Lin *et al. Applied Informatics* (2015) 2:3
DOI 10.1186/s40535-015-0006-6

Applied Informatics
a SpringerOpen Journal

REVIEW Open Access

A geometric viewpoint of manifold learning

Binbin Lin[1*], Xiaofei He[2] and Jieping Ye[1]

*Correspondence:
Binbin.Lin@asu.edu
[1] Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, 727 E. Tyler St., 85287 Tempe, AZ, USA
Full list of author information is available at the end of the article

Abstract

In many data analysis tasks, one is often confronted with very high dimensional data. The manifold assumption, which states that the data is sampled from a submanifold embedded in much higher dimensional Euclidean space, has been widely adopted by many researchers. In the last 15 years, a large number of manifold learning algorithms have been proposed. Many of them rely on the evaluation of the geometrical and topological of the data manifold. In this paper, we present a review of these methods on a novel geometric perspective. We categorize these methods by three main groups: Laplacian-based, Hessian-based, and parallel field-based methods. We show the connection and difference between these three groups on their continuous and discrete counterparts. The discussion is focused on the problem of dimensionality reduction and semi-supervised learning.

Keywords: Manifold learning; Semi-supervised learning; Geometry

Review

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 try to preserve another meaningful distance, that is, diffusion distance on the manifold. 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.

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.

Tangent space-based methods have also received considerable interest recently, such as local tangent space alignment (LTSA; Zhang and Zha 2005), 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 coordinate 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 2012) and parallel field embedding (PFE; Lin et al. 2013) are much recent works which employ the vector fields to study the metric of the manifold.

Among many of these methods, the core ideas of learning the manifold structure are based on differential operators. In this paper, we would like to discuss differential operators defined on functions and on vector fields. The former include Laplacian and Hessian operators, and the latter include the covariant derivative and the connection Laplacian operator. Since there are lots of geometric concepts involved, we first introduce the background of relevant geometric concepts. Then, we discuss the problem of dimensionality reduction and semi-supervised learning by using these differential operators. The discussion not only focuses on their continuous counterpart but also on their discrete approximations. We try to give a rigorous derivation of these methods and provide some new insights for future work.

Background

In this section, we introduce the most relevant concepts.

Tangent spaces and vector fields

Let $\mathcal{M}$ be a d-dimensional Riemannian manifold. As the manifold is locally a Euclidean space, the key tool for studying the manifold will be the idea of *linear approximation*. The fundamental linear structure of the manifold is the tangent space.

Definition 2.1 (Tangent space; Lee 2003)**.** *Let $\mathcal{M}$ be a smooth manifold and let p be a point on $\mathcal{M}$. A linear map $X : C^\infty(\mathcal{M}) \to \mathbb{R}$ is called a derivation at p if it satisfies:*

$$X(fg) = f(p)Xg + g(p)Xf$$

for all smooth functions $f, g \in C^\infty(\mathcal{M})$. The set of all derivations of $C^\infty(\mathcal{M})$ at p is a vector space called the tangent space to $\mathcal{M}$ at p, and is denoted by $T_p\mathcal{M}$. An element of $T_p\mathcal{M}$ is called a tangent vector at p.

The definition of the tangent space is totally abstract. We first take an example in Euclidean space to show that why a tangent vector is a derivation. Let v denote a geometric tangent vector in $\mathbb{R}^m$. Define a map $D_v|_a : C^\infty(\mathbb{R}^m) \to \mathbb{R}$, which takes the directional derivative in the direction v at a:

$$D_v|_a f = D_v f(a) := \frac{d}{dt}|_{t=0} f(a + tv).$$

Clearly, this operation is linear and it satisfies the derivation rule. Therefore, we might write the directional derivative of f in the direction of Y as $Yf = Y(f) = D_Y f = \nabla_Y f$, where ∇ denotes the covariant derivative on the manifold. Next, we show what a tangent space is on the manifold by using local coordinates. Let $\{x^i | i = 1, \ldots, d\}$ denote a local coordinate chart around p. Then, it can be easily verified by definition that $\partial_i|_p := \frac{\partial}{\partial x_i}|_p$ is a tangent vector at p. Moreover, these coordinate vectors $\partial_1|_p, \ldots, \partial_d|_p$ form a basis for $T_p\mathcal{M}$ (Lee 2003). Therefore, the dimension of the tangent space is exactly the same as the dimension of the manifold. For example, consider a two-dimensional sphere embedded in $\mathbb{R}^3$; given any point of the sphere, the tangent space of the sphere is just a two dimensional plane.

For any smooth manifold $\mathcal{M}$, we define the *tangent bundle* of $\mathcal{M}$, denoted by $T\mathcal{M}$, to be the disjoint union of the tangent spaces at all points of $\mathcal{M}$: $T\mathcal{M} = \cup_{p \in \mathcal{M}} T_p\mathcal{M}$. Now, we define the vector field.

Definition 2.2 (Vector field; Lee 2003)**.** *A vector field is a continuous map $X : \mathcal{M} \to T\mathcal{M}$, usually written as $p \mapsto X_p$, with the property that for each $p \in \mathcal{M}$, X_p is an element of $T_p\mathcal{M}$.*

Intuitively, a vector field is nothing but a collection of tangent vectors with the continuous constraint. Since at each point, a tangent vector is a derivation. A vector field can be viewed as a *directional derivative* on the whole manifold. It might be worth noting that each vector X_p of a vector field X must be in the corresponding tangent space $T_p\mathcal{M}$. Let X be a vector field on the manifold. We can represent the vector field locally using the coordinate basis as $X = \sum_{i=1}^d a^i \partial_i$, where each a^i is a function which is often called the coefficient of X. For the sake of convenience, we will use the Einstein summation convention: when an index variable appears twice in a single term, it implies summation of that term over all the values of the index, i.e., we might simply write $a^i \partial_i$ instead of $\sum_{i=1}^d a^i \partial_i$.

Riemannian metric

Next, we discuss the metric tensor of the manifold. 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 $\mathbb{R}^m$, i.e., $g(u, v) = \langle u, v\rangle$ where $\langle\cdot,\cdot\rangle$ denote the canonical inner product in $\mathbb{R}^m$.

Given coordinates $x(p) = (x^1, \ldots, x^d)$ on an open set U of $\mathcal{M}$, we can thus construct bilinear forms $dx^i \cdot dx^j$. Then, the Riemannian metric g can be represented as:

$$g = g(\partial_i, \partial_j)dx^i \cdot dx^j.$$

Because

$$\begin{aligned} g(X, Y) &= g\left(dx^i(X)\partial_i, dx^j(Y)\partial_j\right) \\ &= g\left(\partial_i, \partial_j\right) dx^i(X) \cdot dx^j(Y). \end{aligned}$$

The function $g(\partial_i, \partial_j)$ are denoted by g_{ij}, i.e., $g_{ij} := g(\partial_i, \partial_j)$. This gives us a representation of g in local coordinates as a positive definite symmetric matrix with entries parameterized over U.

Covariant derivative

A vector field can measure the change of functions on the manifold. Now, we consider the question of measuring the change of vector fields. Let $X = a^i\partial_i$ be a vector field in $\mathbb{R}^d$ where ∂_i denotes the standard Cartesian coordinate. Then, it is natural to define the *covariant derivative* of X in the direction Y as:

$$\nabla_Y X = \left(\nabla_Y a^i\right)\partial_i = Y(a^i)\partial_i.$$

Therefore, we measure the change in X by measuring how the coefficients of X change. However, this definition relies on the fact that the coordinate vector field ∂_i is constant vector field. In other words, $\nabla_Y \partial_i = 0$ for any vector field Y. For general coordinate vector fields, they are not always constant. Therefore, we should give a coordinate free definition of the covariant derivative.

Theorem 2.1 (The fundamental theorem of Riemannian geometry; Petersen 1998)**.** *The assignment $X \to \nabla X$ on $(\mathcal{M}, g)$ is uniquely defined by the following properties:*

1. *$Y \to \nabla_Y X$ is a* $(1, 1)$*-tensor:*

$$\nabla_{\alpha v+\beta w}X = \alpha\nabla_v X + \beta\nabla_w X.$$

2. *$X \to \nabla_Y X$ is a derivation:*

$$\begin{aligned} \nabla_Y(X_1 + X_2) &= \nabla_Y X_1 + \nabla_Y X_2, \\ \nabla_Y(fX) &= (Yf)X + f\nabla_Y X \end{aligned}$$

for functions $f : \mathcal{M} \to \mathbb{R}$.

3. *Covariant differentiation is torsion free:*

$$\nabla_X Y - \nabla_Y X = [X, Y].$$

4. *Covariant differentiation is metric:*

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z Y),$$

where Z is a vector field.

Here, $[\cdot,\cdot]$ denotes the Lie derivative on vector fields defined as $[X, Y] = XY - YX$. Any assignment on a manifold that satisfies rules 1 to 4 is called a *Riemannian connection*. This connection is uniquely determined by these four rules.

Let us see what a connection is in local coordinates. Let X and Y be two vector fields on the manifold, we can represent them by local coordinates as $X = a^i\partial_i$ and $Y = b^j\partial_j$. Now, we can compute $\nabla_Y X$ in local coordinates using the four rules as follows:

$$\nabla_Y X = \nabla_{b^i\partial_i} a^j\partial_j = b^i\nabla_{\partial_i} a^j\partial_j = b^i\partial_i\left(a^j\right)\partial_j + b^i a^j\nabla_{\partial_i}\partial_j. \tag{1}$$

The second equality holds due to the first rule of the connection and the third equality holds due to the second rule of the connection. Since $\nabla_{\partial_i}\partial_j$ is still a vector field on the manifold, we can further represent it as $\nabla_{\partial_i}\partial_j = \Gamma^k_{ij}\partial_k$, where γ^k_{ij} are called Christoffel symbols (Petersen 1998). The Christoffel symbols can be represented in terms of the metric.

Laplacian operator and Hessian operator

Let f be a function on the manifold. The one-form $df : T\mathcal{M} \to \mathbb{R}$ measures the change of the function. In local coordinates, we have $df = \partial_i(f)dx^i$. Note that df is independent to the metric of the manifold. However, the gradient field grad$f = \nabla f$ depends on the metric of the manifold.

Definition 2.3 (Gradient field; Petersen 1998). *Let f be a smooth function on the manifold. The gradient vector field* ∇f *of f is the vector field satisfying:*

$$g(X, \nabla f) = df(X), \forall X \in T\mathcal{M}$$

It might be worth noting that we also have $df(X) = Xf$. In local coordinates, we have $\nabla f = g^{ij}\partial_i f\partial_j$.

We know that on $\mathbb{R}^d$, the Laplacian Δ is defined as $\Delta f = -\text{div}\nabla f^1$. The *divergence* of a vector field, div$X$ is defined as:

$$\text{div}X = \text{tr}(\nabla X).$$

In local coordinates, this is as follows:

$$\text{tr}(\nabla X) = dx^i\left(\nabla_{\partial_i} X\right),$$

and with respect to an orthonormal basis

$$\text{tr}(\nabla X) = \sum_{i=1}^{n} g\left(\nabla_{e_i} X, e_i\right).$$

Thus,

Definition 2.4 (Laplacian; Petersen 1998). *The Laplacian operator is defined as follows:*

$$\Delta f = -tr\left(\nabla(\nabla f)\right) = -div(\nabla f).$$

Also, we can define a second-order derivative on a function. First, we give the definition of second-order derivative:

$$\nabla^2 f(X, Y) = \nabla_X\nabla_Y f - \nabla_{\nabla_X Y} f.$$

Then, we have:

Definition 2.5 (Hessian; Petersen 1998)**.** *The Hessian operator of a function f is a (0, 2) tensor which is defined as follows:*

$$Hessf(X,Y) := \nabla^2 f(X,Y) = \nabla_X \nabla_Y f - \nabla_{\nabla_X Y} f, \forall X, Y \in T\mathcal{M}.$$

A geometric viewpoint of dimensionality reduction

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 most of papers, they 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 $\mathbb{R}^d$.

Variational principals

Since the target space is the Euclidean space $\mathbb{R}^d$, we can represent F by its components $F = (f_1, \ldots, f_d)$, where each $f_i : \mathcal{M} \to \mathbb{R}$ is a function on the manifold.

Next, we introduce various variational objective functionals on F or f_i. For simplicity, we first consider the objective for each component. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on the manifold, and let $C^\infty(\mathcal{M})$ denote the space of smooth functions on the manifold. Clearly, $C^\infty(\mathcal{M})$ is a linear space. Then, we can define an inner product on $C^\infty(\mathcal{M})$ as follows:

$$\langle f, g \rangle := \int_{\mathcal{M}} f(x) g(x) dx.$$

Then, the norm $\| \cdot \|$ on $C^\infty(\mathcal{M})$ can be defined as:

$$\|f\|^2 := \langle f, f \rangle.$$

One can show that it is a valid norm. Also, we can define a norm for vector fields. For any two vector fields X and Y, define the inner product $\langle \cdot, \cdot \rangle$ on the space of vector fields as:

$$\langle X, Y \rangle := \int_{\mathcal{M}} g(X, Y) dx.$$

The norm of X can be defined as $\|X\|^2 := \int_{\mathcal{M}} g(X, X) dx$. Therefore, $(\Gamma(T\mathcal{M}), \| \cdot \|)$ is a normed vector space, where $\Gamma(T\mathcal{M})$ denotes the space of vector fields.

The first functional is given as follows:

$$\max_f \|f\|^2 = \int_{\mathcal{M}} f(x)^2 dx, \quad \text{s.t.} \quad \int_{\mathcal{M}} f(x) dx = 0. \tag{2}$$

Note that when the mean of the function $\int_{\mathcal{M}} f(x) dx$ equals to zero, $\|f\|^2$ measures how the function varies on the manifold. If the function varies dramatically, then $\|f\|^2$ is large; if the function varies a little, then $\|f\|^2$ is small. The statistical meaning of $\|f\|^2$ is exactly the covariance of the random variable f on the manifold. Clearly, this problem is not well-defined as the optimal value can be infinity. In other words, one can always construct a function varies as dramatically as possible. But if we restrict the function f to be linear, then we will have a meaningful solution. Assume $f(x) = a^T x$, then:

$$\max_{\mathbf{a}} \|f\|^2 = \int_{\mathcal{M}} a^T x x^T a dx = a^T \left(\int_{\mathcal{M}} x x^T dx \right) a, \quad \text{s.t.} \quad a^T \int_{\mathcal{M}} x dx = 0. \tag{3}$$

Since a cannot be zero, the constraint becomes $\int_{\mathcal{M}} \mathbf{x} dx = 0$. If we approximate the integral by discrete summations over data points, then Equation 3 becomes the objective function of PCA. The solution can be obtained by singular value decomposition (SVD).

Next, we consider the case when f is a nonlinear function. A widely adopted principal is the smoothness principal: if two points x and y are close, then $f(x)$ and $f(y)$ should also be close. It is sometimes also referred as locality preserving property (He and Niyogi 2003). The smoothness principal can be formularized as minimizing the norm of the gradient of the function (Belkin and Niyogi 2001). Formally, we would like to minimize the following:

$$\min_f \|\nabla f\|^2, \quad \text{s.t.} \quad \|f\|^2 = 1. \tag{4}$$

Under certain boundary conditions, by Stoke's theorem, we have the following equation:

$$\|\nabla f\|^2 = \langle \nabla f, \nabla f \rangle = \int_{\mathcal{M}} g\left(\nabla f, \nabla f\right) = \int_{\mathcal{M}} f \cdot \Delta f = \langle f, \Delta f \rangle.$$

Therefore, Equation 4 is equivalent to:

$$\min_f \langle f, \Delta f \rangle, \quad \text{s.t.} \quad \|f\|^2 = 1.$$

If we rewrite it by the method of Lagrange multipliers, then it becomes:

$$\min_f \langle f, \Delta f \rangle + \lambda \left(\|f\|^2 - 1 \right)$$

By taking derivatives with respect to f, the first-order optimality condition implies:

$$\Delta f = -\lambda f.$$

In discrete cases, one often uses graph Laplacian (Chung 1997) to approximate the Laplacian operator. Some theoretical results (Belkin and Niyogi 2005; Hein et al. 2005) also show the consistency of the approximation. One of the most important features of the graph Laplacian is that it is coordinate free. That is, it does not depend on any special coordinate system. The representative methods include Laplacian eigenmaps (LE; Belkin and Niyogi 2001) and locality preserving projections (LPP; He and Niyogi 2003). Note that (Belkin and Niyogi 2001) has showed that the objective function of LLE is equivalent to minimizing $\langle L^2 f, f \rangle$. If we replace L by L^2 in the last equation, we will get LLE. Therefore, LLE also belongs to this category. Generally, we can replace the Laplacian operator by any compact self-adjoint operators.

Next, we consider another variational objective function which uses the second-order information. The first-order information measures the smoothness of the function, and the second-order information measures the metric information of the manifold.

$$\min_f \int_{\mathcal{M}} \|\text{Hess} f\|^2_{\text{HS}} dx, \quad \text{s.t.} \quad \|f\|^2 = 1. \tag{5}$$

The norm $\|\cdot\|$ represents the standard tensor norm. In matrix form, this norm is equivalent to the Frobenius norm. In this case, it is much harder to get the optimality condition. Since Hessian operator is second order, the optimality condition will be a fourth-order equation. However, we can simply study the null space of the objective function. In other words, we would like to study functions that satisfying $\text{Hess} f = 0$.

Proposition 3.1 (Petersen 1998)**.** *Let $f : \mathcal{M} \to \mathbb{R}$ be a continuous function on the manifold. If* $\text{Hess} f \equiv 0$*, then:*

$$\left(f \circ \gamma\right)(t) = f(\gamma(0)) + \alpha t$$

for each geodesic γ.

The function satisfying $\text{Hess}f \equiv 0$ is said to be linear on the manifold. Proposition 3.1 tells us that a linear function on the manifold varies linearly along the geodesics on the manifold. As pointed out by Goldberg et al. (2008), the final embedding may not be isometry due to the fact of normalization. The representative methods are Hessian-based methods including HLLE.

It motivates the development of vector field-based methods. The objective function of parallel field embedding (PFE; Lin 2013) is as follows:

$$\min_{V} \int_{\mathcal{M}} \|\nabla V\|^2_{\text{HS}}, \quad \text{s.t.} \quad \|V_x\| = 1 \; \forall x \in \mathcal{M}. \tag{6}$$

After solving the above function to obtain parallel vector fields, one solves the following:

$$\min_{f} \int_{\mathcal{M}} \|\nabla f - V\|^2. \tag{7}$$

Interestingly, we can represent Equation 6 by a quadratic form as follows:

$$\int_{\mathcal{M}} \|\nabla V\|^2_{\text{HS}} = \int_{\mathcal{M}} g\left(\nabla^* \nabla V, V\right)$$

$\nabla^* \nabla$ is called the connection Laplacian operator on the manifold. Therefore, Equation 6 can be viewed as a vector field generalization of Laplacian eigenmaps. Taking derivatives of Equation 7 with respect to *f*, we get:

$$\Delta f = -\text{div}(V),$$

where div is the divergence operator on the manifold.

Note that $\text{Hess}f = \nabla\nabla f$. If $V = \nabla f$, then, the objective function in Equation 6 becomes the objective function of HLLE. And also $\Delta f = -\text{div}(V)$ holds by noticing that $\Delta f = -\text{div}(\nabla f)$.

Manifold regularization

Besides dimensionality reduction, the functionals introduced in the last section have been widely employed in semi-supervised learning. In semi-supervised learning, one often gives a set of labeled points, and we aim to learn the label on unlabeled points. It is well known that in order to make semi-supervised learning work, there should be some assumptions on the dependency between the prediction function and the marginal distribution of data (Zhu 2006). In the last decade, the *manifold assumption* is widely adopted in semi-supervised learning, which states that the prediction function lives in a low-dimensional manifold of the marginal distribution. Under the manifold assumption, previous studies focus on using differential operators on the manifold to construct a regularization term on the unlabeled data. These methods can be roughly classified into three categories: Laplacian regularization, Hessian regularization, and parallel field regularization.

We first briefly introduce semi-supervised learning methods with regularization. Let $\mathcal{M}$ be a d-dimensional submanifold in $\mathbb{R}^m$. Let $U \subset \mathcal{M}$ denote the set of labeled points and let f_l denote the label function defined on U. Given U and f_l, we aim to learn a function $f : \mathcal{M} \to \mathbb{R}$ defined on the whole manifold. A framework of semi-supervised learning based on differential operators can be formulated as follows:

$$\underset{f \in C^\infty(\mathcal{M})}{\arg\min}\ E(f) = R_0(f, f_l) + \lambda_1 R_1(f)$$

where $C^\infty(\mathcal{M})$ denotes smooth functions on $\mathcal{M}$, $R_0 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the loss function and $R_1(f) : C^\infty(\mathcal{M}) \to \mathbb{R}$ is a regularization functional. For simplicity, we consider R_0 as a quadratic loss function and write it as $R_0 = \|\delta_U(f - f_l)\|^2$, where δ_U is an indicator function. That is $\delta_U(x) = 1$ if $x \in U$, $\delta_U(x) = 0$ otherwise.

R_1 is often written as a functional norm associated with a differential operator, i.e., $R_1(f) = \int_{\mathcal{M}} \|Df\|^2$ where D is a differential operator.

Laplacian regularization

If D is the covariant derivative ∇ on the manifold, then $R_1(f) = \|\nabla f\|^2 = \langle f, \Delta f\rangle$ becomes the Laplacian regularizer. The objective function can be written as follows:

$$\underset{f \in C^\infty(\mathcal{M})}{\arg\min}\ E(f) = \|\delta_U(f - f_l)\|^2 + \lambda \|\nabla f\|^2.$$

Taking derivatives of $E(f)$ with respect to f, we have:

$$\frac{\partial E(f)}{\partial f} = 2\left(\delta_U(f - f_l) + \lambda \Delta f\right).$$

By the optimality condition $\frac{\partial E(f)}{\partial f} = 0$, we have:

$$f = (I_U + \lambda \Delta)^{-1}\, \delta_U(f_l),$$

where I_U is identity operator with support on U. That is $I_U(f)(x) = 1$ if $x \in U$, $I_U(f)(x) = 0$ otherwise. In discrete cases, one constructs a nearest neighbor graph over the labeled and unlabeled data to model the underlying manifold structure and use the graph Laplacian (Chung 1997) to approximate the Laplacian operator. A variety of semi-supervised learning approaches using the graph Laplacian have been proposed (Belkin et al. 2004; Sindhwani et al. 2005; Zhou et al. 2003; Zhu et al. 2003). It might be worth noting that one can also add another regularizer on f defined by its kernel norm. Such ideas have been discussed in Belkin et al. (2006).

Hessian regularization

If D is the Hessian operator, then $R = R_1(f) = \|\mathrm{Hess} f\|^2$ becomes the Hessian regularizer. The objective function can be written as follows:

$$\underset{f \in C^\infty(\mathcal{M})}{\arg\min}\ E(f) = \|\delta_U(f - f_l)\|^2 + \lambda \|\mathrm{Hess} f\|^2.$$

The Hessian-based methods in unsupervised learning were first proposed in Hessian eigenmaps (HLLE; Donoho and Grimes 2003). The important feature of Hessian is that it preserves second-order smoothness, i.e., preserves the distance or linearity. Steinke and Hein (2008) extend the Hessian regularizer to Elles energy, which is applied to the problem of regression between manifolds. Kim et al. (2009) further propose to employ the Hessian regularizer in semi-supervised regression using an alternative implementation for approximating the Hessian operator in HLLE.

The recent theoretical analysis by Lafferty and Wasserman (2008) shows that using the Laplacian regularizer in semi-supervised regression does not lead to faster minimax rates of convergence. They further propose to use the Hessian regularizer when the Hessian of the prediction function is consistent with the Hessian of the marginal distribution. A more recent work (Nadler et al. 2009) shows that when there are infinite unlabeled data but only finite labeled data, the prediction function learned by using the Laplacian regularizer can be globally smooth but locally bumpy, which is meaningless for learning. These results indicate that the smoothness measured by Laplacian, i.e., the first-order smoothness, is way too general in semi-supervised regression.

Parallel field regularization

Although Hessian regularizer might have a faster convergence rate, but the estimation of the Hessian regularizer is very difficult and sensitive to noise. Lin et al. (2011) proposed to ensure the second-order smoothness by penalizing the parallelism of the gradient field of the prediction function. The objective function of parallel field regularization (PFR; Lin et al. 2011) can be written as follows:

$$\underset{f\in C^\infty(\mathcal{M}),V}{\arg\min}\ E(f,V) = \|\delta_U(f-f_l)\|^2 + \lambda_1 R_1(f,V) + \lambda_2 R_2(V), \tag{8}$$

where $R_1(f,V) = \|\nabla f - V\|^2$ and $R_2(V) = \|\nabla V\|^2_{\mathrm{HS}}$.

Taking derivatives of $E(f,V)$ with respect to f and V, we have:

$$\frac{\partial E(f,V)}{\partial f} = 2\left(\delta_U(f-f_l) + \lambda_1\Delta f + \lambda_1 \mathrm{div}(V)\right),$$
$$\frac{\partial E(f,V)}{\partial V} = 2\left(-\lambda_1\nabla f + \lambda_1 V + \lambda_2\nabla^*\nabla\right).$$

By the optimality condition, we can rewrite it as follows:

$$\begin{pmatrix} I_U + \lambda_1\Delta & \lambda_1\mathrm{div} \\ -\lambda_1\nabla & \lambda_1 I + \lambda_2\nabla^*\nabla \end{pmatrix}\begin{pmatrix} f \\ V \end{pmatrix} = \begin{pmatrix} \delta_U f_l \\ 0 \end{pmatrix} \tag{9}$$

In discrete cases, Lin et al. (2011) gives a systematic way to approximate these differential operators. We list them in Table 1. The detailed definitions of discrete matrices can be found in Lin et al. (2011). It might be worth noting that VDM (Singer and Wu 2012) provides another way of the approximation of the connection Laplacian operator. The similarity and differences are discussed in Lin et al. (2013).

Conclusions

In this paper, we discussed differential operators defined on functions and on vector fields. These differential operators include Laplacian, Hessian, covariant derivative, and the connection Laplacian. We introduced the background of relevant geometric concepts. Then, we discussed the problem of dimensionality reduction and semi-supervised learning by using these differential operators. The discussion not only focused on their continuous counterpart but also on their discrete approximations.

Table 1 Discrete approximation of differential operators

	Differential operators	Discrete approximations
Gradient operator	∇	C
Divergence operator	div	$-C^T$
Connection Laplacian	$\nabla^*\nabla$	B
Metric tensor	g	G

Vector field-based methods are developed recently, which have been proved efficient in many applications including multi-task learning (Lin et al. 2012), manifold alignment (Mao et al. 2013), and ranking (Ji et al. 2012). However, there are still many problems unknown and worth to explore. The first is the convergence of the approximation of the differential operators. The second is the theoretical explanation of vector field regularization. Preliminary work indicates that there is a deep connection between the heat flows on vector fields. The study of heat equation on vector fields and machine learning would be an interesting topic.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BL drafted the manuscript. XH and JY participated in its design and coordination and helped to revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by the National Program for Special Support of Top-Notch Young Professionals, in part by the National Natural Science Foundation of China under Grant 61233011, Grant 61125203, and in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400.

Author details

[1]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, 727 E. Tyler St., 85287 Tempe, AZ, USA. [2]State Key Lab of CAD&CG, No. 866 Yu Hang Tang Road, 85201 Hangzhou, Zhejiang, China.

Received: 23 July 2014 Accepted: 22 January 2015
Published online: 12 March 2015

References

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

Belkin M, Niyogi P (2005) Towards a theoretical foundation for Laplacian-based manifold methods. In: COLT. Curran Associates, Inc. pp 486–500

Belkin M, Niyogi P (2007) Convergence of Laplacian eigenmaps. In: Advances in Neural Information Processing Systems 19. Curran Associates, Inc. pp 129–136

Belkin M, Matveeva I, Niyogi P (2004) Regularization and semi-supervised learning on large graphs. In: COLT. Curran Associates, Inc. pp 624–638

Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from examples. J Machine Learning Res 7:2399–2434

Brand M (2003) Charting a manifold. In: Advances in Neural Information Processing Systems 16. Curran Associates, Inc.

Chung FRK (1997) Spectral graph theory. Regional Conference Series in Mathematics, Vol. 92. AMS

Coifman RR, Lafon S (2006) Diffusion maps. Appl Comput Harmonic Anal 21(1):5–30. Diffusion Maps and Wavelets

Cox T, Cox M (1994) Multidimensional scaling. Chapman & Hall, London

Dollár P, Rabaud V, Belongie S (2007) Non-isometric manifold learning: analysis and an algorithm. In: ICML '07: Proceedings of the 24th International Conference on Machine Learning. Curran Associates, Inc. pp 241–248

Donoho DL, Grimes CE (2003) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc Nat Acad Sci USA 100(10):5591–5596

Duda RO, Hart PE, Stork DG (2000) Pattern classification. 2nd edn. Wiley-Interscience, Hoboken, NJ

Goldberg Y, Zakai A, Kushnir D, Ritov Y (2008) Manifold learning: the price of normalization. J Machine Learning Res 9:1909–1939

Ham J, Lee DD, Mika S, Schölkopf B (2004) A kernel view of the dimensionality reduction of manifolds. In: Proceedings of the Twenty-first International Conference on Machine Learning. Curran Associates, Inc. p 47

He X, Niyogi P (2003) Locality preserving projections. In: Advances in Neural Information Processing Systems 16. Curran Associates, Inc.

Hein M, Audibert J, Luxburg UV (2005) From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In: COLT. Springer. pp 470–485

Ji M, Lin B, He X, Cai D, Han J (2012) Parallel field ranking. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12. pp 723–731

Jolliffe IT (1989) Principal component analysis. Springer, New York

Kim KI, Steinke F, Hein M (2009) Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In: Advances in Neural Information Processing Systems 22. Curran Associates, Inc. pp 979–987

Lafferty J, Wasserman L (2008) Statistical analysis of semi-supervised regression. In: Platt JC, Koller D, Singer Y, Roweis S (eds). Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA. pp 801–808

Lafon S, Lee AB (2006) Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans Pattern Anal Machine Intelligence 28:1393–1403

Lee JM (2003) Introduction to smooth manifolds. 2nd edn. Springer, New York

Lin T, Zha H (2008) Riemannian manifold learning. IEEE Trans Pattern Anal Machine Intelligence 30(5):796–809

Lin B, He X, Zhang C, Ji M (2013) Parallel vector field embedding. J Machine Learning Res 14(1):2945–2977

Lin B, Yang S, Zhang C, Ye J, He X (2012) Multi-task vector field learning. In: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. pp 296–304

Lin B, Zhang C, He X (2011) Semi-supervised regression via parallel field regularization. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. Vol. 24. pp 433–441

Mao X, Lin B, Cai D, He X, Pei J (2013) Parallel field alignment for cross media retrieval. In: Proceedings of the 21st ACM International Conference on Multimedia. ACM. pp 897–906

Nadler B, Lafon S, Coifman R, Kevrekidis I (2006) Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In: Advances in Neural Information Processing Systems 18. Curran Associates, Inc. pp 955–962

Nadler B, Srebro N, Zhou X (2009) Statistical analysis of semi-supervised learning: the limit of infinite unlabelled data. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A (eds). Advances in Neural Information Processing Systems 22. Curran Associates, Inc. pp 1330–1338

Petersen P (1998) Riemannian geometry. Springer, New York

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

Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319

Sindhwani V, Niyogi P, Belkin M (2005) Beyond the point cloud: from transductive to semi-supervised learning. In: ICML. Curran Associates, Inc. pp 824–831

Singer A, Wu H-T (2012) Vector diffusion maps and the connection Laplacian. Commun Pure Appl mathematics 65(8):1067–1144

Steinke F, Hein M (2008) Non-parametric regression between manifolds. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds). Advances in Neural Information Processing Systems 21. Curran Associates, Inc. pp 1561–1568

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

Weinberger KQ, Sha F, Saul LK (2004) Learning a kernel matrix for nonlinear dimensionality reduction. In: ICML '04: Proceedings of the Twenty-first International Conference on Machine Learning. Curran Associates, Inc.

Zhang Z, Zha H (2005) Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J Sci Comput 26(1):313–338

Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2003) Learning with local and global consistency. In: Advances in Neural Information Processing Systems 16. Curran Associates, Inc.

Zhu X (2006) Semi-supervised learning literature survey. Comput Sci, University of Wisconsin-Madison 2:3

Zhu X, Ghahramani Z, Lafferty J (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: Proc. of the Twentieth Internation Conference on Machine Learning. Curran Associates, Inc.