_1, \ldots, \partial_d$ be an orthonormal basis of $T_x\mathcal{M}$, then the element of $\nabla V|_x$ would be $g_x(\nabla_{\partial_i} V, \partial_j)$ where $g_x(\cdot,\cdot)$ denote the inner product at point x. By the definition of Hilbert-Schmidt tensor norm (see Defant and Floret, 1993), we have

$$\begin{aligned} \|\nabla V|_x\|_{\mathrm{HS}}^2 &= \sum_{i=1}^{d}\sum_{j=1}^{d} (g_x(\nabla_{\partial_i} V, \partial_j))^2 \\ &= \sum_{i=1}^{d} g_x(\nabla_{\partial_i} V, \nabla_{\partial_i} V). \end{aligned} \tag{3}$$

The second equation uses the fact $g_x(\nabla_{\partial_i} V, \nabla_{\partial_i} V) = \sum_{j=1}^{d} (g_x(\nabla_{\partial_i} V, \partial_j))^2$. It is important to point out that the Hilbert-Schmidt norm $\|\nabla V\|_{\mathrm{HS}}$ is independent to the choice of the basis of tangent space. Thus our objective function $E(V)$ is well defined. According to the above equations, the computation of $\|\nabla V\|_{\mathrm{HS}}$ depends on the computation of the norm of covariant derivative $\nabla_{\partial_i} V$. Next we show the geometrical meaning of covariant derivative. For a given direction Y_x at $x \in \mathcal{M}$,

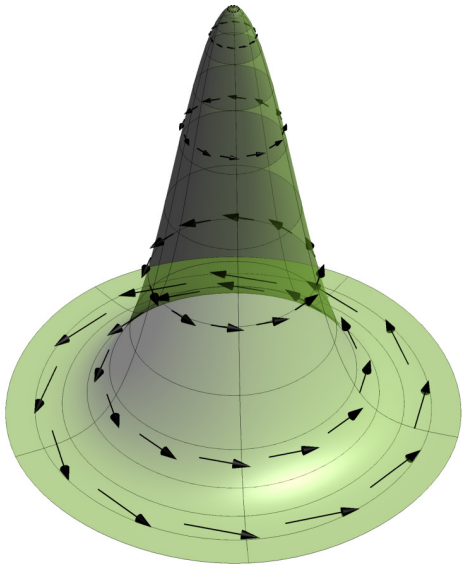


Figure 4: An example of a vector field but not a gradient field. This vector field has loops, thus it cannot be a gradient field for any function.

let $\nabla_Y V|_x$ denote the vector at x of vector field $\nabla_Y V$. Then $\nabla_Y V|_x$ is also a vector at x which is demonstrated in Figure 3.

After finding the parallel vector fields V_i on $\mathcal{M}$, the embedding function can be obtained by minimizing the following objective function:

$$\Phi(f) = \int_{\mathcal{M}} \|\nabla f - V\|^2 dx. \tag{4}$$

The solution of $\Phi(f)$ is not unique, but differs with a constant function.

In the following, we show that if the manifold is flat and diffeomorphic to an open connected subset of Euclidean space $\mathbb{R}^d$, then our method can successfully recover the metric of the manifold.

Theorem 2 *Let $\mathcal{M}$ be a d-dimensional Riemannian manifold embedded in $\mathbb{R}^m$. If there exist a global isometry $\varphi : \mathcal{M} \to D \subset \mathbb{R}^d$, where D is an open connected subset of $\mathbb{R}^d$, then there is an orthonormal basis $\{V_i\}_{i=1}^d$ of the parallel fields on $\mathcal{M}$, and embedding function $f_i : \mathcal{M} \to \mathbb{R}$ whose gradient field satisfies $\nabla f_i = V_i, i = 1, \ldots, d$. Moreover, $F = (f_1, \ldots, f_d)$ is a global isometry.*

Proof In Euclidean space, if a vector field is written in Cartesian coordinates, then it is parallel if and only if it has constant coefficients (Petersen, 1998). Consider a parallel field on D. Since D is open connected, this parallel field is globally constant. Thus the space of parallel fields on D is a d dimensional linear space.

According to Proposition 5, we know that a global isometry preserves parallelism, that is, its differential carries a parallel vector field to another parallel vector field. Thus for a global isometry φ, $d\varphi$ maps parallel fields to parallel fields and $d\varphi$ is an isometric isomorphism, so is $d\varphi^{-1}$. Thus the space of parallel fields on $\mathcal{M}$ is isomorphic to the space of parallel fields on D. Therefore there exists an orthonormal basis $\{V_i\}_{i=1}^d$ of the space of the parallel fields on $\mathcal{M}$. Next we show V_i is

also a gradient field. We first map V_i to D using $d\varphi$. Clearly, $d\varphi(V_i)$ is a parallel field. Note that a parallel field in Euclidean space has constant coefficients in Cartesian coordinates. We can write $d\varphi(V_i)$ in Cartesian coordinates $\partial_j, j = 1,\ldots,d$ as follows:

$$d\varphi(V_i) = \sum_j c_i^j \partial_j,$$

where c_i^j are constant. Since $d\varphi$ is an isometric isomorphism, we can rewrite it as follows:

$$V_i = \sum_j c_i^j d\varphi^{-1}(\partial_j).$$

Here $d\varphi^{-1}(\partial_j)$ actually is an orthonormal basis of $\mathcal{M}$. Since c_i^j are constant, each V_i is a gradient field for some linear function with respect to the coordinates of $d\varphi^{-1}(\partial_j)$. Let f_i be a linear function such that $\nabla f_i = V_i$, $i = 1,\ldots,d$. It is worth noting that such a linear function f_i is not unique but only differs a constant. Then these $\{f_i | i = 1,\ldots,d\}$ constitutes a map F, $F = (f_1,\ldots,f_d) : \mathcal{M} \to \mathbb{R}^d$, which maps the manifold $\mathcal{M}$ to $\mathbb{R}^d$.

Next we show that such F is a global isometry on $\mathcal{M}$. Firstly, F is a local isometry. According to the construction of F, the differential of F $dF = (df_1,\ldots,df_d) = (V_1,\ldots,V_d)$ is orthonormal. Thus F is a local isometry according to Proposition 2. Secondly, F is a differmorphism. Clearly the map F restricted on the manifold $\mathcal{M}$, $F : \mathcal{M} \to F(\mathcal{M})$, is surjective. Next we show F is also injective. If not, assume there exist two distinct points p and q such that $F(p) = F(q)$. Then we have $f_i(p) = f_i(q)$, $i = 1,\ldots,d$. Since f_i is linear with respect to the coordinates of $d\varphi^{-1}(\partial_j)$. We rewrite f_i as follows:

$$f_i(p) = \sum_{j=1}^{d} c_i^j z_j(p) + \varepsilon_i,$$

where $z_j, j = 1,\ldots,d$ represent coordinate functions. Since c_i^j are constant, we have $z_j(p) = z_j(q)$ for $j = 1,\ldots,d$. Since z_j are coordinates, we have $p = q$ which contradicts to the assumption that p and q are distinct points. So far, we have proved F is a homeomorphism. Since each f_i is a linear function, F is clearly smooth. According to the Proposition 5.7 of Lee (2003), F^{-1} is also smooth, so F is a diffeomorphism. Since F is a local isometry and a diffeomorphism, it is a global isometry. ■

When there exists a global isometry, by optimizing our objective functions Equation (1), Equation (2) and Equation (4), Theorem 2 shows that our obtained gradient fields ∇f_i must be parallel. Moreover, the obtained map $F = (f_1,\ldots,f_d)$ is a global isometry. It might be worth noting that there are two variations in finding the global isometry. The first variation is the choice of the orthonormal basis of parallel fields. The second variation is the constant added to each embedding function. Thus the space of global isometry on $\mathcal{M}$ is actually $O(d) \times \mathbb{R}^d$. The first part is the space of orthonormal basis of parallel fields and the second part is the space of constants. Geometrically, the first part represents a rotation of the map and the second part represents a translation of the map.

When there is no isometry between the manifold $\mathcal{M}$ and $\mathbb{R}^d$, our approach can still find a reasonably good embedding function. For example, if the curvature of the manifold is not very high, by minimizing Equation (1), we can still find vector fields are nearly parallel. Consequently, the embedding would be nearly isometric. Please see our experimental results for details. However,

when the curvature of the manifold is extremely high, as shown in Figure 4, the obtained vector fields may have loops or singular points and is no longer a gradient field. The resulting embedding may cause overlap. It would be important to note that, in such cases, the isometric embedding or nearly isometric embedding does not exist.

3.3 Implementation

In real problems, the manifold $\mathcal{M}$ is usually unknown. In this subsection, we discuss how to find the isometric embedding function F from random points.

The implementation includes two steps, first we estimate parallel vector fields on manifold from random points, and then we reconstruct embedding functions by requiring that the gradient fields are as close to the parallel fields as possible. The parallel vector fields are computed one by one under orthogonality constraint. These two steps are described in subsections 3.3.1 and 3.3.2, respectively. After finding d orthonormal vector fields and the corresponding embedding function f_i, the final map F is given by $F = (f_1, \ldots, f_d)$. We discuss how to perform out-of-sample extension in subsection 3.3.3. The detailed algorithmic procedure is presented in subsection 3.3.4.

Given $x_i \in \mathcal{M} \subset \mathbb{R}^m$, $i = 1, ..., n$, we aim to find a lower dimensional Euclidean representation of the data such that the geometrical and topological properties can be preserved. We first construct a nearest neighbor graph by either ε-neighborhood or k nearest neighbors. Let $x_i \sim x_j$ denote that x_i is the neighbor of x_j or x_j is the neighbor of x_i. Let $N(i)$ denote the index set of the neighbors of x_i, that is, $N(i) = \{ j | x_j \sim x_i \}$. For each point x_i, we estimate its tangent space $T_{x_i}\mathcal{M}$ by performing principal component analysis on the neighborhood $N(i)$. We choose the largest d eigenvectors as the bases since $T_{x_i}\mathcal{M}$ is d dimensional. Let $T_i \in \mathbb{R}^{m \times d}$ be the matrix whose columns constitute a orthonormal basis for $T_{x_i}\mathcal{M}$. It is easy to show $P_i = T_i T_i^T$ is the *unique* orthogonal projection from $\mathbb{R}^m$ onto the tangent space $T_{x_i}\mathcal{M}$ (Golub and Loan, 1996). That is, for any vector $a \in \mathbb{R}^m$, we have $P_i a \in T_{x_i}\mathcal{M}$ and $(a - P_i a) \perp P_i a$.

3.3.1 PARALLEL VECTOR FIELD ESTIMATION

Let V be a vector field on manifold. For each point x_i, let V_{x_i} denote the value of the vector field V at x_i and $\nabla V|_{x_i}$ denote the value of ∇V at x_i. According to the definition of vector fields, V_{x_i} should be a tangent vector in the tangent space $T_{x_i}\mathcal{M}$. Thus it can be represented by local coordinates of the tangent space,

$$V_{x_i} = T_i v_i, \tag{5}$$

where $v_i \in \mathbb{R}^d$. We define $\mathbb{V} = \left(v_1^T, \ldots, v_n^T\right)^T \in \mathbb{R}^{dn}$. That is, $\mathbb{V}$ is a dn-dimensional big column vector which concatenates all the v_i's. By discretizing the objective function (1), the parallel field V can be obtained by solving the following optimization problem:

$$\begin{aligned} &\min_{\mathbb{V}} \; E(\mathbb{V}) = \sum_{i=1}^{n} \|\nabla V|_{x_i}\|_{\mathrm{HS}}^2, \\ &s.t. \;\; \|\mathbb{V}\| = 1. \end{aligned} \tag{6}$$

In the following we discuss, for a given point x_i, how to approximate $\|\nabla V|_{x_i}\|_{\mathrm{HS}}$.

Let $\gamma(t)$ be the geodesic connecting x_i and x_j which satisfies $\gamma(0) = x_i$ and $\gamma(d_{ij}) = x_j$, where d_{ij} is the geodesic distance of x_i and x_j. Let $e_{ij} = \gamma'(0)$. Since γ is a geodesic, $e_{ij} \in T_{x_i}\mathcal{M}$ is a unit

vector. Then the covariant derivative of vector field V along e_{ij} is given by (please see Figure 3)

$$\begin{aligned}
\nabla_{e_{ij}} V &= P_i\left(\frac{dV}{dt}|_{t=0}\right) \\
&= P_i \lim_{t\to 0} \frac{V(\gamma(t)) - V(\gamma(0))}{t} \\
&= P_i \frac{(V_{x_j} - V_{x_i})}{d_{ij}} \\
&\approx \sqrt{w_{ij}}(P_i V_{x_j} - V_{x_i}),
\end{aligned}$$

where $d_{ij} \approx 1/\sqrt{w_{ij}}$ approximates the geodesic distance d_{ij} of x_i and x_j. There are several ways to define the weights w_{ij}. Since for neighboring points, Euclidean distance is a good approximation to the geodesic distance, we can define the Euclidean weights as $w_{ij} = \frac{1}{\|x_i - x_j\|^2}$. If the data are uniformly sampled from the manifold, then w_{ij} would be almost constant. So in practice, $0-1$ weights is also widely used which is defined as follows:

$$w_{ij} = \begin{cases} 1, & \text{if } x_i \sim x_j \\ 0, & \text{otherwise.} \end{cases}$$

Since we do not know $\nabla_{\partial_i} V$ for a given basis ∂_i, $\|\nabla V\|^2_{\mathrm{HS}}$ can not be computed according to Equation (3). We define a $(0,2)$ symmetric tensor α as $\alpha(X,Y) = g(\nabla_X V, \nabla_Y V)$, where X and Y are vector fields on manifold. We have

$$\begin{aligned}
\mathrm{Trace}(\alpha) &= \sum_{i=1}^{d} g(\nabla_{\partial i} V, \nabla_{\partial_i} V) \\
&= \|\nabla V\|^2_{\mathrm{HS}},
\end{aligned}$$

where $\partial_1, \ldots, \partial_d$ is an orthonormal basis on tangent space. For the trace of α, we have the following geometric interpretation (see the exercise 1.12 in Chow et al., 2006):

$$\mathrm{Trace}(\alpha) = \frac{1}{\omega_d} \int_{S^{d-1}} \alpha(X,X) d\delta(X),$$

where S^{d-1} is the unit $(d-1)$-sphere, $d\omega_d$ is its volume, and $d\delta$ is its volume form. Thus for a given point x_i, we can approximate $\|\nabla V|_{x_i}\|_{\mathrm{HS}}$ by the following

$$\begin{aligned}
\|\nabla V|_{x_i}\|^2_{\mathrm{HS}} &= \mathrm{Trace}(\alpha)_{x_i} \\
&= \frac{1}{\omega_d} \int_{S^{d-1}} \alpha(X,X)|_{x_i} d\delta(X) \\
&\approx \sum_{j\in N(i)} \|\nabla_{e_{ij}} V\|^2 \\
&= \sum_{j\in N(i)} w_{ij} \left\|P_i V_{x_j} - V_{x_i}\right\|^2. \qquad (7)
\end{aligned}$$

For the third equation, the integral on the left hand side is approximated by the discrete summation on nearest neighbors. It might be worth noting that we implicitly assume that the nearest neighbors

are uniformly sampled. If the sampling is not uniform, then one should use weighted summation to approximate the integral.

Combining Equation (5), the optimization problem Equation (6) reduces to:

$$\begin{aligned}&\min_{\mathbb{V}} E(\mathbb{V}) = \sum_{i\sim j} w_{ij} \left\| P_i T_j v_j - T_i v_i \right\|^2, \\ &s.t.\ \|\mathbb{V}\| = 1.\end{aligned}$$

We first optimize $E(\mathbb{V})$ to find parallel vector fields on manifold, and then re-normalize the vector fields locally.

We will now switch to Lagrangian formulation of the problem. The Lagrangian is as follows:

$$\mathcal{L} = E(\mathbb{V}) - \lambda(\mathbb{V}^T\mathbb{V} - 1).$$

By matrix calculus, we have

$$\begin{aligned}&\frac{\partial E(\mathbb{V})}{\partial v_i} \\ =& -2\sum_{j\in N(i)} w_{ij}(T_i^T(P_iT_jv_j - T_iv_i) - T_i^TP_j(P_jT_iv_i - T_jv_j)) \\ =& 2\sum_{j\in N(i)} w_{ij}((T_i^TT_jT_j^TT_i + I_d)v_i - 2T_i^TT_jv_j) \\ =& 2\sum_{j\in N(i)} w_{ij}((Q_{ij}Q_{ij}^T + I_d)v_i - 2Q_{ij}v_j),\end{aligned}$$

where $Q_{ij} = T_i^TT_j$. Then we have

$$\frac{\partial E(\mathbb{V})}{\partial \mathbb{V}} = 2B\mathbb{V}, B = \begin{pmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{pmatrix},$$

where B is a $dn \times dn$ sparse block matrix. If we index each $d \times d$ block by B_{ij}, then for $i = 1, \ldots, n$, we have

$$B_{ii} = \sum_{j\in N(i)} w_{ij}(Q_{ij}Q_{ij}^T + I), \tag{8}$$

$$B_{ij} = \begin{cases} -2w_{ij}Q_{ij}, & \text{if } x_i \sim x_j \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Requiring that the gradient of $\mathcal{L}$ vanish gives the following eigenvector problem:

$$B\mathbb{V} = \lambda\mathbb{V}.$$

In order to find multiple parallel fields, we just use the eigenvectors corresponding to the smallest eigenvalues of the matrix B. Recall Corollary 1 tells us that if we want to check the orthogonality of two parallel fields at every point, it suffices to compute the integral of the inner product of them

which can be discretely approximated as $\langle \mathbb{V}_i, \mathbb{V}_j \rangle$. Since the matrix B is symmetric, its eigenvectors are mutually orthogonal and, thus, $\langle \mathbb{V}_i, \mathbb{V}_j \rangle = 0$.

Recall Corollary 2 tells us for any point $p \in \mathcal{M}$, the tangent vector V_p of a parallel field V has a constant length. In our objective function (6), we only need to add a global normalization constraint $\|\mathbb{V}\| = 1$ for the sake of simplicity. After finding d vector fields $V_j, j = 1, \ldots, d$, we can further ensure local normalization as follows:

$$\|V_j|_{x_i}\| = 1, \forall i = 1, \ldots, n, j = 1, \ldots, d.$$

3.3.2 EMBEDDING

Once the parallel vector fields V_i are obtained, the embedding functions $f_i : \mathcal{M} \to \mathbb{R}$ can be constructed by requiring their gradient fields to be as close to V_i as possible. Recall that, if the manifold is isometric to Euclidean space, then the vector field computed via Equation (1) is also a gradient field. However, if the manifold is not isometric to Euclidean space, V may not be a gradient field. In this case, we try to find the optimal embedding function f in a least-square sense. This can be achieved by solving the following minimization problem:

$$\Phi(f) = \int_{\mathcal{M}} \|\nabla f - V\|^2 dx.$$

In order to discretize the above objective function, we first discuss the Taylor expansion of f on the manifold.

Let $\exp_x$ denote the exponential map at x. The exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ maps the tangent space $T_x\mathcal{M}$ to the manifold $\mathcal{M}$. Let $a \in T_x\mathcal{M}$ be a tangent vector. Then there is a *unique* geodesic γ_a satisfying $\gamma_a(0) = x$ with initial tangent vector $\gamma_a'(0) = a$. The corresponding exponential map is defined by $\exp_x(ta) = \gamma_a(t)$, $t \in [0, 1]$. Locally, the exponential map is a diffeomorphism.

Note that $f \circ \exp_x : T_x\mathcal{M} \to \mathbb{R}$ is a smooth function on $T_x\mathcal{M}$. Then the following Taylor expansion of f holds:

$$f(\exp_x(a)) \approx f(x) + \langle \nabla f(x), a \rangle, \tag{10}$$

where $a \in T_x\mathcal{M}$ is a sufficiently small tangent vector. In discrete case, let $\exp_{x_i}$ denote the exponential map at x_i. Since $\exp_{x_i}$ is a diffeomorphism, there exists a tangent vector $a_{ij} \in T_{x_i}\mathcal{M}$ such that $\exp_{x_i}(a_{ij}) = x_j$. We approximate a_{ij} by projecting the vector $x_j - x_i$ to the tangent space, that is, $a_{ij} \approx P_i(x_j - x_i)$. Therefore, Equation (10) can be rewritten as follows:

$$f(x_j) = f(x_i) + \langle \nabla f(x_i), P_i(x_j - x_i) \rangle. \tag{11}$$

Since f is unknown, ∇f is also unknown. In the following, we discuss how to compute $\|\nabla f(x_i) - V_{x_i}\|$ discretely. We first show that the vector norm can be computed by an integral on a unit sphere, where the unit sphere can be discretely approximated by a neighborhood.

Let e be a unit vector on tangent space $T_x\mathcal{M}$, then we have (see the exercise 1.12 in Chow et al., 2006)

$$\frac{1}{\omega_d} \int_{S^{d-1}} \langle X, e \rangle^2 d\delta(X) = 1,$$

where S^{d-1} is the unit $(d-1)$-sphere, $d\omega_d$ its volume, and $d\delta$ its volume form. Let ∂_i, $i = 1, \ldots, d$, be an orthonormal basis on $T_x\mathcal{M}$. Then for any vector $b \in T_x\mathcal{M}$, it can be written as $b = \sum_{i=1}^{d} b^i \partial_i$.

Furthermore, we have

$$\begin{aligned}\|b\|^2 &= \sum_{i=1}^{d}(b^i)^2 \\ &= \sum_{i=1}^{d}(b^i)^2\frac{1}{\omega_d}\int_{S^{d-1}}\langle X,\partial_i\rangle^2 d\delta(X) \\ &= \frac{1}{\omega_d}\int_{S^{d-1}}\langle X,b\rangle^2 d\delta(X).\end{aligned}$$

From Equation (11), we see that

$$\langle \nabla f(x_i), P_i(x_j - x_i)\rangle = f(x_j) - f(x_i).$$

Thus, we have

$$\begin{aligned}&\|\nabla f(x_i) - V_{x_i}\|^2 \\ &= \frac{1}{\omega_d}\int_{S^{d-1}}\langle X, \nabla f(x_i) - V_{x_i}\rangle^2 d\delta(X) \\ &\approx \sum_{j\in N(i)}\langle e_{ij}, \nabla f(x_i) - V_{x_i}\rangle^2 \\ &= \sum_{j\in N(i)}\frac{1}{d_{ij}^2}\langle a_{ij}, \nabla f(x_i) - V_{x_i}\rangle^2 \\ &\approx \sum_{j\in N(i)} w_{ij}\langle P_i(x_j - x_i), \nabla f(x_i) - V_{x_i}\rangle^2 \\ &= \sum_{j\in N(i)} w_{ij}\left((P_i(x_j - x_i))^T V_{x_i} - f(x_j) + f(x_i)\right)^2,\end{aligned}$$

where e_{ij} is a unit vector and w_{ij} is the weight, and both of which are the same in Section 3.3.1. In the second equation, the integral is approximated by the discrete summation on nearest neighbors which is the same in Equation (7). In the fourth equation, the vector a_{ij} is approximated by the projection vector $P_i(x_j - x_i)$. Recall that In Section 3.3.1 d_{ij} is approximated by $\frac{1}{\sqrt{w_{ij}}}$ and we have $\|a_{ij}\| = d_{ij}$. Next we show these two approximations are coincide as long as $w_{ij} = O(\frac{1}{\|x_j - x_i\|^2})$. Let us take the Euclidean weight $w_{ij} = \frac{1}{\|x_j - x_i\|^2}$ as an example. We have

$$\lim_{x_j \to x_i}\frac{\|P_i(x_j - x_i)\|^2}{\|x_j - x_i\|^2} = \lim_{\theta\to 0}\cos(\theta)^2 = 1,$$

where θ is the angle between vector $P_i(x_j - x_i)$ and vector $x_j - x_i$.

Let $y_i = f(x_i)$ and $y = (y_1, \ldots, y_n)^T$. The objective function $\Phi(f)$ can be discretely approximated by $\Phi(y)$ as follows:

$$\Phi(y) = \sum_{i\sim j} w_{ij}\left((P_i(x_j - x_i))^T V_{x_i} - y_j + y_i\right)^2.$$

By setting $\partial\Psi(y)/\partial y = 0$, we get

$$-\sum_{i\sim j} w_{ij}s_{ij}(x_j - x_i)^T P_i V_{x_i} + \left(\sum_{i\sim j} w_{ij}s_{ij}s_{ij}^T\right) y = 0,$$

where s_{ij} is an all-zero vector except the i-th element being -1 and the j-th element being 1. Let $L = \sum_{i\sim j} w_{ij} s_{ij} s_{ij}^T$ and $c = \sum_{i\sim j} w_{ij} s_{ij} (x_j - x_i)^T P_i V_{x_i}$. Then we can rewrite the above linear system as follows

$$Ly = c. \tag{12}$$

Algorithm 1 PFE (Parallel Field Embedding)

Input: Data sample $X = (x_1, \ldots, x_n) \in \mathbb{R}^{m\times n}$
Output: $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{d\times n}$
  for $i = 1$ to n **do**
    Compute tangent spaces $T_{x_i}M$
  end for
  Construct matrix B according to Equation (8), Equation (9)
  Find the smallest d eigenvalues $\lambda_1, \ldots, \lambda_d$ and the associated eigenvectors $\mathbb{V}_1, \ldots, \mathbb{V}_d$ of B
  for $l = 1$ to d **do**
    Construct vector field V_l from local representations $\mathbb{V}_l$
    Normalize $V_l(x_i)$ to unit-norm for each x_i
  end for
  Solve linear system $LY = c$
  return Y

It is easy to verify that L is a graph Laplacian matrix (Chung, 1997) and its rank is $n-1$. Thus, the solution of Equation (12) is not unique. If y^* is an optimal solution, $y^* + \text{constant}$ is also an optimal solution. By fixing $y_1 = 0$ in Equation (12), we get a unique solution of y. This is consistent with the continuous case. If f is an optimal solution of Equation (4), then $f + \text{const}$ is also an optimal solution.

3.3.3 OUT-OF-SAMPLE EXTENSION

Most nonlinear manifold learning algorithms do not have straightforward extension for out-of-sample examples. Some efforts (Bengio et al., 2003) are made to generalize existing nonlinear manifold learning algorithms to novel points. We will show that our PFE algorithm has a natural out-of-sample extension.

Given n training points $x_1, \ldots, x_n$ and n' new points $x_{n+1}, \ldots, x_{n+n'}$. Our task is to estimate the embedding results of the new points. For each new point x_j, we first find its k nearest points in the whole data set. Then we compute its tangent space $T_{x_j}\mathcal{M}$ by performing PCA on its neighborhood. We choose the largest d eigenvectors as the bases of $T_{x_j}\mathcal{M}$. Let $T_j \in \mathbb{R}^{m\times d}$ be the matrix whose columns constitute a orthonormal basis for $T_{x_j}\mathcal{M}$. For each vector V_{x_j}, it can be represented by local coordinates of the tangent space. That is, $V_{x_j} = T_j v_j$.

Since $\partial E/\partial \mathbb{V} = 2B\mathbb{V}$ and B is a symmetric matrix, we have

$$E(V) = \mathbb{V}^T B \mathbb{V}. \tag{13}$$

Let $\mathbb{V} = (\mathbb{V}_o^T, \mathbb{V}_n^T) = \left(v_1^T, \ldots, v_n^T, v_{n+1}^T, \ldots, v_{n+n'}\right)^T$, where $\mathbb{V}_o$ denotes the tangent vectors on the *original* training points and $\mathbb{V}_n$ denotes the tangent vectors on *new* points. Then Equation (13) can

be written as

$$E(\mathbb{V}) = \mathbb{V}^T B \mathbb{V} = \begin{bmatrix} \mathbb{V}_o^T & \mathbb{V}_n^T \end{bmatrix} \begin{bmatrix} B_{oo} & B_{on} \\ B_{no} & B_{nn} \end{bmatrix} \begin{bmatrix} \mathbb{V}_o \\ \mathbb{V}_n \end{bmatrix}.$$

Requiring $\partial E(\mathbb{V})/\partial \mathbb{V}_n = 0$, we obtain the following linear system:

$$B_{nn}\mathbb{V}_n = -B_{no}\mathbb{V}_o.$$

After solving this linear system, we get the optimal $\mathbb{V}_n$. Actually, one can compute f_n in the same way, where f_n denotes the function values on new points. We first construct the Laplacian matrix involving all points, including both new and old ones. Then taking derivatives with respect to f_n, we obtain the following linear system:

$$L_{nn}f_n = -L_{no}f_o,$$

where f_o denotes the function values on old points, and L_{nn} and L_{no} are block matrices which are defined in the same way as B_{nn} and B_{no}. Note that the procedure described above involves only local computation on each neighborhood of new samples and solving two sparse linear systems. Therefore our out-of-sample extension algorithm is quite efficient.

3.3.4 ALGORITHM

The PFE algorithm consists of three steps, which is summarized in Algorithm 1.

3.4 Related Work and Discussion

In this section, we would like to discuss the relationship between our work and related work which are based on Laplacian, Hessian and connection Laplacian operators.

The approximation of the Laplacian operator using the graph Laplacian Chung (1997) has enjoyed a great success in the last decade. Several theoretical results (Belkin and Niyogi, 2005; Hein et al., 2005) also showed the consistency of the approximation. One of the most important features of the graph Laplacian is that it is coordinate free. That is, the definition of the graph Laplacian does not depend on any special coordinate systems. The Laplacian operator based methods are motivated by the smoothness criterion, that is, the norm of the gradient $\int_{\mathcal{M}} \|\nabla f\|$ should be small. In the continuous case, with appropriate boundary conditions we have

$$\int_{\mathcal{M}} \|\nabla f\|^2 = \int_{\mathcal{M}} fL(f)dx. \tag{14}$$

Most of previous work focuses on approximating the continuous Laplacian operator. Next we show our method provides a direct way to approximate the integral on the left hand side of Equation (14). First note that

$$\|\nabla f\|^2 = \frac{1}{\omega_d} \int_{S^{d-1}} \langle X, b\rangle^2 d\delta(X).$$

From Equation (11), we see that

$$\langle \nabla f(x_i), P_i(x_j - x_i)\rangle = f(x_j) - f(x_i).$$

Therefore we have

$$\begin{aligned}\|\nabla f(x_i)\|^2 &= \frac{1}{\omega_d}\int_{S^{d-1}} \langle X, \nabla f(x_i)\rangle^2 d\delta(X) \\ &\approx \sum_{j\in N(i)} \langle e_{ij}, \nabla f(x_i)\rangle^2 \\ &= \sum_{j\in N(i)} \frac{1}{d_{ij}^2}\langle a_{ij}, \nabla f(x_i)\rangle^2 \\ &\approx \sum_{j\in N(i)} w_{ij}\langle P_i(x_j - x_i), \nabla f(x_i)\rangle^2 \\ &= \sum_{j\in N(i)} w_{ij}\left(f(x_i) - f(x_j)\right)^2.\end{aligned}$$

It can be seen that this result is consistent with traditional graph Laplacian methods.

Our method is also closely related to the approximation of Hessian operator. Note that if we replace V by ∇f in Equation (1), $E(V)$ becomes Hessian functional (Donoho and Grimes, 2003). It is evident by noticing that $\text{Hess} f = \nabla\nabla f$. The estimation of Hessian operator is very difficult and challenging. Previous approaches (Donoho and Grimes, 2003; Kim et al., 2009) first estimate normal coordinates on the tangent space, and then estimate the first order derivative of the function at each point, which turns out to be a matrix pseudo-inversion problem. One major limitation of this is that when the number of nearest neighbors k is larger than $d + \frac{d(d+1)}{2}$, where d is the dimension of the manifold, the estimation will be inaccurate and unstable (Kim et al., 2009). This is contradictory to the asymptotic case, since it is not desirable that k is bounded by a finite number when the data is sufficiently dense. In contrast, we directly estimate the norm of the second order derivative instead of trying to estimate its coefficients, which turns out to be an integral problem over the nearest neighbors. We only need to do simple matrix multiplications to approximate the integral at each point, but do not have to solve matrix inversion problems. Therefore, asymptotically, we would expect our method to be much more accurate and robust for the approximation of the norm of the second order derivative.

The most related work is the Vector Diffusion Maps (VDM, Singer and Wu, 2011) as we both focus on vector fields rather than embedding functions. VDM is based on the heat kernel for vector fields rather than for functions over the data. It first constructs the heat kernel and orthogonal transformations from the weighted graph. Then VDM defines an embedding of the data via full spectral decomposition of the heat kernel. VDM tries to preserve the newly defined vector diffusion distance for the data when doing embedding. Singer and Wu (2011) also provided theoretical analysis that the construction for the kernel essentially defines the discrete type of the *connection Laplacian operator* and they proved the consistency of such an approximation. We first show that the objective function of finding parallel vector fields of PFE is the same as VDM. According to the Bochner technique (please see Section 3.2 in Chapter 7 of Petersen, 1998), with appropriate boundary conditions we have

$$\int_{\mathcal{M}} \|\nabla V\|_{\text{HS}}^2 = \int_{\mathcal{M}} \langle \nabla^*\nabla V, V\rangle, \tag{15}$$

where $\nabla^*\nabla$ is the connection Laplacian operator. VDM approximated the connection Laplacian operator by generalizing the graph Laplacian operator. We propose to directly approximate the integral on the left hand side of Equation (15). The approximations of PFE and VDM share several

similar important features but also differ in several aspects. Firstly, we use the same way to represent vector fields by using local coordinates of tangent spaces. Intuitively, this is the most natural way to represent vector fields as long as the data are embedded in Euclidean space, or in other words the data has features. It would be very interesting to consider the problem of how to represent vector fields on the graph data, that is, the data that do not have features. Secondly, the approximation of the covariant derivative is similar but different. The idea of computing the covariant derivative is to find a way to compute the difference between vectors on different tangent spaces. VDM proposed an intrinsic way to compute covariant derivative using the concept of *parallel transport*. They first transported the vectors to the same tangent space using the parallel transport, and then compute the difference of vectors on the tangent space. The way of finding the parallel transport between two points is to compute the orthogonal transformation between two corresponding tangent spaces. It turns out that computing the parallel transport is a singular value decomposition problem for each edge of the nearest neighbor graph. Our approach first computes the directional derivative using the (parallel) transport of vectors on Euclidean space, then projects the directional derivative to the corresponding tangent space. The main computational cost is the projection which is the multiplication of matrices with vectors. In continuous cases, they are two different ways to define the covariant derivative. Thirdly, the discrete type connection Laplacian operators (matrix $D^{-1}S-I$ in VDM and matrix B in PFE) are different. The difference is that the transformation matrix O_{ij} in VDM is orthogonal but the transformation matrix Q_{ij} in PFE is not. It is because that we use different ways to approximate the covariant derivative. It might be also worth noting that if we use the orthogonal transformation matrix to compute the covariant derivative, the resulted connection Laplacian matrix followed by our discrete approximation methods would be the same as VDM. Overall, VDM uses vector fields to define the vector diffusion distance, while PFE uses vector fields to find the isometry. Although the procedures of computing vector fields are similar, the motivation and objective are different.

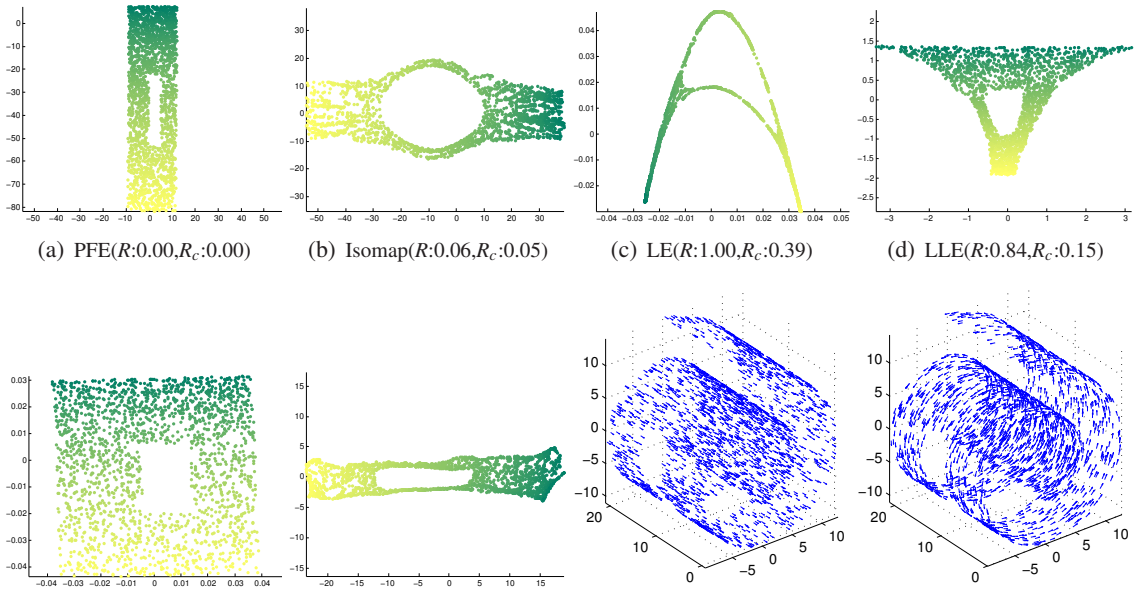
4. Experiments

In this section, we evaluate our algorithm on several synthetic manifold examples and two real data sets.

4.1 Topology

In this example, we study the effectiveness of different manifold learning algorithms for isometric embedding. The data set contains 2000 points sampled from a swiss roll with a hole, which is a 2D manifold embedded in $\mathbb{R}^3$. The swiss roll is a highly nonlinear manifold. Classical linear algorithms like PCA cannot preserve the manifold structure. On the other hand, the swiss roll is a *flat* manifold with zero-curvature everywhere and thus can be isometrically embedded in $\mathbb{R}^2$. We compare our algorithm with several state-of-the-art nonlinear dimensionality reduction algorithms: Isomap, Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE), Hessian Eigenmaps (HLLE), and Maximum Variance Unfolding (MVU).

In all of our experiments, we use a binary-weighted k nearest neighbor graph for each algorithm that needs to construct a graph. Since the manifold learning algorithms usually rely heavily on the choice of the number of nearest neighbors for graph construction, we run each algorithm with the number of nearest neighbors (k) varying in $\{4, 5, \ldots, 20\}$ and show the best results of each algorithm according to the R-score (to be introduced shortly).



(a) PFE(R:0.00,R_c:0.00) (b) Isomap(R:0.06,R_c:0.05) (c) LE(R:1.00,R_c:0.39) (d) LLE(R:0.84,R_c:0.15)

(e) HLLE(R:1.00, R_c:0.21) (f) MVU(R:0.38,R_c:0.11) (g) V_1 (h) V_2

Figure 5: Isometric embedding of 2000 points on a Swiss Roll with a hole. The number of nearest neighbors (k) is set to the best among $\{4, 5, \ldots, 20\}$ for each algorithm when constructing the neighborhood graph. (a)-(f) show the embedding results of various algorithms. (g)-(h) visualize the two vector fields obtained by our algorithm.

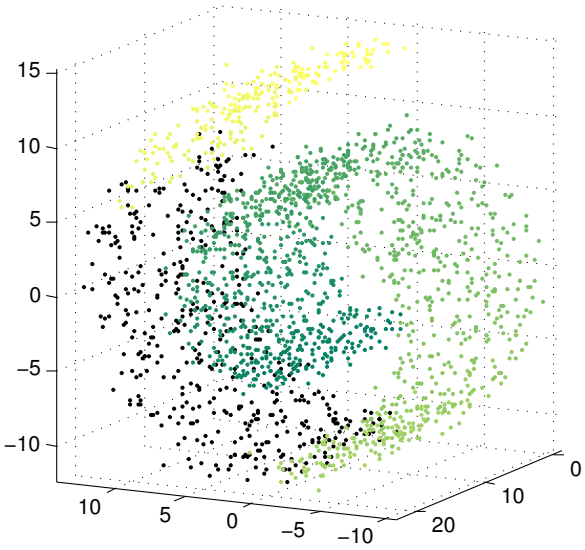


Figure 6: 2000 points are randomly sampled from the Swiss roll with noise. The black points are used for visualization of the gradient field.

The embedding results for all algorithms are shown in Figure 5. Both LE and LLE fail to recover the intrinsic rectangular shape of the original manifold. Isomap performs well at the two ends of the swiss roll but generate a big distorted hole in the middle. This is due to the fact that Isomap can not handle non-convex data. MVU also roughly preserved the overall rectangular shape, but it still generates distortions around the hole. Both HLLE and PFE give very good results. However, it would be important to note that our PFE algorithm generates an *isometric* embedding while the result of HLLE fail to preserve the scale of the coordinates. In order to measure the faithfulness of the embedding in a quantitative way, we employ the normalized R-score (Goldberg and Ritov, 2009):

$$R(X,Y) = \frac{1}{n}\sum_{i=1}^{n} G(X_i,Y_i)/\|HX_i\|_F^2.$$

Here $G(X_i,Y_i)$ is the (normalized) *Procrustes statistic* (Sibson, 1978) which measures the distance between two configurations of points (the original data X_i and the embedded Y_i). $H = I - \frac{1}{k}\mathbf{1}\mathbf{1}^T$ is the centering matrix. A local isometry preserving embedding is considered faithful and thus would get a low R-score. We also report a variant R_c-score (Goldberg and Ritov, 2009)

$$R_c(X,Y) = \frac{1}{n}\sum_{i=1}^{n} G_c(X_i,Y_i)/\|HX_i\|_F^2,$$

where $G_c(X_i,Y_i)$ allows not only rotation and translation, but also rescaling when measuring the distance between two configurations of points. This score can be used to measure conformal map. Please refer to Goldberg and Ritov (2009) for the details of R-score and R_c-score.

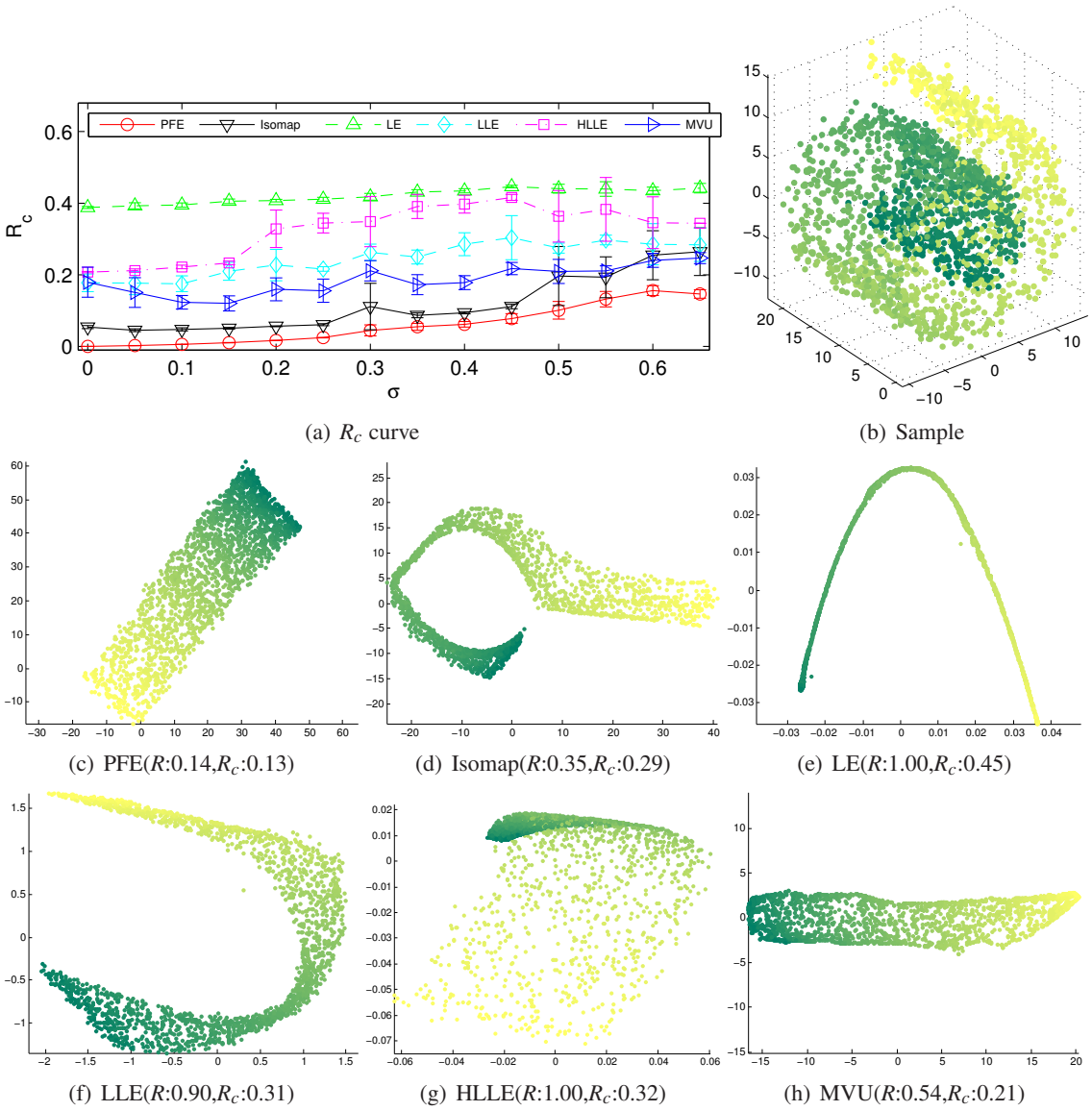
We give the measures of R-score and R_c-score for each algorithm in Figure 5. As can be seen, PFE outperforms all the other algorithms by achieving the minimum R-score and R_c-score. For R-score, except for PFE, Isomap and MVU, all the other three algorithms give excessively high value because their normalization significantly change the scale of the original coordinates. These three algorithm also performs worse than MVU, Isomap and especially PFE in terms of R_c-score.

4.2 Noise

In this example, we compare the performance of different algorithms on noisy data. 2000 points are randomly sampled from the swiss roll without hole, as shown in Fig 6. Then we add random Gaussian noises $\mathcal{N}(0,\sigma^2)$ to each dimension. For each given σ^2, we repeat our experiment 10 times with random noise and the average and standard deviation of the R_c-scores are recorded for each algorithm. The results are shown in Figure 7(a). As can be seen, our PFE method consistently outperforms other algorithms and is relatively stable even under heavy noise. Isomap performs the second best. It also achieves very small standard deviation under small noises, but it becomes very unstable when $\sigma \geq 0.5$.

Figure 7(b) shows the sample points when $\sigma = 0.65$ and Figure 7(c)$\sim$(h) show the embedding results of different algorithms. As can be seen, under such heavy noise, Isomap, LE, and LLE distort the original Swiss roll. HLLE can expand the manifold correctly to some extent, but there is overlap at the top. MVU unfolds the manifold correctly, but does not preserve the isometry well. Our PFE algorithm can successfully recover the intrinsic structure of the manifold.

Note that, algorithms other than PFE do not calculate vector fields explicitly. However, once the embedding result is obtained, we can reversely estimate the gradient field at each point x_i, which is



(a) R_c curve (b) Sample

(c) PFE(R:0.14,R_c:0.13) (d) Isomap(R:0.35,R_c:0.29) (e) LE(R:1.00,R_c:0.45)

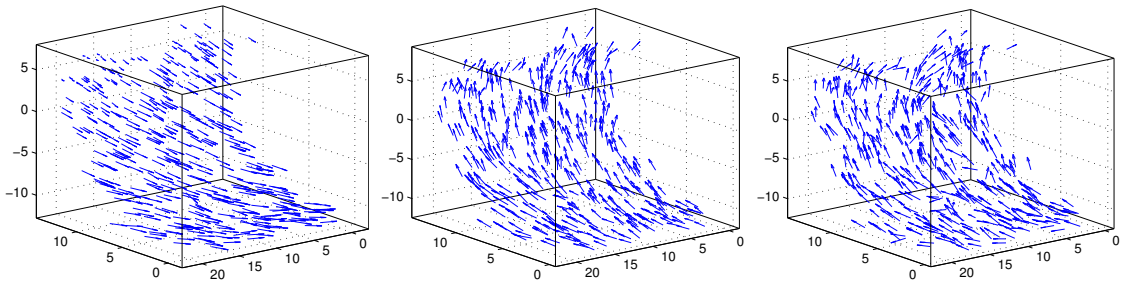
(f) LLE(R:0.90,R_c:0.31) (g) HLLE(R:1.00,R_c:0.32) (h) MVU(R:0.54,R_c:0.21)

Figure 7: Swiss roll with noise. (a) shows the R_c-scores of the five algorithms. (b) shows the 2000 random points sampled from the swiss roll with noise ($\sigma = 0.65$). (c∼h) show the embedding results of the six algorithms.

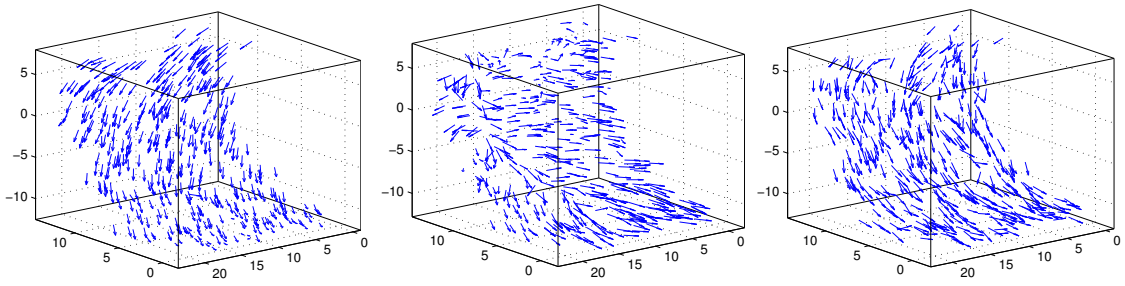
denoted as $\nabla f(x_i)$, by minimizing the following objective function at the local neighborhood of x_i :

$$\sum_{j\sim i}(f(x_j) - f(x_i) - (P_i(x_j - x_i)) \cdot \nabla f(x_i))^2.$$

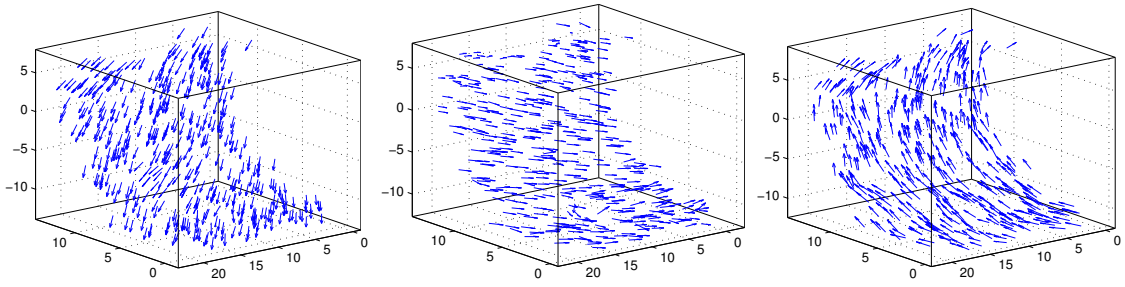
Figure 8 shows the gradient fields obtained by various algorithms. For the purpose of better visualization, we only show the gradient field at part of the Swiss roll (i.e., the black points in Figure 6).



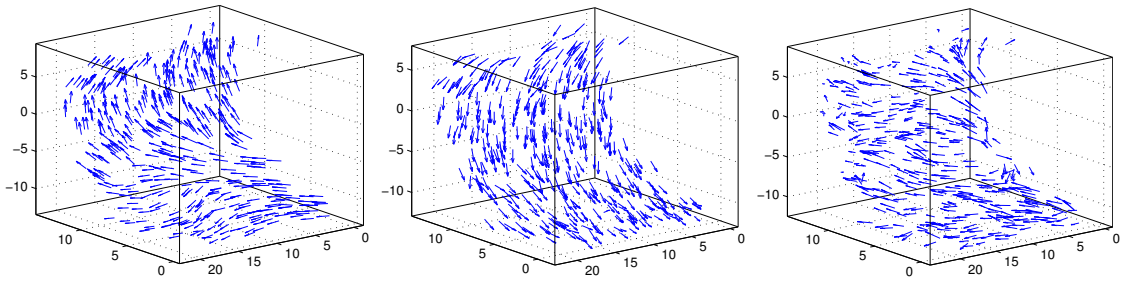
(a) PFE-1 (b) Isomap-1 (c) LE-1



(d) PFE-2 (e) Isomap-2 (f) LE-2

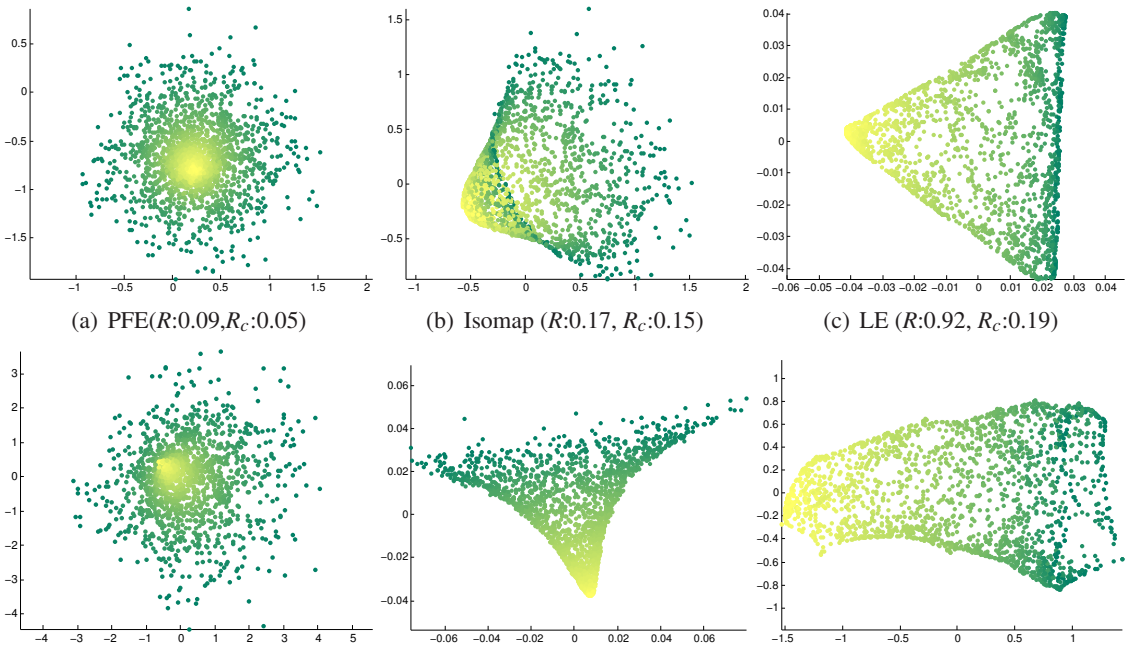


(g) LLE-1 (h) HLLE-1 (i) MVU-1



(j) LLE-2 (k) HLLE-2 (l) MVU-2

Figure 8: The gradient fields obtained by the six algorithms for the Swiss roll. For the sake of better visualization, only part of the gradient fields is shown.



(a) PFE(R:0.09,R_c:0.05) (b) Isomap (R:0.17, R_c:0.15) (c) LE (R:0.92, R_c:0.19)

(d) LLE (R:1.79, R_c:0.08) (e) HLLE (R:0.92, R_c:0.21) (f) MVU (R:0.58, R_c:0.11)

Figure 9: Embedding results of Gaussian. (a∼f) show the embedding results of the six algorithms.

As can be seen, even with noise, our PFE algorithm can accurately find two orthogonal smooth vector fields.

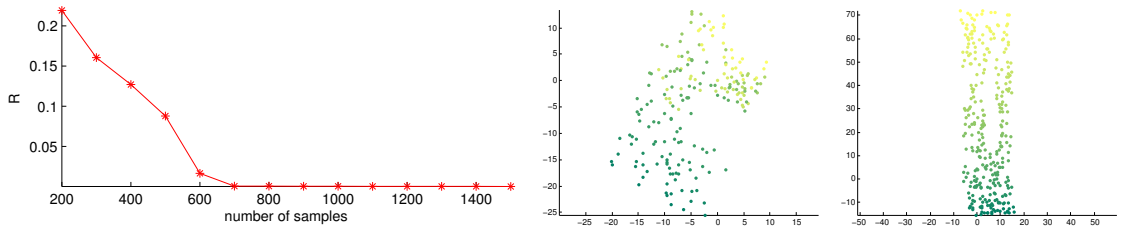
4.3 Curvature

Our PFE algorithm tries to find an isometric embedding. When the underlying manifold cannot be isometrically embedded in $\mathbb{R}^d$, PFE seeks for a least square approximation. The data set used in this experiment is a Gaussian surface. Note that, there is no isometric map from this surface to $\mathbb{R}^2$. The embedding results of the six algorithms are shown in Figure 9. As can be seen, except PFE and LLE, for all the other four algorithms, there is big distortion or overlap. LLE still produces small distortions at the center of the embedding. Moreover, PFE has the advantage over LLE that it optimally preserves the global isometry, and therefore achieving much smaller R-score. Even when measured by the R_c score, PFE outperforms LLE and all the other algorithms.

4.4 Sampling Rate

In this subsection, we test the robustness of our algorithm with respect to the density of data points. Specifically, we use the swiss roll with a hole manifold, and test our algorithm under different sampling rates. As shown in Figure 10(a), we run our algorithm on a data set with the number of data points varying from 200 to 1500 and report the corresponding R-score.

As can be seen, our algorithm produces near-optimal results when the number of points reaches 700. The R-score keeps increasing when the sampling rate reduces. This agrees with the common sense that it is difficult to analyze a small number of samples. However, the R-score does not increase too much even with small sampling rate. As we visualized in Figure 10(b) and (c),



(a) R-score curve (b) 200 points (c) 300 points

Figure 10: Robustness with respect to the sampling rate. (a) shows the R-score achieved by PFE on the swiss roll with hole data set, with respect to different numbers of samples (i.e., different sampling rates). (b) and (c) give the visualization of the embedding obtained from 200 samples and 300 samples, respectively.

although the result of 200 points is distorted and overlapped, the result of 300 points is good. This demonstrates the robustness of our algorithm with respect to different sampling rates.

4.5 Face Manifold

We consider the application of our algorithm on a real-world face data set. The data set[2] contains 698 64×64 gray-scale face images. So each face image is represented as a point in the 4096-dimensional Euclidean space. However, the intrinsic dimensionality may be very low, because the face images are rendered with variations on two pose parameters (up-down and left-right) and one lighting condition parameter. By applying our PFE algorithm, the face images are embedded in a two-dimensional space as shown in Figure 11. We show some representative images selected along the boundary of the embedding. As can be seen, the new coordinates can accurately reflect the face variations.

4.6 Classification after Embedding

In many scenarios, calculating the low-dimensional embedding of the data manifold is not the final step. So another popular evaluation of dimensionality reduction algorithms is to measure the performance of general machine learning algorithms on the low-dimensional representation. In this subsection, we compare the dimensionality reduction algorithms by evaluating the accuracy of 1NN classification on the low-dimensional data.

Specifically, we work with the CMU PIE face data set (Sim et al., 2003). This data set contains 68 subjects with 41,368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. We use the frontal pose (C27) subset with varying lighting and illumination in this experiment. Each face image is of the size 32×32, and we simply work with the raw pixel features without preprocessing. For each algorithm, we calculate the low-dimensional embedding, and then run face recognition with simple Nearest-neighbor classifier. The classification results for 2, 3, 4 and 5 training labels for each person[3] are

2. It is available online from `http://isomap.stanford.edu/face_data.mat.Z`.
3. Labels are randomly chosen and 10 repetitions are run to calculate the averaged performance.

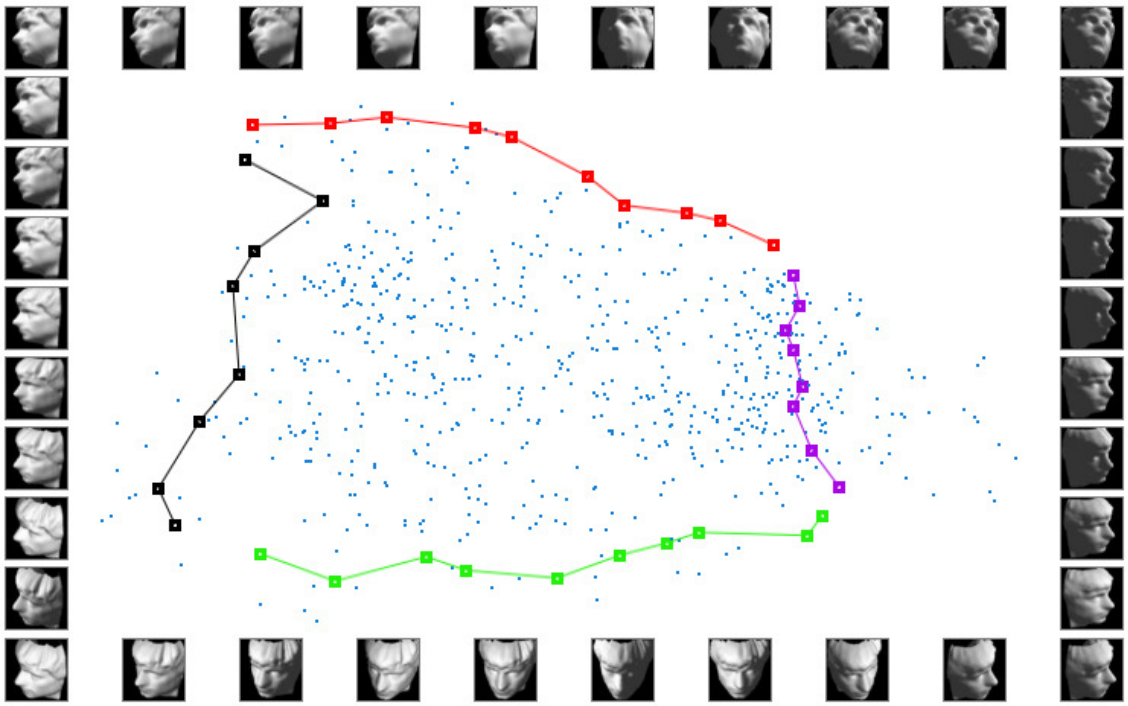


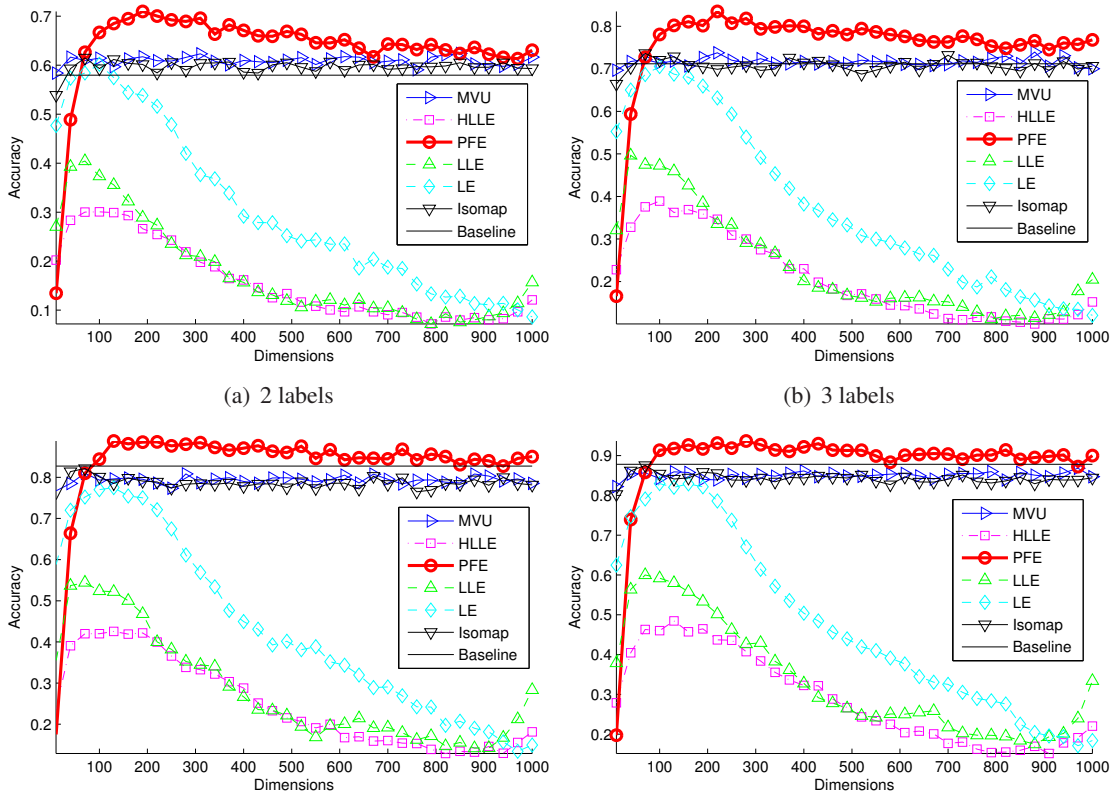
Figure 11: Face manifold embedding. A face data set of 698 images is embedded in the 2-dimensional Euclidean space. Some representative face images are selected along the boundary.

shown in Figure 12. We also show the baseline performance, which corresponds to classification in the original feature space without dimensionality reduction.

As we can see, PFE performs best around dimension 250. It is interesting to see that PFE does not outperform other algorithm when the dimension is very low. But a significant performance gap over other algorithms as well as the baseline can be observed when comparing the peak performance.

4.7 Out-of-Sample Extension

In this experiment, we evaluate the effectiveness of our PFE algorithm for out-of-sample extension. The data set used here is the face data set used in Section 4.5. Specifically, we leave 4 points (corresponding to 4 different poses) out and compute the embedding of the remaining 694 points. The embedding result is shown in Figure 13, which looks almost identical to the previous result in Figure 11. Then we compute the embedding of the 4 testing points by using our out-of-sample extension algorithm. The 4 testing points associated with their face images are shown in Figure 13. As can be seen, these testing samples optimally find their coordinates which reflect their intrinsic property, that is, pose variation. Note beside poses, there is one extra degree of freedom (light condition) in this data set. By embedding the data points in a plane, we ignore the factor of light condition.



(a) 2 labels (b) 3 labels

(c) 4 labels (d) 5 labels

Figure 12: Classification after embedding. Baseline shows the performance without dimensionality reduction.

We also demonstrate the performance of out-of-sample extension by using synthetic data. Specifically, 300 points are uniformly sampled from the swiss roll and embedded in a two-dimensional space by using our PFE algorithm. Then we randomly sample 5000 other points from the same manifold and embed them by using our out-of-sample extension algorithm. The 300 training points and the final embedding results of all samples are shown in Figure 14(a) and Figure 14(b), respectively. As can be seen, our algorithm can find a faithful embedding for both training and testing points.

5. Conclusion

We have introduced a novel local isometry based dimensionality reduction method from the perspective of vector field. Learning on manifold has three most important aspects: geometry, topology and functions on the manifold. Interestingly, there is strong connection between these three aspects and the vector fields on the manifold. In this paper, we are particularly interested in the connection between the geometry and the vector fields on the manifold. A variational method is proposed to find vector fields as parallel as possible on the manifold. The property of such vector field reflects the in-

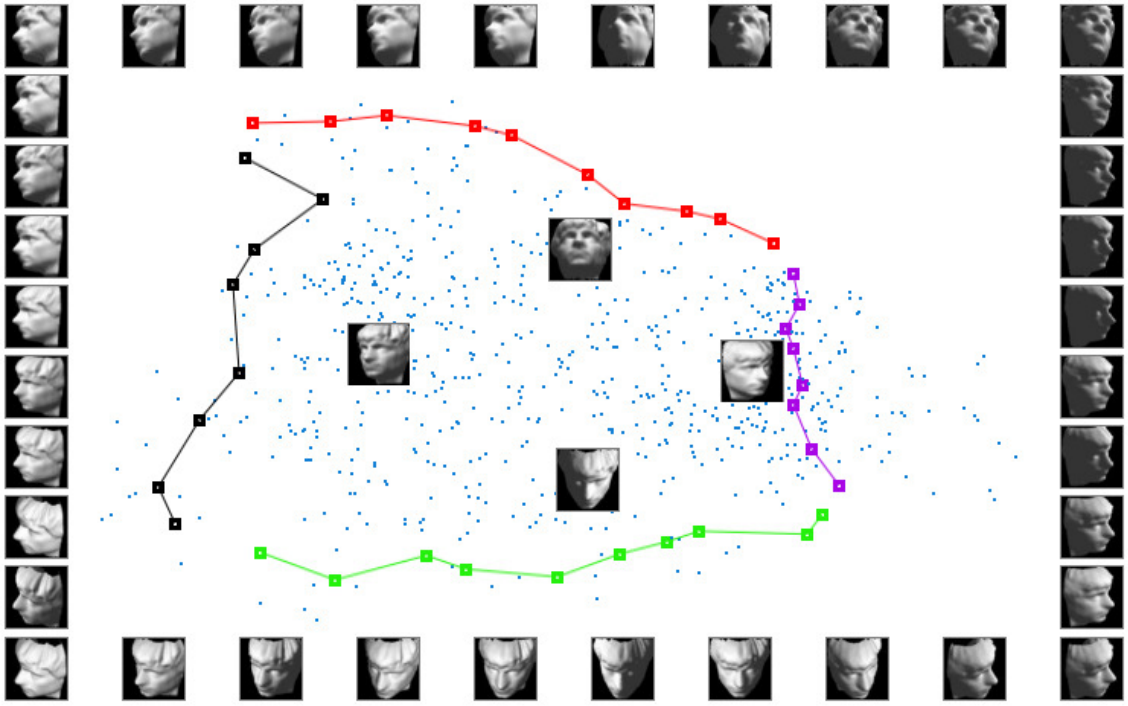
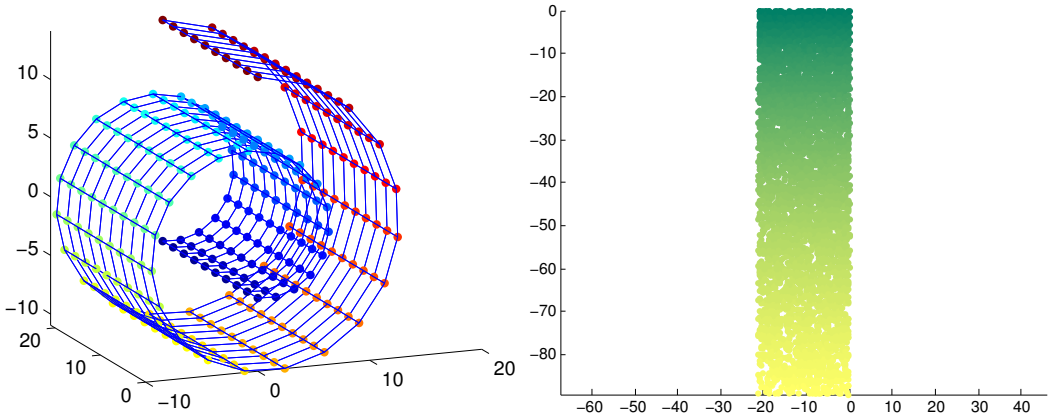


Figure 13: Out-of-sample extension for the face data set. As can be seen, these four testing samples optimally find their coordinates which reflect their intrinsic property, that is, pose variation.



(a) Training Samples (b) Embedding

Figure 14: Out-of-sample extension on the Swiss Roll. (a) shows the 300 uniformly sampled training points. (b) shows the embedding results of both the 300 training points and the 5000 testing points. Both R-score and R_c-score equal to 0.00 in this experiment.

trinsic geometry and topology of the manifold. If the manifold is isometric to Euclidean space, then the obtained vector field is parallel. If the manifold has high curvature or complex topology, then the obtained vector field may be twisted and may have loops or singular points. These properties can be used to study the intrinsic structure of the manifold.

Parallel fields play a central role for finding an isometry. Moreover, parallel fields provide a natural parametric representation of the manifold. Besides dimensionality reduction, they are also useful for other learning problems on the manifold. For example, we can perform classification and regression on the manifold by requiring that the function varies smoothly along the vector fields.

Acknowledgments

This work is supported by National Basic Research Program of China (973 Program) (Grant No: 2012CB316400) and National Natural Science Foundation of China (Grant No: 61125203, 61233011, 90920303).

References

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, Cambridge, MA, 2001.

M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. In *COLT*, pages 486–500, 2005.

M. Belkin and P. Niyogi. Convergence of laplacian eigenmaps. In *Advances in Neural Information Processing Systems 19*, pages 129–136. 2007.

Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems 16*, 2003.

M. Brand. Charting a manifold. In *Advances in Neural Information Processing Systems 16*, 2003.

B. Chow, P. Lu, and L. Ni. *Hamilton's Ricci Flow*. AMS, Providence, Rhode Island, 2006.

F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. AMS, 1997.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21 (1):5 – 30, 2006. Diffusion Maps and Wavelets.

T. Cox and M. Cox. *Multidimensional Scalling*. Chapman & Hall, London, 1994.

A. Defant and K. Floret. *Tensor Norms and Operator Ideals*. North-Holland Mathematics Studies, North-Holland, Amsterdam, 1993.

P. Dollár, V. Rabaud, and S. Belongie. Non-isometric manifold learning: Analysis and an algorithm. In *Proceedings of the Twenty-fourth International Conference on Machine Learning*, pages 241–248, 2007.

D. L. Donoho and C. E. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10):5591–5596, 2003.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2000.

Y. Goldberg and Y. Ritov. Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. *Machine Learning*, 77(1):1–25, 2009.

Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov. Manifold learning: The price of normalization. *The Journal of Machine Learning Research*, 9:1909–1939, 2008.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

J. Ham, D. D. Lee, Sebastian M., and Bernhard S. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

M. Hein, J. Audibert, and U. V. Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In *COLT*, pages 470–485. Springer, 2005.

I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1989.

K. I. Kim, F. Steinke, and M. Hein. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *Advances in Neural Information Processing Systems 22*, pages 979–987. 2009.

S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1393–1403, 2006.

J. M. Lee. *Introduction to Smooth Manifolds*. Springer Verlag, New York, 2nd edition, 2003.

T. Lin and H. Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in Neural Information Processing Systems 18*, pages 955–962. 2006.

P. Petersen. *Riemannian Geometry*. Springer, New York, 1998.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, (10):1299–1319, 1998.

R. Sibson. Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):234–238, 1978.

T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

A. Singer and H.-t. Wu. Vector diffusion maps and the connection Laplacian. *ArXiv e-prints*, 2011.

J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

K. Q. Weinberger, F. Sha, and L. K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1), 2004.