Submitted to the Annals of Statistics

MANIFOLD FITTING AND PROJECTION VIA QUADRATIC APPROXIMATION

BY FIRST AUTHOR[??], SECOND AUTHOR[??,??] AND THIRD AUTHOR[??,??]

National University of Singapore

In this paper, we propose to fit the manifold by a quadratic function defined on the tangent space with a specific form. Compared with the existing linear approximation methods, the quadratic approximation approach can fit the unknown manifold with higher precisions. Because more complicated function is adopted in our manifold fitting process, the pulling back approach turns from the linear least square problem into a nonlinear quartic minimization problem. By bringing in the auxiliary function, we solve the quartic by repeatedly solving a series of quadric minimization problems. Numerical experiments demonstrate that our method has a very strong recovery capability compared with the current works.

1. Introduction. The data residing in the high-dimensional ambient space often have some low-dimensional structure, because of the principal factors which affect the intrinsic generation process are often very limited. The locations of earthquakes or volcanos which caused by the crustal movement can be thought as the points residing on the one-dimensional manifold or the principal flow Panaretos, Pham and Yao (2014); Davenport et al. (2010).

In deep learning approach Goodfellow et al. (2014), using a network with fine-tuning parameters, we can transfer a very low-dimensional data into a very complicated image which some specific meaning. Also, we can use a simple sentence to generate a vivid picture by a well-trained network. These phenomena indicate that very huge amount of data in our daily life are low-dimensional.

To handle the high-dimensional data, there are mainly two approaches. First, reconstruct the data in the low-dimensional space, such that, the reconstructed data in the low-dimensional space also keeps the relation (such as distance, geometry, affine transformation etc) in high dimensional space. This approach containing lots of classical dimension reduction works, such as LLE Roweis and Saul (2000), LTSA Zhang and Zha (2004), Isomap Tenenbaum, De Silva and Langford (2000), Eigenmap Belkin and Niyogi (2003).

The other approach try to fit a new smooth manifold by reducing the noise existing in the data. After the fitting process, the true signal will be strengthened and the noise factors will be diminished. Kernel density estimation Genovese et al. (2014); Ozertem and Erdogmus (2011) is a very popular tool for the manifold fitting by transforming the manifold fitting problem into the ridge estimation problem. The ridge is defined as the set of points which satisfy some relation on gradient and Hessian of the density function. Overall, all the manifold fitting process can be considered with two steps. First, estimate the tangent space which is often implemented through the eigenvalue-decomposition of the covariance matrix. Second, estimate an attraction force which is a direction which points to somewhere near the manifold, this direction often constructed via the weighted shift-mean approach. By using the subspace constraint mean shift algorithm, we can move the outlier onto the manifold to some degree.

The approximation error for the local affine space will yield an approximate error of $\|\tau\|_2^2$, where τ is the local coordinate. In other words, when we restrict the error subjects to the

MSC2020 subject classifications: Primary 00X00, 00X00; secondary 00X00.

Keywords and phrases: Manifold fitting, manifold projection, quadratic approximation.

condition $\|\tau\|_2^2 \leq \epsilon$, the maximum radius of the valid region is with the order of $\epsilon^{1/2}$. In this paper, we propose to approximate the manifold via the quadric function, by adopting this strategy, the approximation error will become $\|\tau\|_2^3$. With the same precision requirement ϵ, we have the maximum radius of our interested region is of the order $\epsilon^{1/3}$. Normally, ϵ is some scalar less than 1, in this way, the quadratic approximation is a better approximation compared with the tangent affine plane.That is to say, to achieve the same precision, we will need much less times of fitting process.

A good approximation of the underlying manifold is the first step in our manifold fitting problem. The second important step needs to project the noisy data onto the fitted manifold which we obtained in the first step. Here, the term 'project' has two meanings. First, the distance should be minimum, second, the direction of the projection should be perpendicular with the tangent place of the fitted manifold. A major reason for the linear approximation to be very widely used is the form of projection yields a very simple linear least square problem.

1.1. *Fitting Model.* In this paper, we are not concentrating on finding the representation of $\phi(\tau)$ under the noiseless assumption. Instead, we assume to have the observations drawn from some low-dimensional manifold and disturbed by some noise, i.e

$$x_i = \tilde{x}_i + \epsilon_i, \quad \tilde{x}_i \in \mathcal{M}$$

where ϵ_i is some noise which obeys some distribution, such as multi-dimensional gaussian noise.

Since the observations $\{x_i, i = 1 : n\}$ are discrete distributed, the idea of manifold fitting aims at generalizing the discrete data and obtaining a low-dimensional approximation of the dataset. The manifold fitting approach can be written as a parameter estimation problem under the observation and the constrained model $\mathcal{G}$.

$$\theta_* = \arg\min_{\theta} \sum_i \text{Loss}(x_i, \mathcal{G}, \theta), \tag{1}$$

where $\mathcal{G}$ represents the abstract model and θ represents the parameters within the model $\mathcal{G}$. Different models (such as linear or nonlinear) correspond to different $\mathcal{G}$. When we obtain our best parameter θ_* from (1), we can use the model $\mathcal{G}(\theta_*)$ to refine the outlier x using $P_{\mathcal{G}(\theta_*)}(x)$ via solving the following minimization problem:

$$P_{\mathcal{G}(\theta_*)}(x) = \arg\min_{y \in \mathcal{G}(\theta_*)} \|x - y\|_2.$$

The works such as Genovese et al. (2014); Ozertem and Erdogmus (2011) all focus on how to get a better affine space to locally approximate the distribution of the data, i.e, $\mathcal{G}(\theta)$ is a linear model. However, at far as we know, the result of linear approximation approach relies heavily on the selection of the origin point (the red dot in Figure 2). An origin selected with good quality will surly improve the ability of recovery of the projection.

1.2. *Manifold Parameterization.* For any manifold $\mathcal{M}$ and any point $x_0 \in \mathcal{M}$, there is a corresponding twice differentiable function $\phi_{x_0}(\tau)$

$$\phi_{x_0}(\tau) : \mathbb{R}^d \to \mathbb{R}^{D-d},$$

such that every point within a local domain $\mathcal{D}_{x_0}(r)$ of $\mathcal{M}$, can be written with a parameterization form of

$$x(\tau) = x_0 + U_{x_0}\tau + U_{x_0}^{\perp}\phi_{x_0}(\tau), \tag{2}$$

where the columns of U_{x_0} are the basis on the tangent space and the columns in $U_{x_0}^{\perp}$ are the basis on the normal space.

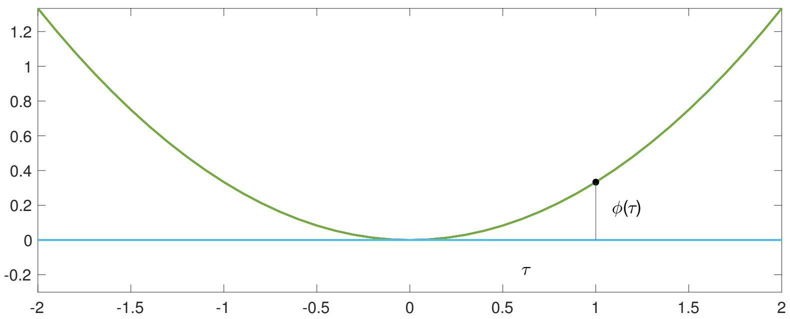

FIG 1. *Geometric interpretation of τ and $\phi(\tau)$ for a 1-D manifold embedded in 2-D*

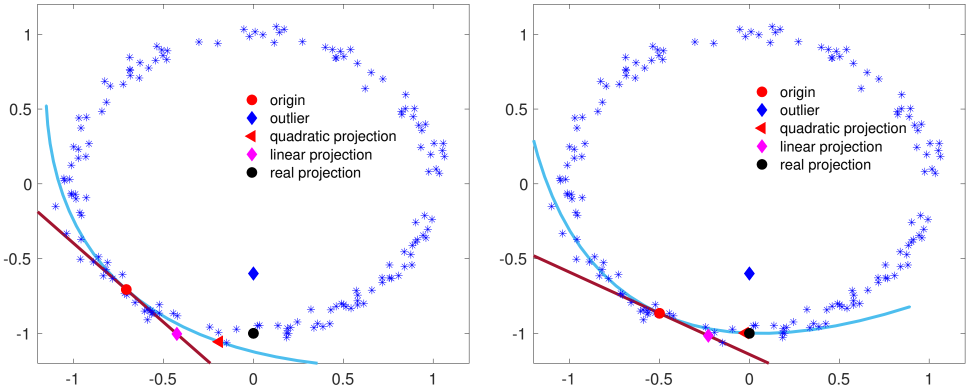

FIG 2. *The reliance of the origin point in the process of manifold fitting and projection*

In other words, corresponding to x_0 and $\mathcal{M}$, there is some radius r, such that:

$$\mathcal{M} \cap \mathcal{D}_{x_0}(r) = \{y | y = x_0 + U_{x_0}\tau + U_{x_0}^{\perp}\phi_{x_0}(\tau), \tau \in \mathbb{R}^d\} \cap \mathcal{D}_{x_0}(r). \tag{3}$$

where $\mathcal{D}_{x_0}(r) = \{x | \|x - x_0\| \leq r\}$, $\phi_{x_0}(\tau) = [\phi_{x_0}^{d+1}(\tau), ..., \phi_{x_0}^{D}(\tau)]^T$ is the function defined on the tangent space, the range of $\phi_{x_0}(\tau)$ represents the coordinate in the normal space and τ is the local coordinate in the tangent space.

To demonstrate the difference behaviors corresponding to the linear and quadratic fitting approaches, we give a toy fitting case in Figure 2 with different origins. The origin x_0 (the red dot) in the left figure is farther away from the true projection (the black dot) than that in the right figure. This example shows the effect of the different fitting errors, which is $O(\|\tau\|_2^2)$ and $O(\|\tau\|_2^3)$ corresponding to the linear and quadratic forms, respectively. From this case, we know that since the linear approximation approach yields a lower-order error, the linear approach relies on a good origin heavily than the quadratic form.

In this paper, we transfer our manifold fitting problem into finding an approximation version of $\phi_{x_0}(\tau)$ through a deep analysis on the characteristic of it. We also show the dominant term for Taylor expansion of $\phi_{x_0}(\tau)$ can be written as a quadratic form of a tensor acting on τ, i.e, $\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}(\tau,\tau)$, where $\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}$ is a third-order tensor with shape of $d \times d \times (D - d)$.

In addition, we show that, by adopting the dominant term in the Taylor expansion of $\phi_{x_0}(\tau)$, the nonlinear function $\phi_{x_0}(\tau)$ can be locally simplified as a quadratic form of $\mathcal{A}(\tau,\tau)$, where $\mathcal{A}$ is the empirical estimation of $\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}$. The unknown parameters $\mathcal{A}$ can be obtained via a linear least square problem. After obtaining the representation of $x_{\mathcal{A}}(\tau)$, we also develop a projection strategy to refine the outlier point $\bar{x}$ by projecting it onto our fitted manifold $\mathcal{M}_{\mathcal{A}}$. Furthermore, we show the projection of $\bar{x}$ onto $\mathcal{M}_{\mathcal{A}}$ can be achieved by repeatedly solving a series of linear least square problems.

2. Related work.

2.1. *Manifold approximation by Linear Methods.* Fitting the manifold with linear methods corresponds to setting $\phi_{x_0}(\tau)$ in (2) equals to zero. Then, the problem becomes to find x_0 and U_{x_0} such that we can approximate the manifold linear around x_0 by a linear parameterization form

$$x(\tau) = x_0 + U_{x_0}\tau$$

Note that, the function $x(\tau)$ is a local approximation of $\mathcal{M}$ at x_0. The choice of x_0 depending on our interest area, such that, if we want to project an outlier $\bar{x}$ onto the manifold, x_0 can be selected as the nearest point x_0 on $\mathcal{M}$, such that

$$\|x_0 - \bar{x}\|_2 = \min_{x_0 \in \mathcal{M}} \|x_0 - \bar{x}\|_2$$

There are lots of works which approximate the basis U_{x_0} in the tangent space by a linear space approach. One approach is to directly fit the manifold by finding the eigenspace of some matrices, e.g the covariance matrix or Hessian of KDE. Then using the eigenvalue-decomposition to get the eigenspace corresponding to the d largest eigenvalues. The Hessian of the the classical KDE approach shares the same eigenspace of

$$C(\bar{x}) = \sum_i K_h(\bar{x}, x_i)(x_i - \bar{x})(x_i - \bar{x})^T$$

If transformed by the concave increasing function $\log$ with respect to the KDE function, the deterministic term in the Hessian matrix will become

$$C(\bar{x}) = \sum_i K_h(\bar{x}, x_i)(x_i - c(\bar{x}))(x_i - c(\bar{x}))^T$$

where $c(\bar{x})$ is the weighted shift mean vector. In this case, using the basis (consisting of the columns of U_d) corresponding to the largest d eigenvalues of the covariance matrix $C(\bar{x})$.

When we use the locally shift mean to replace the origin x_0, the affine space yields the form as

$$x(\tau) = c(\bar{x}) + U_d\tau,$$

where τ is the coordinate in the d-dimensional space. The projection onto the $x(\tau)$ of $\bar{x}$ is

$$\bar{x} + U_\perp U_\perp^T(c(\bar{x}) - \bar{x}) = c(\bar{x}) + U_d U_d^T(\bar{x} - c(\bar{x}))$$

There are also complex methods such as Fefferman et al. (2018); Yao and Xia (2019) by approximating the normal space at x as a weighted combination of the normal space at each neighbor sample x_i as

$$A = \sum_i \alpha_i \Pi_i^\perp$$

where $\Pi_i^\perp$ is the estimated normal space at the observation x_i and α_i is the weight for x_i. Then the normal space can be obtained from eigenvalue decomposition of A and picking up the eigenvectors corresponding to the $D - d$ largest eigenvalues.

2.2. *Manifold approximation by Nonlinear Least-square Approach.* Using the moving least-square (MMLS) projection approach to approximate the manifold is proposed by Sober and Levin (2019). In their work, they choose to approximate the manifold with two steps.

1. Given the outlier x, find the local coordinates by solving a minimization problem
$$(\hat{q}(x), \hat{\mathcal{H}}(x)) = \arg \min_{x-q \perp \mathcal{H}} \sum_{i=1}^{n} d^2(x_i, \mathcal{H})\theta(\|x_i - q\|),$$
where $\mathcal{H}$ is the d-dimensional affine subspace and q is the origin. The squared distance from x_i to $\mathcal{H}$ is $d^2(x_i, \mathcal{H}) = \|V_{\mathcal{H}}^T(x_i - q)\|_2^2$. The parameter $\hat{q}$ and $\hat{H}$ is achieved by a repeated minimizing procedure.
2. Using the local coordinate $\tau_i = V_{\mathcal{H}}^T(x_i - q)$, fit the polynomial by minimizing
$$\hat{g} = \min_{p \in \Pi_m^d} \sum_{i=1}^{n} \|p(\tau_i) - f_i\|_2^2 \theta(\|x_i - q\|),$$
where Π_m^d is the polynomial function class up to order m defined in the d-dimensional space. Then, the projection is defined as: $P_m(r) = \hat{g}(0)$.

2.3. *Difference.* Our approach differ with the former approaches in three respects.

1. We do not require the origin satisfy $r - q \perp H$ in our approach by simplifying the first step. Since our method is not sensitive to the initial point q, by using the initial guess of q, H can be obtained by PCA very easily.
2. We restrict the form of our polynomial which will result in a with very concise form and save lots of extra parameters. The manifold fitting by a second-order paraboloid function has a more concrete form by defining a tensor which has the clear meaning by representing the curvature of our manifold. The tensor can be obtained by solve a linear least square problem in each of the normal dimensions.
3. The projection of outlier r onto the fitted manifold $\mathcal{M}_{\mathcal{A}}$ is not restraint to be the origin point of $\mathcal{M}_{\mathcal{A}}$. It can be any points $\mathcal{M}_{\mathcal{A}}$ as long as the minimum criteria is achieved.

The strengths of our method can be summarized as:

1. By fitting a function with the quadratic form, the curvature of the manifold is considered in our algorithm. Because of the existing of the curvature, the fitted manifold $\mathcal{M}_{\mathcal{A}}$ can approximate the true manifold in a relatively larger scope.
2. Our algorithm does not rely on the accuracy of the estimated origin point too much, which means a relatively rough estimation of the origin is accessible in our algorithm. Instead of the origin, we just need to get a point (nearby x) which is also supposed to be next to the true $\mathcal{M}$.
3. The solution of our algorithm has a clear geometric interpretation via the optimized $\hat{\tau}$. When we got the local coordinate $\hat{\tau}$, we can think of our projected point $x(\hat{\tau})$ (the projection onto $\mathcal{M}_{\mathcal{A}}$) as a modification of the origin x_0, through,
$$x(\tau) = U_{\perp}\mathcal{A}(\tau, \tau) + U\tau + x_0.$$
where $U\tau$ is the modification of x_0 in the tangent space and $\mathcal{A}(\tau, \tau)$ is the modification of x_0 in the normal space.

3. Preliminary: Tensor Operation. In this section, we give some preliminary knowledge with tensor operation and tensor differentiation. Using tensor notations will bring in lots of convenience in our notations and discussion.

3.1. *Multiplication.* For a three-order tensor $\mathcal{A}$ with the shape of $d \times d \times (D-d)$, we can also think of $\mathcal{A}$ as an operator acting on the d or $(D-d)$ dimensional vector space, besides regarding $\mathcal{A}$ as a particular array of numbers. For a vector $\tau \in \mathbb{R}^d$, the tensor $\mathcal{A}$ acting on τ will result in a matrix denoted as $\mathcal{A}(\tau)$ of shape $d \times (D-d)$, of which the j, k-th element is

$$\{\mathcal{A}(\tau)\}_{jk} = \sum_i \tau_i \mathcal{A}_{i,j,k},$$

which is a weighted combination of the slices in the first dimension of the tensor $\mathcal{A}$. Similarly, for two vectors $\tau, \eta \in \mathbb{R}^d$, the tensor $\mathcal{A}$ acting on τ, η will result in a vector denoted as $\mathcal{A}(\tau, \eta) \in \mathbb{R}^{D-d}$, whose k-th element is

$$\{\mathcal{A}(\tau, \eta)\}_k = \sum_{i,j} \tau_i \eta_j \mathcal{A}_{i,j,k}.$$

Clearly, we have the vector $\mathcal{A}(\tau, \eta)$ can be written as a matrix-vector product by: $\mathcal{A}(\tau, \eta) = \mathcal{A}(\tau)^T \eta$ and the summation with index i, j can also be regarded as an inner product of matrix, i.e, $\{\mathcal{A}(\tau, \eta)\}_k = \langle \tau \eta^T, \mathcal{A}_{\cdot\cdot k} \rangle$.

3.2. *Differentiation.* For the $(D-d)$ dimensional vector $\mathcal{A}(\tau, \tau)$, we can take the differential operation with respect to τ. The first derivative of $\mathcal{A}(\tau, \tau)$ with respect to τ will result in a matrix of size $d \times (D-d)$ and the second derivative of $\mathcal{A}(\tau, \eta)$ with respect to τ and η will get the tensor $\mathcal{A}$,

$$\begin{aligned} \nabla_\tau \mathcal{A}(\tau, \tau) &= 2\mathcal{A}(\tau) \in \mathbb{R}^{d \times (D-d)}, \\ \nabla_\eta \nabla_\tau \mathcal{A}(\tau, \eta) &= \mathcal{A} \in \mathbb{R}^{d \times d \times (D-d)}, \\ \nabla_\tau \nabla_\tau \mathcal{A}(\tau, \tau) &= 2\mathcal{A} \in \mathbb{R}^{d \times d \times (D-d)}. \end{aligned}$$

Noticing that $\mathcal{A}(\tau, \tau, \iota)$ is a scalar, taking the derivative for τ twice will result in a symmetric matrix (similar with Hessian)

$$\begin{aligned} \nabla_\eta \nabla_\tau \mathcal{A}(\eta, \tau, \iota) &= \mathcal{A}(\iota) = \sum_k \iota_k \mathcal{A}_{\cdot\cdot k} \in \mathbb{R}^{d \times d}, \\ \nabla_\tau \nabla_\tau \mathcal{A}(\tau, \tau, \iota) &= 2\mathcal{A}(\iota) = 2\sum_k \iota_k \mathcal{A}_{\cdot\cdot k} \in \mathbb{R}^{d \times d}, \end{aligned}$$

where $\mathcal{A}(\iota)$ is a matrix obtained via folding the tensor in the third dimension with the weight ι, i.e, the i, j-th element is:

$$\mathcal{A}(\iota)_{i,j} = \sum_k \iota_k \mathcal{A}_{i,j,k}.$$

The tensor multiplication operations and differentiations will be used frequently in our manifold projection section of this manuscript. Working with the tensor can make our writing more concise.

4. Fitting Model. There are two meanings of the fitting model. The first refers to locally fitting a complicated function by a simple form such as Taylor expansion. The second indicates we use a generalized representation such that the measurement with respect to the observations and the representation is optimal under some criteria.

For a complicate nonlinear function $\phi(\tau)$, we can fit locally via the lower order Taylor expansion such as

$$\phi_q(\tau) = \phi(\tau_0) + \langle \nabla_\tau \phi(\tau)|_{\tau=\tau_0}, \tau - \tau_0 \rangle + \frac{1}{2} \nabla_\tau \nabla_\tau \phi(\tau)|_{\tau=\tau_0}(\tau - \tau_0, \tau - \tau_0). \tag{4}$$

1. The linear function $\phi(\tau_0)+\langle\nabla_\tau\phi(\tau)|_{\tau=\tau_0},\tau-\tau_0\rangle$ can be thought as a locally fitting model of $\phi(\tau)$ at $\tau=\tau_0$ with the error of order $O(\|\tau-\tau_0\|_2^2)$.
2. The quadratic function (4) can be thought as a locally fitting model of $\phi(\tau)$ at $\tau=\tau_0$ with the error of order $O(\|\tau-\tau_0\|_2^3)$.

In this paper, we solve the manifold fitting problem via the local quadratic approximation under the noisy observations. To simplify the problem, we first assume:

ASSUMPTION 4.1. *Assume the observations $\{x_i, i=1,...n\}$ and the origin x_0 are drawn exactly from some unknown d-dimensional manifold $\mathcal{M}$ which is embedded in the D-dimensional ambient space.*

4.1. *Quadratic surface in 3-D space.*

A manifold is a topological space that locally resembles Euclidean space. Except for a few special cases, we cannot have global representation to parameterize the manifold. As a result, we try to approximate the manifold locally within an interested area. First, we show a simple demo for our motivation.

EXAMPLE 1. The data of blue circles evenly distributed on a 2-D sphere as shown in (3) is a toy example of manifold. Fixing a point $x_0\in\mathbb{S}^2$, we can have a tangent place. The yellow circles are the projection of the data onto the tangent plane. The continuous elliptical paraboloid surface is the structure what we wanted.

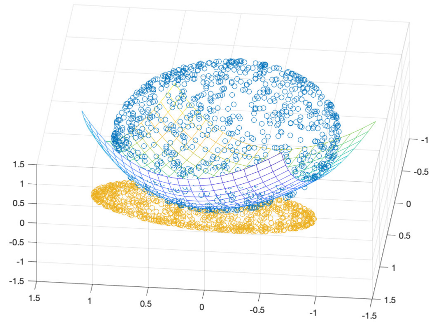

FIG 3. *Fitting a manifold with the elliptical paraboloid surface*

Obviously, from the Figure 3 we can see that the mesh surface is a better structure to approximate $\mathbb{S}^2$ compared with the tangent plane. Here, the parameterized function of the quadratic form $x_A(\tau):\mathbb{R}^2\to\mathbb{R}^3$ yields:

$$x_A(\tau)=x_0+U_{x_0}\tau+U_{x_0}^{\perp}\tau^T A\tau$$

where A is matrix of size 2×2 which is used to control the shape and the direction of the paraboloid surface. The eigenvalues of A represents the curvature or flatness in the corresponding eigenvector direction.

REMARK. In 3-D space, obviously, when the paraboloid surface is at one side of the tangent plane, the matrix A is positive definite or negative definite.

4.2. *Quadratic surface in high-dimensional space* . Instead of fitting the manifold by the tangent plane, we bring a unknown function $\phi_{x_0}(\tau)$ from the tangent space to the normal space

$$\phi_{x_0}(\tau) : \mathbb{R}^d \to \mathbb{R}^{D-d}.$$

Using $\phi_{x_0}(\tau)$, we know any manifold $\mathcal{M}$ locally at x_0 can be written as the range of function $x(\tau)$

$$x(\tau) = x_0 + U_{x_0}\tau + U_{x_0}^{\perp}\phi_{x_0}(\tau), \tag{5}$$

where $\phi_{x_0}(\tau) = [\phi_{x_0}^{d+1}(\tau), ..., \phi_{x_0}^{D}(\tau)]^T$ is the function from tangent space to the normal space. τ is the local coordinate in the tangent space. From (5), we know that there is a one-to-one correspondence between $\phi_{x_0}(\tau)$ and $x(\tau)$ as:

$$\phi_{x_0}(\tau) = {U_{x_0}^{\perp}}^T (x(\tau) - x_0) \tag{6}$$

From (6), we know if we are given the structure of $\phi_{x_0}(\tau)$ with some parameters, we can have the representation of $\phi_{x_0}(\tau)$.

1, x_0 is any point on the manifold we are interested in.
2, U_{x_0} is the basis of principal space at x_0 which can be obtained from the local PCA.
3, $U_{x_0}^{\perp}$ is the orthogonal directions at x_0 which can also be obtained from the local PCA.
4, The only remaining part is to $\phi_{x_0}(\tau)$.

Note that, to implement the local PCA, the local covariance matrix at x_0 is defined as $C(x_0) = \sum_i K_h(x_0, x_i)(x_i - x_0)(x_i - x_0)^T$. The basis of the tangent space can be obtained from the eigenvalue decomposition as

$$C(x_0) = [U_d, U_{D-d}] \begin{bmatrix} \Lambda_1, & 0 \\ 0 & \Lambda_2 \end{bmatrix} [U_d, U_{D-d}]^T,$$

in which from the decomposition, we can set $U_{x_0} = U_d, U_{x_0}^{\perp} = U_{D-d}$. In the local domain $\mathbb{R}_{x_0}^D(r) \cap \mathcal{M}$, $\phi_x(\tau)$ can be approximated by the polynomial of order 2. Thus, we will get

$$\phi_{x_0}(\tau) = \phi_{x_0}(0) + \nabla\phi_{x_0}(\tau)|_{\tau=0}(\tau) + \frac{1}{2}\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}(\tau,\tau) + O(\|\tau\|_2^3) \tag{7}$$

where $\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}$ stands for the second fundamental form which is a three order tensor with the size $d \times d \times (D-d)$. Acting on τ twice will result in a vector of size $D-d$. The linear operator $\nabla\phi_{x_0}(\tau)|_{\tau=0} \in \mathbb{R}^{(D-d)\times d}$, acting on τ, will result in a vector size of $D-d$.

THEOREM 4.2. *If the columns of U_{x_0} are the basis of tangent space of $\mathcal{M}$ at x_0, then, the nonlinear function $\phi_{x_0}(x)$ satisfies:*

$$\phi_{x_0}(0) = 0, \quad \nabla\phi_{x_0}(0) = 0.$$

PROOF. Recall that: $x(\tau) = x_0 + U_{x_0}\tau + U_{x_0}^{\perp}\phi_{x_0}(\tau)$. Let $\tau = 0$, then, we have

$$U_{x_0}^{\perp}\phi_{x_0}(0) = 0$$

Because of $U_{x_0}^{\perp}$ is a orthonormal matrix, $U_{x_0}^{\perp}\phi_{x_0}(0) = 0$ implies $\phi_{x_0}(0) = 0$.

For any curve $\gamma(t) \in \mathbb{R}^d, \gamma(0) = 0, \gamma'(0) = v$. If we take the direction derivative for $x(\tau)$:

$$\begin{aligned} \partial_v x(\tau) &= \frac{dx(\gamma(t))}{dt}|_{t=0} = \lim_{t\to 0}\frac{x(\tau + vt) - x(\tau)}{t} \\ &= \lim_{t\to 0}\frac{U_{x_0}(\tau + tv) - U_{x_0}(\tau) + U_{x_0}^{\perp}\phi_x(\tau + tv) - U_{x_0}^{\perp}\phi_x(\tau)}{t} \\ &= U_{x_0}v + U_{x_0}^{\perp}\nabla\phi_x(\tau)v \end{aligned}$$

Because $\partial_v x(\tau) \in \mathcal{T}_{\mathcal{M}}(x_0)$, we have $U_{x_0}^{\perp} \nabla \phi(\tau) \in \mathcal{T}_{\mathcal{M}}(x_0)$.Thus, we have

$$U_{x_0}^{\perp} \nabla \phi(\tau) = 0,$$

which implies $\nabla \phi(\tau) = 0$ because the columns of the orthonormal matrix $U_{x_0}^{\perp}$ are linear independent . □

Because of Theorem 4.2, recalling that (7), we know $\mathcal{M}$ can be parameterized with the remainder term:

$$x(\tau) = x_0 + U_{x_0}\tau + \frac{1}{2} U_{x_0}^{\perp} \nabla\nabla \phi_{x_0}(\tau)|_{\tau=0}(\tau,\tau) + O(\|\tau\|_2^3)$$

For notational convenience, we denote the tensor as $\mathcal{A}_{x_0} = \frac{1}{2}\nabla\nabla\phi_{x_0}(\tau)|_{\tau=0}$ and define a new manifold $\mathcal{M}_{\mathcal{A}}$ derived from $y_{\mathcal{A}}(\tau)$ as

$$\mathcal{M}_{\mathcal{A}} = \{z | z = x_0 + U_{x_0}\tau + U_{x_0}^{\perp} \mathcal{A}_{\phi_{x_0}}(\tau,\tau), \tau \in \mathbb{R}^d\} \tag{8}$$

From (8), we know $\mathcal{M}_{\mathcal{A}}$ is derived from

$$y_{\mathcal{A}}(\tau) = x_0 + U_{x_0}\tau + U_{x_0}^{\perp} \mathcal{A}_{\phi_{x_0}}(\tau,\tau).$$

Because of $y_{\mathcal{A}}(\tau) - x(\tau) = O(\|\tau\|_2^3)$, we know $\mathcal{M}_{\mathcal{A}}$ approximates $\mathcal{M}$ well when τ_i is relatively small.

REMARK. In most of the manifold fitting cases, both of the true manifold $\mathcal{M}$ and the function $\phi_{x_0}(\tau)$ are unknown, as a result, it is impossible for us to get the second order parameter $\mathcal{A}_{x_0}$ through differentiate $\phi_{x_0}(\tau)$.

Instead of get the accurate $\phi_{x_0}(\tau)$, we can estimated it by using the data (observations) which is supposed to be drawn from some low-dimensional manifold. In this case, we need to estimate the $\mathcal{A}_{\phi_{x_0}}$ from the observations to obtain $\hat{\mathcal{A}}$, as a result, the range $y_{\hat{\mathcal{A}}}(\tau)$ can be assumed a local smoothed approximation of $\mathcal{M}$ around x_0 to some degree.

From now on, we will abbreviate $\hat{\mathcal{A}}$ as $\mathcal{A}$, to stand for the estimated second order parameter.

4.3. *Fitting Model.* In real case, when we have samples from the manifold, we need to determine $U_{x_0}, \mathcal{A}_{\phi_{x_0}}(0)$ by knowing $x_0, \{x_i\}$.

Using local principal analysis, we know for each observation x_i, there is a local coordinate $\tau_i = \tau(x_i, x_0, U_{x_0})$ in the tangent space. When we use x_0 as the origin of the coordinate. By projecting onto the tangent space, we have the local coordinate has the closed-form solution as

$$\tau_i = U_{x_0}^T(x_i - x_0). \tag{9}$$

When having $\{\tau_i, U_{x_0}, x_0\}$, we can determine the tensor $\mathcal{A}$ by a least square problem. The global coordinate x_i has a second order approximation as

$$x_i - (x_0 + U_{x_0}\tau_i + U_{x_0}^{\perp}\mathcal{A}(\tau_i,\tau_i)) = o(\|\tau_i\|_2^2). \tag{10}$$

We should find a tensor $\mathcal{A}$ such that the remainder is a higher order item. Substitute (9), into (10), we have

$$x_i - (x_0 + U_{x_0}\tau_i + U_{x_0}^{\perp}\mathcal{A}(\tau_i,\tau_i)) = U_{x_0}^{\perp}({U_{x_0}^{\perp}}^T(x_i - x_0) - \mathcal{A}(\tau_i,\tau_i)) = o(\|\tau_i\|_2^2)$$

which is equivalent to

$${U_{x_0}^{\perp}}^T(x_i - x_0) - \mathcal{A}(\tau_i,\tau_i) = o(\|\tau_i\|_2^2) \tag{11}$$

Noticing that (11) is a vector form, split (11) into each dimension, e.g, for the k-th dimension in the normal space, we have

$$(u_{x_0}^k)^T(x_i - x_0) = \mathcal{A}_{\cdot\cdot k}(\tau_i, \tau_i) + o(\|\tau_i\|_2^2) = \tau_i^T S^k \tau_i + o(\|\tau_i\|_2^2) \tag{12}$$

In (12), because of the symmetric position of τ_i, we know that the best fitted S^k is symmetric, i.e, each slice of $\mathcal{A}_{\cdot\cdot k}$ is a symmetric matrix. Therefore, the unknown parameters in S^k yields a total number of $d(d+1)/2$.

In (12), the only unknown is $\mathcal{A}$, finding the best approximation of $\mathcal{A}$ in the $\|\cdot\|_2^2$ criteria is a linear regression or least square problem. To use the linear algebra tools, we should vectorize each slice of $\mathcal{A}$.

In the following section, we will show how to vectorize each slice of $\mathcal{A}$ and how to make the vectorized result into the matrix form. Also, we show the quadratic form equals to the inner product of two vectors as

$$\tau_i^T S^k \tau_i = 2\mathrm{vech}(\tau_i\tau_i^T, 1)^T \mathrm{vech}(S^k, 1/2)$$

4.4. *Closed-form of S^k by Vectorization.* In this section, we vectorize the matrix, such that we could obtain S^k by using a least square problem with samples $\{x_i\}$.

For any symmetric matrix A, we know $A_{ij} = A_{ji}$. Therefore, we can only need $d(d+1)/2$ elements with the corresponding order to restore it. As a result, we can only vectorize the upper-triangle elements in the matrix A as a vector

$$\mathrm{vech}(A, t)_{\frac{(2d-i)(i-1)}{2}+j-i+1} = \begin{cases} A_{ij}, & j > i \\ tA_{ij}, & j = i \end{cases}$$

where the diagonal elements multiplied by a scalar t, which will bring us convenience for our following notations. When $t = 1$, the vector is constructed by picking the upper-triangle elements of A including the diagonal ones, i.e.,

$$\mathrm{vech}(A, 1) = [A_{11}, A_{12}, ...A_{1d}, ...A_{dd}]^T.$$

When $t = 1/2$, the vector is constructed by picking the upper-triangle elements of A, and half of the diagonal elements, i.e.,

$$\mathrm{vech}(A, 1/2) = [A_{11}/2, A_{12}, ...A_{1d}, ...A_{dd}/2]^T.$$

Note that, we can easily recover the matrix A from the vector $\mathrm{vech}(A, 1/2)$ by

$$A = \mathrm{Mat}(\mathrm{vech}(A, 1/2)) + \mathrm{Mat}^T(\mathrm{vech}(A, 1/2))$$

where $\mathrm{Mat}(y)$ is an operator constructed by realigning the elements in y into a upper-triangle matrix $\mathrm{Mat}(y)$ such that the i, j-th elements equals

$$\mathrm{Mat}(y)_{i,j} = \begin{cases} y_{(2d-i)(i-1)/2+j-i+1}, & j \geq i \\ 0, & j < i \end{cases}$$

With the above notations, the quadratic form $x^T Ax$ can be written in the vectorized version as

$$x^T Ax = 2\mathrm{vech}(xx^T, 1)^T \mathrm{vech}(A, 1/2)$$

where $\mathrm{vech}(xx^T, 1)$ is a vectorization of the symmetric matrix xx^T including the diagonal ones.

REMARK. The vectorization process for a symmetric matrix is invertible, i.e, if we vectorize a symmetric matrix into a vector, we can also turn the vector into a matrix which is identical with the former one.

Because of the symmetric of the matrix $\tau_i\tau_i^T$ and S^k, using the above notations, it can be easily verified that

$$\tau_i^T S^k \tau_i = \langle \tau_i\tau_i^T, S^k\rangle = 2\mathrm{vech}(\tau_i\tau_i^T, 1)^T \mathrm{vech}(S^k, 1/2)$$

For notational convenience, we denote $\theta_k = \mathrm{vech}(S^k, 1/2)$ and $g_i = \mathrm{vech}(\tau_i\tau_i^T, 1)$. Using the vector notations, the equation in (12) can be converted as:

$$g_i^T\theta_k - \frac{1}{2}(u_{x_0}^k)^T(x_i - x_0) = o(\|\tau_i\|_2^2). \tag{13}$$

For notational convenience, we use z_i^k to stand for the local coordinate in the k-th normal dimension corresponding to x_i, i.e.,

$$z_i^k = \frac{1}{2}(u_{x_0}^k)^T(x_i - x_0)$$

From (13), we know that

$$g_i^T\theta_k - z_i^k = o(\|\tau_i\|_2^2).$$

To determine the $d(d+1)/2$ parameters in θ_k, we need $d(d+1)/2$ linear independent equations. In other words, we need at least $d(d+1)/2$ samples to construct $d(d+1)/2$ linear independent $\{g_i, i = 1, ..., d(d+1)/2\}$.

Suppose, we have m samples on the manifold. Denote the matrix G and the vector ℓ_k as

$$G = [g_1, ..., g_m]^T, \quad \ell_k = [z_1^k, ..., z_m^k]^T.$$

REMARK. Idealy, we only need $d(d+1)/2$ samples to fix $\{S^k, k = d+1, D\}$ if and only if the rows of G can span the whole space of the $d(d+1)/2$-dimensional space, i.e. $\mathrm{rank}(G) = d(d+1)/2$. If we have less than $d(d+1)/2$ samples, we have multiple choice for each of the S^k in $\{S^k, k = d+1, ..., D\}$.

4.5. *Sample Related Weights.* Using the samples to estimate θ_k in (13), it should be noted that we want a locally fitting model, which means we want $\mathcal{M}_{\mathcal{A}}$ be defined as the range of $x_{\mathcal{A}}(\tau)$

$$\mathcal{M}_{\mathcal{A}} = \{x : x_{\mathcal{A}}(\tau) = U_\perp\mathcal{A}(\tau,\tau) + U\tau + x_0, \tau \in \mathbb{R}^d\}, \tag{14}$$

to fit $\mathcal{M}$ well when $\|\tau\|_2$ is relatively small. To achieve this goal, the points which reside nearby x_0 should have a larger weight. Then, by using a nonlinear kernel function $K_h(\cdot)$, the optimal minimization problem with respect to the k-th dimension becomes

$$\min_{\theta_k}\sum_{i=1}^m K_h(x_i - x_0)\{g_i^T\theta_k - \frac{1}{2}(u_{x_0}^k)^T(x_i - x_0)\}^2 = \|W_h^{1/2}(G\theta_k - \ell_k)\|_2^2, \tag{15}$$

where W_h is a diagonal matrix and the i-th element is $\{W_h\}_{ii} = K_h(x_i - x)$. Denote the column space $\mathcal{C} = \mathrm{span}\{W_h^{1/2}G_1, ..., W_h^{1/2}G_{d(d+1)/2}\}$

$$\|W_h^{1/2}(G\theta_k - \ell_k)\|_2^2 = \underbrace{\|W_h^{1/2}G\theta_k - P_{\mathcal{C}}(W_h^{1/2}\ell_k)\|_2^2}_{(e.1)} + \underbrace{\|P_{\mathcal{C}}^{\perp}(W_h^{1/2}\ell_k)\|_2^2}_{(e.2)} \tag{16}$$

REMARK. The term $(e.1)$ determines whether we have multiple choice of parameters corresponding the form of $x(\tau)$ which could yields the same least approximation error. The second term $(e.2)$ indicates whether $x(\tau)$ can go through the samples, if the term $(e.2)$ equals zero, we can infer that we can have $x(\tau)$ go through the samples exactly.

4.6. *Solution Behavior for S^k(or θ_k).* The behavior of $\psi(\tau)$ can be observed from the two parts in (16) as

Case i: $\|P_{\mathcal{C}}^{\perp}(W_h^{1/2}\ell_k)\|_2^2 = 0$. By evaluating $\|P_{\mathcal{C}}^{\perp}(W_h^{1/2}\ell_k)\|_2^2$, we can confirm whether $\psi(\tau)$ can go through the m samples. If $\|P_{\mathcal{C}}^{\perp}(W_h^{1/2}\ell_k)\|_2^2 = 0$, the minimum value of the problem (49) is zeros which implies $\phi_{\mathcal{A}}(\tau)$ pass the m samples on the manifold.

Case ii: $\mathrm{rank}(G) = d(d+1)/2$. The column-rank equals to the rank of G. If there is $\mathrm{rank}(G) = d(d+1)/2$, then, we know the columns of G are linear independent. Since $W_h^{1/2}$ is a full-rank matrix, the linear independent of G implies the linear independent of $W_h^{1/2}G$. As a consequence, there is a unique representation of $P_{\mathcal{C}}(W_h^{1/2}\ell_k)$ in the space spanned by the columns of $W_h^{1/2}G$, which is the parameter θ_k in (16). The unique solution of (49) is

$$\hat{\theta}_k = (G^T W_h G)^{-1} G^T W_h^{1/2} P_{\mathcal{C}}(W_h^{1/2}\ell_k) \tag{17}$$

Because there is an one-to-one corresponding between $\hat{S}^k$ and $\hat{\theta}_k$, the uniqueness solution of $\hat{\theta}_k$ implies we have a unique $\hat{S}^k$.

Case iii: $\mathrm{rank}(G) < d(d+1)/2$. The rank of G equals the column rank of $W_h^{1/2}G$ because $W_h^{1/2}$ is a full rank matrix. As a result, $\mathrm{rank}(G) < d(d+1)/2$ implies the columns of $W_h^{1/2}G$ are linear dependent. Therefore, the linear system

$$W_h^{1/2} G\alpha = 0, \tag{18}$$

has nonzero solutions. Since the dimension of the space $\mathcal{Q}$ spanned by the rows of $W_h^{1/2}G$ is $\mathrm{rank}(G)$, any vector in the orthogonal space $\mathcal{Q}^{\perp}$ is the solution of the linear system (18). Thus, in this case, $\hat{\theta}_k$ has multiple solutions, which can be written as:

$$\hat{\theta}_k = \tilde{\theta}_k + \sum_{u=1}^{r} c_r e_r$$

where $r = d(d+1)/2 - \mathrm{rank}(G)$, $\{c_r\}$ is any nonzero parameter and $\{e_r\}$ is the basis of the subspace of $\mathcal{Q}^{\perp}$. $\tilde{\theta}_k$ is any special solution for the linear system

$$W_h^{1/2} G\theta_k = P_{\mathcal{C}}(W_h^{1/2}\ell_k).$$

Because $P_{\mathcal{C}}(W_h^{1/2}\ell_k)$ is in the column space of $W_h^{1/2}G$, it is certain that there is some $\tilde{\theta}_k$ can satisfy the above equation.

4.7. *Perturbation Analysis.* In the above discussion, we assume x_0 is some point on the manifold. In real manifold fitting cases, we only have data points $\{x_i\}$ and some outlier x with we want to project onto the underline manifold $\mathcal{M}$. The process to estimate x_0 from x will be discussed in section 5.3. In the construction of $\{U, U_{\perp}\}$ only depending on the outlier $\bar{x}$ and the observation $\{x_i\}$.

As a result, we know x_0 can not be exactly on $\mathcal{M}$. Suppose the estimated x_0 be written as

$$x_0 = \tilde{x}_0 + e_0$$

where $\tilde{x}_0$ is the projection of x_0 onto $\mathcal{M}$ and e_0 is the approximation error in the normal space $\mathcal{N}_{\mathcal{M}}(\tilde{x}_0)$. The real relationship of (12) should be:

$$\tilde{z}_i^k = (u^k)^T (x_i - \tilde{x}_0) = \mathcal{A}_{\cdot\cdot k}(\tau_i, \tau_i) + o(\|\tau_i\|_2^2) \tag{19}$$

However, because of the unknown $\tilde{x}_0$, we approximate through the relation

$$z_i^k = (u^k)^T (x_i - x_0) = \mathcal{A}_{\cdot\cdot k}(\tau_i, \tau_i) + o(\|\tau_i\|_2^2) \tag{20}$$

Because of (19) and (20), we have the error is not related with the i-th sample, i.e,

$$\tilde{z}_i^k - z_i^k = (u^k)^T e_0.$$

Recalling the definition of ℓ_k, we have the true $\tilde{\ell}_k = [\tilde{z}_1^k, ..., \tilde{z}_m^k]$ and the $\ell_k = [z_1^k, ..., z_m^k]$ has a relation as

$$\tilde{\ell}_k = \ell_k + (u^k)^T e_0 \mathbb{1}.$$

The stability for the solutions corresponding to the two linear systems with respect toℓ_k and $\tilde{\ell}_k$ is highly related with the condition number of $G^T W_h G$

$$W_h^{1/2} G \tilde{\theta}_k = P_{\mathcal{C}}(W_h^{1/2} \tilde{\ell}_k); \quad W_h^{1/2} G \theta_k = P_{\mathcal{C}}(W_h^{1/2} \ell_k).$$

The error is

$$\tilde{\theta}_k - \theta_k = (G^T W_h G)^{-1} G^T W_h^{1/2} P_{\mathcal{C}}(W_h^{1/2} u^k e_0 \mathbb{1}).$$

4.8. *Geometric Interpretation of S^k.* Since S^k is a symmetric matrix, its eigenvalues all exist in real field. Here, we will show the definite property of S^k determines the shape of our fitted function. It is natural that we can give an analysis depending on the definite property of S^k.

Case i: positive (or negative) definite S^k. Note that, in our quadratic function of (14), the positive-define property of k-th slice S^k of $\mathcal{A}$ indicates that the quadratic term of the kth component satisfies for all $\tau \in \mathbb{R}^d$

$$\tau^T S^k \tau \geq 0.$$

Furthermore, if each of the slice S^k in $\mathcal{M}_{\mathcal{A}}$ is positive-definite, we can conclude that our quadratic surface $\mathcal{M}_{\mathcal{A}}$ resides only on one side of the tangent plane $\{x | x = U_d \tau + x_0\}$.

Case ii: indefinite S^k. Recall that in 3-dimensional space, the function of a saddle surface go through $(0, 0, 0)$ yields a form as

$$(x, y, z) = (\tau_1, \tau_2, \tau_1^2/a + \tau_2^2/b),$$

where $ab < 0$. Denote $\tau = (\tau_1, \tau_2)$. When we rotate the coordinate with an orthogonal matrix $[U, U_\perp]$ and move the origin $(0, 0)$ to x_0, the above parametric function becomes

$$(x(\tau), y(\tau), z(\tau)) = U\tau + U_\perp(\tau^T \begin{bmatrix} a, 0 \\ 0, b \end{bmatrix} \tau) + x_0.$$

Note that, in this case, the second order parameter $A = \begin{bmatrix} a, 0 \\ 0, b \end{bmatrix}$ is an indefinite matrix because of $ab < 0$.

For general case in higher dimension, an indefinite matrix S^k will result in a hyperbolic paraboloid in the corresponding dimension k in normal space. Suppose S^k yields a eigen-decomposition of the form

$$S^k = [V_1, V_2] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} [V_1, V_2]^T$$

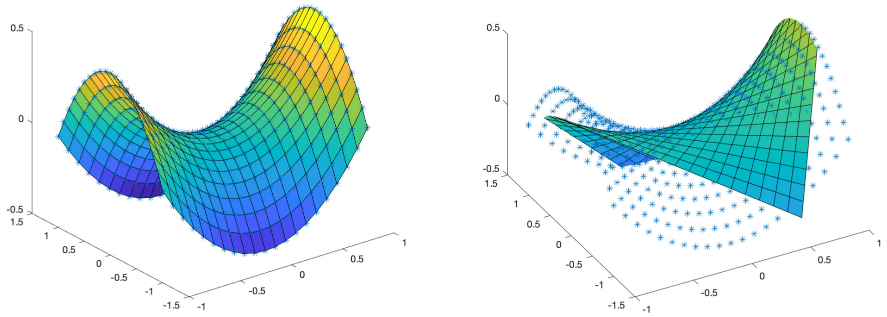

FIG 4. *Illustration of hyperbolic fitting case: Left: fitting at* $(0,0,0)$*, Right: fitting at* $(1/2,1/2,0)$*. Data drawn from function* $(x,y,z) = (\tau_1, \tau_2, \tau_1^2/2 - \tau_2^2/2)$ *without distrubed by noise:*

where the diagonal elements of Λ_1 are positive, and diagonal ones of Λ_2 are negative. Thus, we know the space spanned by the columns of $U_\perp V_1$ is second order positive, i.e., any direction restricted in this space is a subspace-restricted local minimum point. Conversely, the space spanned by the columns of $U_\perp V_2$ is second order negative, i.e., any direction restricted in this space is a subspace-restricted local maximum point. Thus, recalling that the second order surface is tangent to the first-order plane $\{x(\tau)|x(\tau) = x_0 + U\tau\}$ at $x = x_0$, the origin point x_0 is a saddle point.

In our manifold fitting problem, because we do not has any prior information on the manifold besides the observations, we do not restrict S^k to be a positive (or negative) matrix. The definite (or indefinite) property is decided by our fitting model and the inputed data.

4.9. *Asymptotic Properties.* In this section, we give the asymptotic properties of our fitted $\hat{S}^k$ and the true S^k (Hessian of $\phi_k(\tau)$ at 0) with the behaviors of n and h.

THEOREM 4.3. *For the estimated $\hat{\theta}_k$, the error between $\hat{\theta}_k$ and the true θ_k is with the order of*

$$\hat{\theta}_k - \theta_k = O(h) + O_p(\frac{1}{\sqrt{nh^{D-2}}})$$

where n is the number of observations and h is the bandwidth selected in our locally least square problem.

REMARK. Obviously, the optimum order of h is $h = O(\frac{1}{n^{1/D}})$. With the increasing of observations, we could choose a relatively small h such that the estimated error is relatively small.

5. Manifold Fitting via Nonlinear Projection.

In this section, we show how to implement a nonlinear projection onto our fitted manifold, also give some theoretical result to ensure the convergence properties.

The manifold fitting problem can be viewed as projecting data x to a local approximated structure. The works before considers this local structure as an affine plane (parameter with τ), such as

$$x(\tau) = U\tau + x_0$$

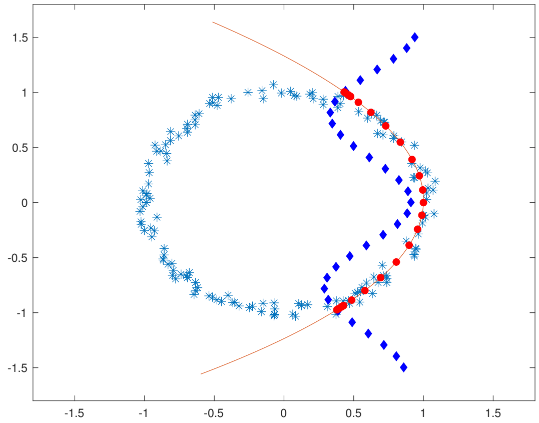

FIG 5. *Demo for Nonlinear Projection*

In our work, we can have a fitted paraboloid surface, with the form of a quadratic form as

$$x_{\mathcal{A}}(\tau) = U_{\perp}\mathcal{A}(\tau,\tau) + U\tau + x_0 \tag{21}$$

where x_0 is the origin (center for the local coordinate).

5.1. *Projection onto Manifold.* Suppose we have a locally fitted manifold which has the parameterized form written as $x(\tau):\mathbb{R}^d \to \mathbb{R}^D$. Since the range of the function $x_{\mathcal{A}}(\tau)$ generated a simple manifold, we denote it as

$$\mathcal{M}_{\mathcal{A}} = \{y : y = x_{\mathcal{A}}(\tau), \tau \in \mathbb{R}^d\}.$$

Given any $z \notin \mathcal{M}_{\mathcal{A}}$, we define the projection $P_{\mathcal{M}_{\mathcal{A}}}(z)$ of z onto $\mathcal{M}_{\mathcal{A}}$ as

$$P_{\mathcal{M}_{\mathcal{A}}}(z) = \arg\min_{w\in\mathcal{M}_{\mathcal{A}}} \|w - z\|_2. \tag{22}$$

Using the function $x_{\mathcal{A}}(\tau)$, we know the points on $\mathcal{M}_{\mathcal{A}}$ and the Euclidean space R^d is one-to-one. As a result, instead of find w, in (22), we just need to find τ

$$\hat{\tau} = \arg\min_{\tau\in R^d} \|z - x_{\mathcal{A}}(\tau)\|_2^2. \tag{23}$$

Finally, using the explicit form $x_{\mathcal{A}}(\tau)$, we know

$$P_{\mathcal{M}_{\mathcal{A}}}(z) = x_{\mathcal{A}}(\hat{\tau}) = x(\arg\min_{\tau\in R^d} \|z - x_{\mathcal{A}}(\tau)\|_2^2).$$

REMARK. In (23), we minimize the squared of 2-norm instead of 2-norm for two reasons. First, optimizing with respect to $\|\cdot\|_2^2$ and $\|\cdot\|_2$ is equivalent because of $f^2(x)$ and $f(x)$ share the same monotonous property in the intervals of $\{x|f(x) > 0\}$. Second, $\|\cdot\|_2^2$ is a polynomial, which has easier form with respect to derivatives for any order.

Overall, the projection onto $\mathcal{M}_{\mathcal{A}}$ consists of the following two steps:

1. Solve the nonlinear problem: Given z, find τ such as
$$\hat{\tau} = \min_{\tau} \|z - (U_{\perp}\mathcal{A}(\tau,\tau) + U\tau + x_0)\|_2^2 \tag{24}$$
2. Construct the coordinate in the ambient space
$$\hat{z} = U_{\perp}\mathcal{A}(\hat{\tau},\hat{\tau}) + U\hat{\tau} + x_0$$

To solve the nonlinear projection problem (24), we can simplify it as

$$\begin{aligned}
&\|z-(U_{\perp}\mathcal{A}(\tau,\tau)+U\tau+x_0)\|_2^2\\
=&\|P_U(z-x_0)-U\tau\|_2^2+\|P_{U_{\perp}}(z-x_0)-(U_{\perp}\mathcal{A}(\tau,\tau))\|_2^2\\
=&\|U(U^T(z-x_0)-\tau)\|_2^2+\|U_{\perp}(U_{\perp}^T(z-x_0)-\mathcal{A}(\tau,\tau))\|_2^2\\
=&\|U^T(z-x_0)-\tau\|_2^2+\|U_{\perp}^T(z-x_0)-\mathcal{A}(\tau,\tau)\|_2^2
\end{aligned} \tag{25}$$

REMARK. The optimization problem in (25) is nonlinear, with the highest order term corresponding to τ is quartic. As far as we know, there is no closed-form solution to directly minimize (25). The difficulty originates from the relatively higher order. If we can decrease the order with respect to τ to 2, we will have a closed form.

In next section, we show how to obtain the optimum τ via solving a series of quartic optimization problems.

5.2. *Quadratic Approximation.* In this section, we solve the quartic optimization problem by repeatedly implementing the quadratic approximation. Firstly, for notational convenience, let's denote

$$s=U^T(z-x_0)\in\mathbb{R}^d,\quad c=U_{\perp}^T(z-x_0)\in\mathbb{R}^{D-d}$$

where s,c can be obtained from solving the linear PCA at x_0. Then the subproblem (25) can be converted into:

$$f(\tau)=\|s-\tau\|_2^2+\|c-\mathcal{A}(\tau,\tau)\|_2^2 \tag{26}$$

Let the auxilliary function $g(\tau_1,\tau_2)$ defined as

$$g(\tau_1,\tau_2)=\frac{1}{2}\|s-\tau_1\|_2^2+\frac{1}{2}\|s-\tau_2\|_2^2+\|c-\mathcal{A}(\tau_1,\tau_2)\|_2^2$$

Because of the symmetric property of $g(\tau_1,\tau_2)$, we have

$$g(\tau_1,\tau_2)=g(\tau_2,\tau_1).$$

Also, $g(\tau_1,\tau_2)$ is a good approximation of $f(\tau)$, as $g(\tau,\tau)=f(\tau)$. We could derive a Cauchy sequence of $\{\tau_n,n=1,2...,\}$ as

$$\tau_{n+1}=\arg\min_{\tau}g(\tau_n,\tau)=\frac{1}{2}\|s-\tau_n\|_2^2+\frac{1}{2}\|s-\tau\|_2^2+\|c-\mathcal{A}(\tau_n)\tau\|_2^2. \tag{27}$$

For notational convenience, define $\phi(\tau_n)$ as the optimal minimum solution of (27), i.e,

$$\phi(\tau_n)=(2\mathcal{A}(\tau_n)\mathcal{A}(\tau_n)^T+I_d)^{-1}(2\mathcal{A}(\tau_n)c+s).$$

Next, we give the Theorem 5.1 by showing that $\phi(\tau_n)$ is the global minimum for the quadratic function $g(\tau_n,\tau)$.

THEOREM 5.1. *The global minimum point of $g(\tau_n,\tau)$ is $\tau=\phi(\tau_n)$.*

PROOF. Take the derivative of $g(\tau_n,\tau)$ with respect to τ, $\nabla_{\tau}g(\tau_n,\tau)=0$ for τ leads to

$$\nabla_{\tau}g(\tau_n,\tau)=2(\mathcal{A}(\tau_n)\mathcal{A}(\tau_n,\tau)-\mathcal{A}(\tau_n)c)+(\tau-s).$$

Letting $\nabla_{\tau}g(\tau_n,\tau)=0$ and solve the linear equation with respect to τ, we obtain that

$$\tau=(2\mathcal{A}(\tau_n)\mathcal{A}(\tau_n)^T+I_d)^{-1}(2\mathcal{A}(\tau_n)c+s)$$

Taking the twice differential operation with $g(\tau_n,\tau)$, we obtain the Hessian $H_g(\tau)$ as

$$H_g(\tau) = \nabla_\tau \nabla_\tau g(\tau_n,\tau) = 2\mathcal{A}(\tau_n)\mathcal{A}^T(\tau_n) + I_d.$$

Since $H_g(\tau)$ is a constant positive definite matrix which does not rely on τ, thus, it implies that $\tau = \phi(\tau_n)$ is the global minimum of $g(\tau_n,\tau)$. □

Using $\phi(\tau_n)$, we can define a vector sequence $\{\tau_n, n = 1, 2..., \}$ via the fixed point iteration by $\tau_{n+1} = \phi(\tau_n)$. Next, we give the convergence property of the sequence $\{\tau_n, n = 1, 2..., \}$.

THEOREM 5.2. *The sequence $\{a_n = g(\tau_n, \tau_n + 1), n = 1, 2..., \infty\}$ is monotonously decreasing with a positive lower-bound, thus it is a convergent sequence!*

PROOF. Using the symmetric property of function g, we have:

$$g(\tau_n,\tau_{n-1}) = g(\tau_{n-1},\tau_n). \tag{28}$$

Meanwhile, because of the optimal minimizing relation $\tau_{n+1} = \arg\min g(\tau_n,\tau)$, we know

$$g(\tau_n,\tau_{n+1}) \leq g(\tau_n,\tau_{n-1}). \tag{29}$$

Combining the above two relations in (28),(29), we have:

$$g(\tau_n,\tau_{n+1}) \leq g(\tau_n,\tau_{n-1}) = g(\tau_{n-1},\tau_n).$$

Therefore, the sequence $\{g(\tau_n,\tau_{n+1})\}$ decreases with n. Also $\{g(\tau_n,\tau_{n+1})\}$ has a lower bound because it is positive. Using the monotonous-bound theorem, we know the sequence $\{g(\tau_n,\tau_{n+1})\}$ converges! □

Because the sequence $\{g(\tau_n,\tau_{n+1})\}$ converge, denote the unique limit of the sequence as

$$\gamma = \lim_{n\to\infty} g(\tau_n,\tau_{n+1})$$

Next, we will show, under some conditions, γ is also the minimum of $f(\tau)$. This result relies on the convergence property of the vector sequence $\{\tau_n, n = 1, ..., \infty\}$. As the sequence of τ_n is generated from the fixed point iteration, to guarantee the convergence, we can give a stricker condition by requiring $\phi(\tau)$ to be a contraction map.

Similarly with the matrix, define the norm of the tensor as

$$\|\mathcal{A}\|_2 = \max_{\|c\|=1} \|\mathcal{A}(c)\|_2 = \max_{\|c\|=1, \|\tau\|=1} \mathcal{A}(\tau,\tau,c)$$

THEOREM 5.3. *If $L = (4\beta^2\|\mathcal{A}\|_2^3\|c\|_2 + 2\|s\|_2\|\mathcal{A}\|_2^2) < 1$, then, the function $\phi(\tau)$ is a contraction map, i.e,*

$$\|\phi(\tau_{n-1}) - \phi(\tau_n)\|_2 \leq L\|\tau_{n-1} - \tau_n\|_2,$$

where β is the upper bound of $\|\tau_n\|_2$, i.e., $\beta = \sup_n \|\tau_n\|_2$.

The proof is left in the appendix A.3. Recall $s = U^T(z - x_0) \in \mathbb{R}^d$ is the local coordinate of z in the tangent space. $c = U_\perp^T(z - x_0) \in \mathbb{R}^{D-d}$ is the coordinate in the normal space. When we select a good origin x_0, we can have the norm of s and c to be in small scale simultaneously, in order to make L as small as possible.

REMARK. $\|\mathcal{A}\|_2$ represents the maximum curvature of underlining manifold which can be obtained through solving the fitting model from the observations. It can be seen that if the manifold becomes more and more flatter, L will turn more smaller which will makes the requirement of our assumption easily to be satisfied.

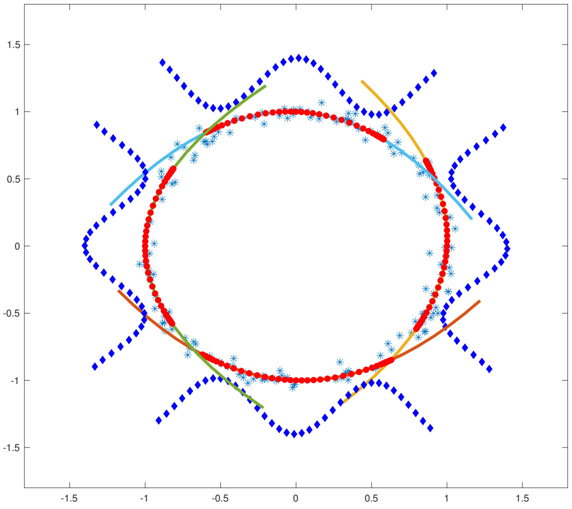

FIG 6. *Projecting the outlier points onto the fitted manifold*

REMARK. The contraction map $\phi(\tau)$ is important for us to show $\{\tau_n, n = 1,2,...\}$ is a Cauchy sequence, furthermore, it converges to some point τ^*.

THEOREM 5.4. *The convergence point τ^* of $\{\tau_n, n = 1,2....\}$ is a stationary point of $f(\tau)$. Furthermore, if*

$$I_d + 2\mathcal{A}(\mathcal{A}(\tau^*,\tau^*) - c) + 4\mathcal{A}(\tau^*)\mathcal{A}^T(\tau^*)$$

is positive definite, τ^* is also an optimal minimal of $f(\tau)$.

PROOF. Since $\{\tau_n, n = 1,2...\}$ is a Cauchy sequence, taking the limit of n leads to

$$\lim_{n\to\infty} g(\tau_n,\tau_{n-1}) = g(\tau^*,\tau^*) = f(\tau^*).$$

From the optimal of $\tau_{n+1} = \min_\tau g(\tau_n,\tau)$, we have $\nabla_\tau g(\tau_n,\tau)|_{\tau=\tau_{n+1}} = 0$, i.e.,

$$2(\mathcal{A}(\tau_n)\mathcal{A}(\tau_n,\tau_{n+1}) - \mathcal{A}(\tau_n)c) + (\tau_{n+1} - s) = 0 \tag{30}$$

In (30), taking the limit with respect to n, we have $\tau^* = \lim \tau_n$ which satisfying the normal equation as

$$2(\mathcal{A}(\tau^*)\mathcal{A}(\tau^*,\tau^*) - \mathcal{A}(\tau^*)c) + (\tau^* - s) = 0 \tag{31}$$

Recall that $\nabla f(\tau) = 2(\tau - s) + 4(\mathcal{A}(\tau,\tau) - c)\mathcal{A}(\tau)$. Obviously, we know τ^* also satisfies the KKT condition for $f(\tau)$, i.e., $\nabla f(\tau^*) = 0$. Thus, τ^* is a stationary point $f(\tau)$. Next, to check τ^* is a local minimum, we need to derive the second order derivative, $H_f(\tau)$,

$$H_f(\tau) = 2I + 4\mathcal{A}(\mathcal{A}(\tau,\tau) - c) + 8\mathcal{A}(\tau)\mathcal{A}^T(\tau).$$

Clearly, τ^* is an optimal minimum of $f(\tau)$ if and only if $H_f(\tau^*)$ is a positive definite matrix. □

THEOREM 5.5. *Under the same fitting error tolerance ϵ, the numbers of the fitting structure needed for $x_\ell(\tau)$ and $x_{\mathcal{A}}(\tau)$ yields*

$$\frac{N(x_\ell(\tau),\mathcal{M},\epsilon)}{N(x_{\mathcal{A}}(\tau),\mathcal{M},\epsilon)} = O(\epsilon^{-d/6})$$

PROOF. Define $\mathcal{M}(x_0,\delta)=\{x|x\in\mathcal{M}, d(x,x_0)\le\delta\}$., For $x\in\mathcal{M}\cap\mathcal{D}_{x_0}(\delta)$, we know:

$$d(x,\hat{\mathcal{M}}(x_{\mathcal{A}}(\tau),x_0)) = O(\|\tau\|_2^3)$$

Similarly, for the linear fitting function $x_\ell(\tau)=U_d\tau+x_0$ of $\mathcal{M}$, we have the projection error is:

$$d(x,\hat{\mathcal{M}}(x_\ell(\tau),x_0)) = O(\|\tau\|_2^2)$$

where $\tau=U_d^T(x-x_0)$ which is the local coordinate of x. If we set the maximum tolerance to be

$$d(x,\hat{\mathcal{M}}(x_{\mathcal{A}}(\tau),x_0))\le\epsilon, \quad d(x,\hat{\mathcal{M}}(x_\ell(\tau),x_0))\le\epsilon,$$

the maximum radius is of order $\epsilon^{1/3}$ and $\epsilon^{1/2}$ for $x_{\mathcal{A}}(\tau)$ and $x_\ell(\tau)$ respectively. The corresponding volume of

$$\mathrm{Vol}\{x|x\in\mathcal{M}, d(x,\hat{\mathcal{M}}(x_\ell(\tau),x_0))\le\epsilon\} = O(\epsilon^{d/2})$$
$$\mathrm{Vol}\{x|x\in\mathcal{M}, d(x,\hat{\mathcal{M}}(x_{\mathcal{A}}(\tau),x_0))\le\epsilon\} = O(\epsilon^{d/3})$$

As a result, the number of the fitting structure needed is $O(\epsilon^{-d/2})$ and $O(\epsilon^{-d/3})$ for $x_\ell(\tau)$ and $x_{\mathcal{A}}(\tau)$ respectively. □

THEOREM 5.6. *For any x, denote the projection of x onto $\mathcal{M}_{\mathcal{A}}$ as x^* and correspondingly there is the local corrdinate τ^* such that $x(\tau^*)=x^*$, then we have*

$$x-x^*\perp\mathcal{T}_{\mathcal{M}_{\mathcal{A}}}(\tau^*)$$

PROOF. Recall $x_{\mathcal{A}}(\tau)=U_\perp\mathcal{A}(\tau,\tau)+U(\tau)+x_0$, then the tangent space $\mathcal{S}^*$ at $x_{\mathcal{A}}(\tau)|_{\tau^*}$ is

$$\mathcal{S}^*=\mathrm{span}\{\nabla_{\tau_1}x_{\mathcal{A}}(\tau)|_{\tau=\tau^*}, \nabla_{\tau_2}x_{\mathcal{A}}(\tau)|_{\tau=\tau^*},\dots,\nabla_{\tau_d}x_{\mathcal{A}}(\tau)|_{\tau=\tau^*}\},$$

which corresponding to each columns of the Jacobi matrix $J(x_{\mathcal{A}}(\tau))|_{\tau^*}$. Note that, the Jacobi $J(x_{\mathcal{A}}(\tau))|_{\tau^*}$ is

$$J(x_{\mathcal{A}}(\tau))|_{\tau=\tau^*}=2U_\perp\mathcal{A}(\tau^*)+U.$$

$\mathcal{S}^*$ is also the space spanned by the columns of $J(x_{\mathcal{A}}(\tau))|_{\tau=\tau^*}$. To prove $x-x^*\perp\mathcal{T}_{\mathcal{M}_{\mathcal{A}}}(x^*)$, we just need to prove $x-x^*$ is orthogonal with each of the columns of $J(x_{\mathcal{A}}(\tau))|_{\tau=\tau^*}$. Next, we give the form of $x-x^*$, recall that

$$\begin{aligned}x-x^* =& x-(U_\perp\mathcal{A}(\tau^*,\tau^*)+U\tau^*+x_0)\\ =& U_\perp(U_\perp^T(x-x_0)-\mathcal{A}(\tau^*,\tau^*))+U(U^T(x-x_0)-\tau^*)\\ =& U_\perp(c_x-\mathcal{A}(\tau^*,\tau^*))+U(s_x-\tau^*).\end{aligned}$$

It is easy to verify

$$(x-x^*)^TJ(x(\tau))|_{\tau=\tau^*}=-\{(\mathcal{A}(\tau^*)\mathcal{A}(\tau^*,\tau^*)-\mathcal{A}(\tau^*)c_x)+(\tau^*-s_x)\}=0,$$

where the last equality is obtained from the optimal condition in (31), that is to say, $x-x^*$ is orthogonal with the space $\mathcal{S}^*$. □

5.3. *Origin Point Selection.* The selection of the origin point x_0 is important for our fitting. If our observations $\{x_i\}$ is sampled exactly from $\mathcal{M}$ (no noise contaminated), we can select x_0 to be the point which is nearest to $\bar{x}$.

$$x_0 = \arg\min_{\{x_i\}} \|\bar{x} - x_i\|_2$$

Otherwise, if the observations $\{x_i\}$ are drawn from $\mathcal{M}$ with noise, we can select x_0 as the weighted-mean as:

$$x_0 = \frac{1}{\sum_{i\in\mathcal{N}_k(\bar{x})} \phi_h(\bar{x}, x_i)} \sum_{i\in\mathcal{N}_k(\bar{x})} \phi_h(\bar{x}, x_i) x_i \tag{32}$$

where $\phi_h(\bar{x}, x_i) = K_h(x_i, \bar{x})$ and $\mathcal{N}_k$ controls the neighbor size and h is the bandwidth parameter, which affect the bias and smoothness. If we further denote

$$w_h(\bar{x}, x_i) = \phi_h(\bar{x}, x_i)/(\sum_{i\in\mathcal{N}_k(\bar{x})} \phi_h(\bar{x}, x_i))$$

From (32), we see that x_0 is a convex combination of samples $\{x_i\}$ in the neighborhood $\mathcal{N}_k(\bar{x})$ of $\bar{x}$ as $x_0 = \sum_{i\in\mathcal{N}_k(\bar{x})} w_h(\bar{x}, x_i) x_i$

ASSUMPTION 5.7. *Assume the observations are drawn from some manifold, which is parameterized by the form of*

$$x(\tau) = U\tau + U_\perp \mathcal{A}(\tau, \tau) + x_0^*.$$

Thus, $x_i = U\tau_i + U_\perp \mathcal{A}(\tau_i, \tau_i) + x_0^* + \sigma\epsilon_i$, where ϵ_i is the independent noise vector such that $\epsilon_i \sim \mathcal{N}(0, I)$ and σ is the scale of the disturbed noise.

THEOREM 5.8. *For any x_0 defined as a summation of point in* (32)*, the squared distance from x_0 to the manifold $\mathcal{M}_\mathcal{A}$ is upper-bounded by*

$$\min_{y\in\mathcal{M}_\mathcal{A}} d(x_0, y) = \|\mathcal{A}\|_2^2 O(h^4)$$

The proof is left in the appendix (A.2).

6. Algorithm: Projection by Repeated Nonlinear Least Square. Because the underlying manifold is unknown, in real computational cases, we cannot find a point $x_0 \in \mathcal{M}$ which is also next to our interested outlier $\bar{x}$. To solve this problem, we use an iteration method to find x_0. When $\bar{x}$ is far away from $\mathcal{M}$, the inaccurate of x_0 is also acceptable. With the process of $\bar{x}$ approaching $\mathcal{M}$, we need the accuracy of x_0 to be improved.

Our algorithm is consisted of steps by repeatedly implementing the fitting and projection procedures.

Repeat: from A1 to B3

The fitting procedure contains the following four steps:

A1. For outlier $\bar{x}$, compute the shift mean x_0 from (32) and use x_0 as the origin of our local coordinate to implement our fitting and projection process.

A2. Given x_0, neighborhood size k and bandwidth parameters h, get the local coordinate $\{\tau_i, \iota_i\}$ for each x_i by applying the eigenvalue decomposition.

$$\tau_i = U_d^T (x_i - x_0).$$

Using τ_i. construct G from $\{\tau_i\}$ as $G = [g_1, ..., g_m]^T$, where each $g_i = \text{vech}(\tau_i \tau_i^T, 1)$.

A3. Solve the manifold fitting problem, for the k-th dimension in the normal space:

$$\min_{\theta_k}\sum_{i=1}^{m} K_h(x_i-\bar{x})\{g_i^T\theta_k-(u_{x_0}^k)^T(x_i-x_0)/2\}^2=\|W_h^{1/2}(G\theta_k-\ell_k/2)\|_2^2,$$

where the m dimensional vector ℓ_k equals to $[u_k^T(x_1-x_0), u_k^T(x_1-x_0), ..., u_k^T(x_n-x_0)]$.

A4. Transform vector θ_k into matrix by putting the elements of θ_k onto the upper-diagonal:

$$S_k=\mathrm{Mat}(\theta_k)+\mathrm{Mat}(\theta_k)^T,$$

By aligning each slice of S_k, we obtain the tensor $\mathcal{A}$, i.e., $\mathcal{A}_{\cdot\cdot k}=S_k$. Here, we get a simple manifold $\mathcal{M}_\mathcal{A}$ to fit the complicated $\mathcal{M}$, where our simple $\mathcal{M}_\mathcal{A}$ yields a form as

$$x(\tau)=U_d^\perp\mathcal{A}(\tau,\tau)+U_d\tau+x_0$$

The projection procedure contains the following three steps:

B1. For an outlier $\bar{x}$, set $\tau_0=U_d^T(\bar{x}-x_0)$ and apply the fix point iteration, to get the convergence point τ^* of the sequence $\{\tau_n\}$. The fix point iteration is:

$$\phi(\tau_n)=(2\mathcal{A}(\tau_n)\mathcal{A}(\tau_n)^T+I_d)^{-1}(2\mathcal{A}(\tau_n)c+s).$$

B2. Put τ^* onto the fitted function to obtain the point $\hat{x}$, which is the projection of $\bar{x}$ onto $\mathcal{M}_\mathcal{A}$ as

$$\hat{x}=P_{\mathcal{M}_\mathcal{A}}(x)=U_\perp\mathcal{A}(\tau^*,\tau^*)+U\tau^*+x_0$$

B3. Check whether $\|\hat{x}-\bar{x}\|\leq\epsilon$, if true, stop and output $\hat{x}$, otherwise set $\bar{x}=\hat{x}$ and repeat the steps from A1 to B3.

7. Higher-degree extension. In this section, we give a third-order generalization our manifold fitting algorithm and even higher order generalization approach is similar with the generalization from order-two to order-three.

We split the extension into two parts, fitting a higher order $\mathcal{M}_\mathcal{A}$, and the projection by solving the nonlinear least square problem details. We show that the higher-degree fitting corresponds to solve a linear least square with more variables than before.

7.1. *Higher-order manifold fitting.* Recall that for any manifold $\mathcal{M}$, there is a corresponding $\phi(\tau)$ such that

$$x(\tau)=x_0+U\tau+U_\perp\phi(\tau,\tau).$$

In the above discussion, we approximate $\phi(\tau,\tau)$ by a quadratic form $\mathcal{A}(\tau_i,\tau_i)$. Besides the quadratic term we also can approximate $\phi(\tau,\tau)$ with the higher order such as

$$\phi(\tau,\tau)\approx\mathcal{A}_2(\tau_i,\tau_i)+\mathcal{A}_3(\tau_i,\tau_i,\tau_i),$$

where $\mathcal{A}_2$ and $\mathcal{A}_3$ are the third and forth order tensor of shape $d\times d\times(D-d)$ and $d\times d\times d\times(D-d)$, respectively. Similarly, we can also define g_i by vectorizing $\tau_i\tau_i^T$ and $\tau_i\otimes\tau_i\otimes\tau_i$. Note that, because of the symmetric property of the tensor, for the k-th slice of $\mathcal{A}_2$ and $\mathcal{A}_3$ there are $d(d+1)/2$ and $d(d^2+1)/2$ free parameters to be determined. In the third order approximation, both of θ_k and g_i are with the total number of free parameters $(d^3+d^2+2d)/2$. Similarly, with the samples $\{x_i\}$, we have

$$\min_{\theta_k}\sum_{i=1}^{m} K_h(x_i-x_0)\{g_i^T\theta_k-\frac{1}{2}(u_{x_0}^k)^T(x_i-x_0)\}^2=\|W_h^{1/2}(G\theta_k-\ell_k)\|_2^2,$$

By realigning the elements in θ_k, we can obtain the k-th slice for the tensors $\mathcal{A}_2$ and $\mathcal{A}_3$.

7.2. *Projection.* Since $\mathcal{M}_\mathcal{A}$ is a smooth approximated manifold, the projection onto the fitted manifold $\mathcal{M}_\mathcal{A}$ is equivalent to solve the minimization problem

$$\begin{aligned}&\min_\tau \|z - U_\perp((\mathcal{A}_2(\tau,\tau) + \mathcal{A}_3(\tau,\tau,\tau)) + U\tau + x_0)\|_2^2 \\ =&\min_\tau \|U^T(z-x_0) - \tau\|_2^2 + \|U_\perp^T(z-x_0) - \mathcal{A}_2(\tau,\tau) - \mathcal{A}_3(\tau,\tau,\tau)\|_2^2.\end{aligned} \tag{33}$$

Denote $s = U^T(z-x_0), c = U_\perp^T(z-x_0)$, the problem (33) turns to

$$\min_\tau g(\tau) = \|s-\tau\|_2^2 + \|c - \mathcal{A}_2(\tau,\tau) - \mathcal{A}_3(\tau,\tau,\tau)\|_2^2.$$

Define the auxiliary quadratic function $f(\tau_1,\tau_2)$ to approximate $g(\tau)$ as

$$f(\tau_1,\tau_2) = \frac{1}{2}\|s-\tau_1\|_2^2 + \frac{1}{2}\|s-\tau_2\|_2^2 + \|c - \mathcal{A}_2(\tau_1,\tau_2) - \mathcal{A}_3(\tau_1,\tau_1,\tau_2)\|_2^2.$$

Then, the alternating minimization iteration will become

$$\tau_{n+1} = \arg\min_\tau \frac{1}{2}\|s-\tau\|_2^2 + \|c - \mathcal{A}_2(\tau_n,\tau) - \mathcal{A}_3(\tau_n,\tau_n,\tau)\|_2^2. \tag{34}$$

Since $\mathcal{A}_2(\tau_n,\tau)$ and $\mathcal{A}_3(\tau_n,\tau_n,\tau)$ can also be written as the matrix-vector multiplication as

$$\mathcal{A}_2(\tau_n,\tau) = \mathcal{A}_2(\tau_n)^T\tau, \quad \mathcal{A}_3(\tau_n,\tau_n,\tau) = \mathcal{A}_3(\tau_n,\tau_n)^T\tau.$$

Note that, both of $\mathcal{A}_2(\tau_n)$ and $\mathcal{A}_3(\tau_n,\tau_n)$ are the $d\times(D-d)$ matrix , then, the above minimization problem (34) has the closed-form solution as

$$\phi(\tau_n) = (2\mathcal{A}_2(\tau_n)\mathcal{A}(\tau_n)^T + 2\mathcal{A}_3(\tau_n,\tau_n)\mathcal{A}_3(\tau_n,\tau_n)^T + I_d)^{-1}(2(\mathcal{A}_2(\tau_n) + \mathcal{A}_3(\tau_n,\tau_n))c + s).$$

Using $\phi(\tau_n)$, we can also build the fix point iteration as $\tau_{n+1} = \phi(\tau_n)$ and obtain the convergence point as the projection onto the third-order manifold fitted function.

From analyzing the two steps of fitting and projection, we can see our model can be easily generalized into a higher-order approximation form. However, because of the number of unknown parameters increases with the speed of d^s, where s is the order and d is the dimension of the tangent space, we need large amount of effective data to fix the higher-order model. Otherwise, too small dataset and quit complicate model will lead to the overfitting problem, which will also diminish the performance of our algorithm.

8. Simulation.

In this section, we compare our nonlinear manifold fitting approach with the existing manifold fitting methods on various occasions. We consider the manifolds having constant curvature and varying curvature. The numerical results show our model can handle all the cases by leading a very promising result.

8.1. *Data Recovery Capability.* In this section, we construct some artificial manifolds e.g triangle cure, circle and the swiss-roll, which can be written as a parameterization form as $\begin{cases} x = \phi(t) \\ y = \psi(t)\end{cases}$ Thus, we have the tangent space is spanned by $(\phi'(t),\psi'(t))$ and the normal space is spanned by the vector $(\psi'(t), -\phi'(t))$. The curvature at t can be calculated as

$$\kappa(t) = \frac{\phi'(t)\psi''(t) - \phi''(t)\psi'(t)}{((\phi'(t))^2 + \psi'(t)^2)^{3/2}} \tag{35}$$

Especially, when the curve parameterized as $(x, f(x))$, the curvature yields quite a simple form as $\kappa(x) = \frac{|f''(x)|}{(1+f'(x))^{3/2}}$. Next, we give three examples and show the performance of our algorithm.

EXAMPLE 2. By substituting the derivatives of the 2-D circle's parametric equation into (35). we know that the curvature of the circle $\begin{cases} x = \cos(\theta) \\ y = \sin(\theta) \end{cases}$ is the constant $\kappa(\theta) = 1$ everywhere,

EXAMPLE 3. For the swiss-roll in the 2-D space, the parametric equation is $\begin{cases} x = \theta\cos(\theta) \\ y = \theta\sin(\theta) \end{cases}$. Thus, we have the first derivatives and the second derivatives:

$$\begin{aligned} &x'(\theta) = \cos(\theta) - \theta\sin(\theta), \quad y'(\theta) = \sin\theta + \theta\cos\theta \\ &x''(\theta) = -2\sin\theta - \theta\cos\theta, \quad y''(\theta) = 2\cos(\theta) - \theta\sin(\theta). \end{aligned} \tag{36}$$

Substitute (36) into the equation of curvature, we have $\kappa(\theta) = \frac{2+\theta^2}{(1+\theta^2)^{3/2}}$ and $\kappa'(\theta) = \frac{-\theta(\theta^2+4)}{(1+\theta^2)^{5/2}}$. For any $\theta \in \mathbb{R}^+$, the curvature decreases with the increasing of θ.

EXAMPLE 4. Recalling the function of the triangle curve $\begin{cases} x(\theta) = \theta \\ y(\theta) = \sin\theta \end{cases}$ and taking the derivatives with respect to x, we know that

$$x'(\theta) = 1, \quad y'(\theta) = \cos(\theta), \quad x''(\theta) = 0, \quad y''(\theta) = -\sin(\theta)$$

By substituting the derivatives into the equation of the curvature, we know $\kappa(\theta)$ of the triangle curve $(\theta, \sin\theta)$ at θ is $\kappa(\theta) = \frac{|\sin(\theta)|}{(1+\cos(\theta))^{3/2}}$, which achieves the maximum $\kappa(\theta) = 1$ when $\theta = k\pi + \frac{1}{2}\pi, k \in Z$ and $\kappa(\theta)$ achieves the minimum $\kappa(\theta) = 0$ when $\theta = k\pi, k \in Z$, which could also be observed from the curve of the function.

REMARK. The first 2-D circle example is a curve with constant curvature and the remaining two examples are curves with varying curvature. The swiss-roll is an example which has the curvature monotonously decreasing with the parameter θ. The triangle curve is an example which has periodic curvature with respect to the parameter.

It is difficult to fit the data drawn from a manifold which is assumed to have varying curvature using a linear model or a tangent plane. Because for x in an area which leads $\kappa(\theta)$ small, the manifold can be approximated by a low-dimensional affine space (a flat plane). However, in a area which has the large curvature $\kappa(\theta)$, such as $\theta = \pi/2$ in our triangle curve case, it is impossible to approximate $y = \sin(\theta)$ by a flat structure. Therefore, it is quit necessary to use a more complicated model to approximate.

Noting that, our nonlinear second-order fitting function $x_{\mathcal{A}}$ is a generalization of the linear function x_ℓ. When the second order parameter $\mathcal{A}$ equal to zero, the derived manifold $\mathcal{M}_{\mathcal{A}}$ will degenerate into $\mathcal{M}_\ell$. By learning $\mathcal{A}$, we automatically considering the curvature information hidden in the manifold.

Since our nonlinear second-order fitting model $\mathcal{M}_{\mathcal{A}}$ is more complicated by having more parameters compared with $\mathcal{M}_\ell$, we need to use more data to solve our model. Solving the least square problem with too few data will lead to a phenomenon called 'overfitting'. When the overfitting phenomenon occurs, the model will not only learn from the true signal of the underlining manifold, but also the noise factors which will distort our model.

In our fitting model, the number of points used is controlled by the bandwidth parameter h. Because the kernel weight is decreased fast with the radius, the contribution of the points resides far from our interested area is ignorable. As a result, our model works well with a relatively larger h, which can be seen in the rightmost partition of Figure (7).

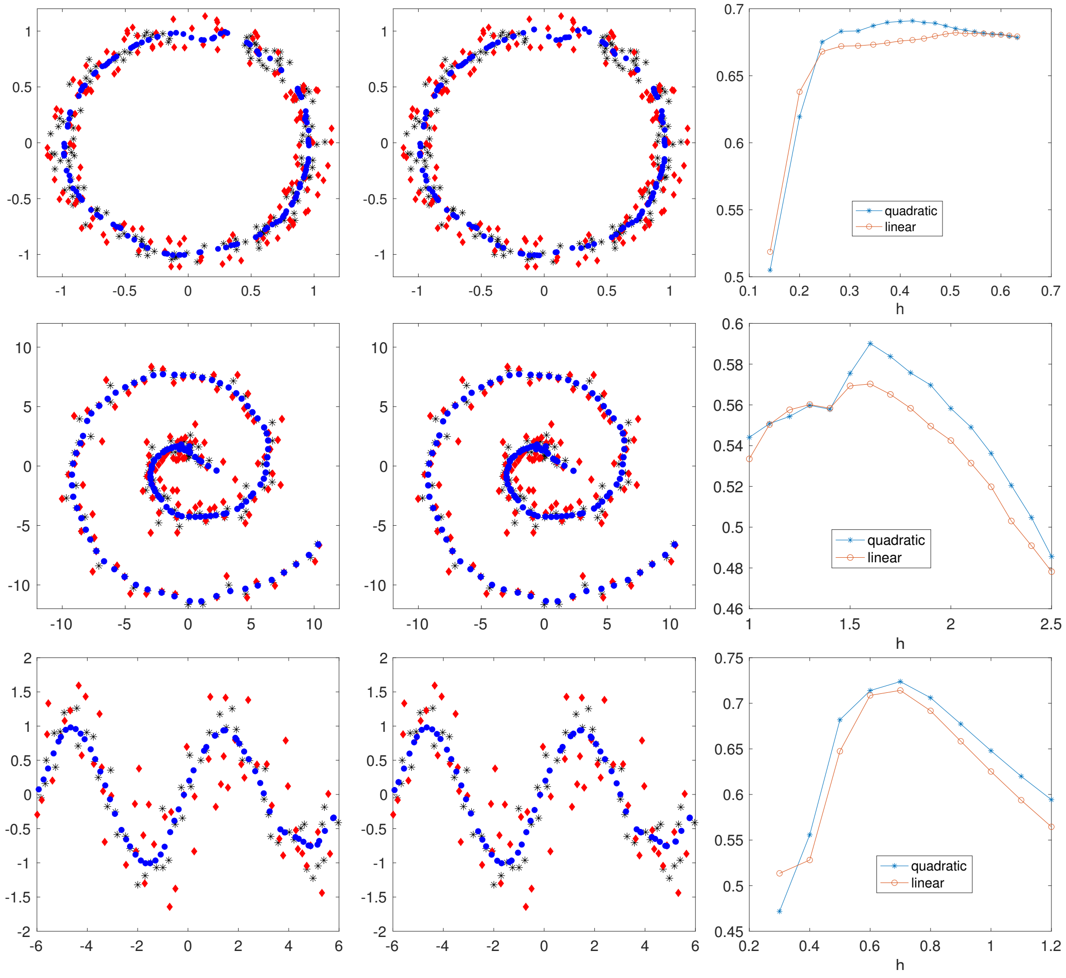

FIG 7. *Comparison between nonlinear projection and linear projection. Left: nonlinear approach. Right: linear approach*

In Figure (7), we randomly samples the black stars '$*$' as $x_i = \tilde{x}_i + \sigma_1 \epsilon_i$ and the red diamonds '$\diamond$' as $x_i = \tilde{x}_i + \sigma_2 \epsilon_i$, where $\tilde{x}_i$ is on the underlining $\mathcal{M}$ and ϵ_i is in the normal space of $\mathcal{M}$ at $\tilde{x}_i$. In our experimental setting, we have $\sigma_1 = 0.2$ and $\sigma_2 = 0.5$ and the length of ϵ_i obeys the normal distribution of $N(0,1)$. The blue dots '$\circ$' represent the result, which is the projection onto the fitted structure. The leftmost figure is the result obtained from the projection onto the nonlinear $\mathcal{M}_{\mathcal{A}}$ and the middle figure stands for the result obtained from the projection onto $\mathcal{M}_{\ell}$.

To evaluate the performance of different results, we define the criteria $c(\hat{\mathcal{M}})$ which represents the percentage of improvement of the corresponding algorithm

$$c(\hat{\mathcal{M}}) = 1 - \frac{d(\hat{\mathcal{M}}, \mathcal{M})}{d(\mathcal{D}, \mathcal{M})},$$

where $\mathcal{D}$ stands for the set corresponding to '$\diamond$' which is the outlier we want to pull towards the underlining $\mathcal{M}$ and $\hat{\mathcal{M}}$ stands for the set corresponding to '$\circ$' which is the result of different methods. The distance of $d(\mathcal{D}, \mathcal{M})$ is defined as:

$$d(\mathcal{D}, \mathcal{M}) = \frac{1}{n} \sum_{x_i \in \mathcal{D}} \|x_i - \tilde{x}_i\|_2 = \frac{1}{n} \sum_{x_i \in \mathcal{D}} \sigma_2 \|\epsilon_i\|_2$$

and by replacing $\mathcal{D}$ with $\hat{\mathcal{M}}$, we could similarly get $d(\hat{\mathcal{M}}, \mathcal{M})$.

The rightmost figure shows that with the increasing of the bandwidth h, the nonlinear projection result has a better performance (under the measurement of $c(\hat{\mathcal{M}})$) in data recovery aspect compared with the linear projection.

8.2. *Orthogonal Property.*

In this section, we give a simple case to show the effectiveness of nonlinear projection compared with the linear projection.

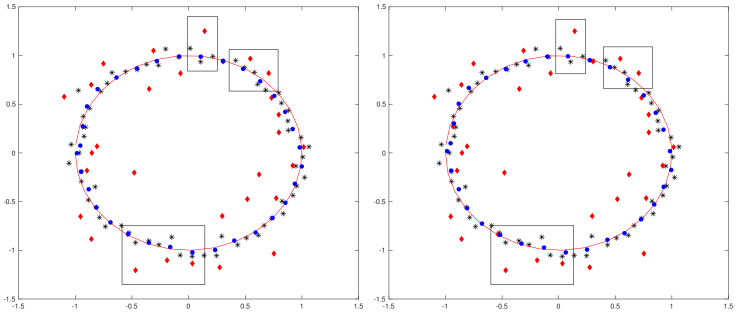

FIG 8. *Comparison between nonlinear projection and linear projection. Left: nonlinear approach. Right: linear approach*

Here, we give a brief explanation about the graph. The black stars $*$ represent the observations which is supposed to sampled from some manifold $\mathcal{M}$. The red diamonds $\diamond$ represent the outlier points which we want to project onto $\mathcal{M}$. Because $\mathcal{M}$ is unknown, we need to fit it locally with two different approaches. The red curve represents the true manifold. The red circles $\circ$ represent the projection of the diamonds $\diamond$ projected on the fitted locally defined manifold.

The procedure of projection can be seen in section 6, which can be described in a few steps.

A. *Locate the origin* : For each outlier point x, determine the origin x_0 corresponding to x, such as in (32).
B. *Fit the manifold* : Corresponding to x, fit with the observations to obtain the local defined function $\tilde{x}(\tau)$ and $x_\ell(\tau)$, i.e, get the respective parameters $\mathcal{A}$, U_d, $U_d^\perp$.
C. *Project backwards* : Project x onto the fitted manifold $\mathcal{M}_s$ defined by the function $\tilde{x}(\tau)$ by the repeatedly nonlinear least square. Project x onto the manifold $\mathcal{M}_\ell$ defined by the function $x_\ell(\tau)$ using the linear projection.

For fairness of the comparison, we use the same central point x_0 as (32). The difference between this two methods owns to the different function which locally approximate the manifold. The two functions can be summarised as the function of the linear affine space and the function of high-dimensional surface:

$$x_{\mathcal{A}}(\tau) = U_d^\perp \mathcal{A}(\tau,\tau) + U_d\tau + x_0, \quad x_\ell(\tau) = U_d\tau + x_0$$

The criteria for a better projection is to require the projection implement exactly in the normal space. We want the vector $x - P_{\mathcal{M}}(x)$ to be perpendicular with the tangent space of $\mathcal{M}$ at the projected point $P_{\mathcal{M}}(x)$, i.e.,

$$x - P_{\mathcal{M}}(x) \perp \mathcal{T}_{\mathcal{M}}(P_{\mathcal{M}}(x))$$

From the above two figures, we can see the left figures is with high quality, that is, almost all points projected onto $\mathcal{M}$ within the normal space, especially for the points emphasized in the rectangle.

8.3. *Parameters Reliance.* There are two parameters (the bandwidth h and neighborhood size k), which affect the performance of our algorithm. In this section, we show the relationship between the fitting property and the parameters. We use the color to represent the average approximation error $\mathrm{Error}(\mathcal{M}_{\mathcal{A}}, \mathcal{M}, \{x_i\})$, which is defined by

$$\mathrm{Error}(\mathcal{M}_{\mathcal{A}}, \mathcal{M}, \{x_i\}) = \frac{1}{n}\sum_i \|P_{\mathcal{M}_{\mathcal{A}}}(x_i) - P_{\mathcal{M}}(x_i)\|_2$$

where $P_{\mathcal{M}}(x_i)$ is the projection of x_i onto $\mathcal{M}$ and $P_{\mathcal{M}_{\mathcal{A}}}(x_i)$ is the projection of x_i onto the fitted manifold $\mathcal{M}_{\mathcal{A}}$.

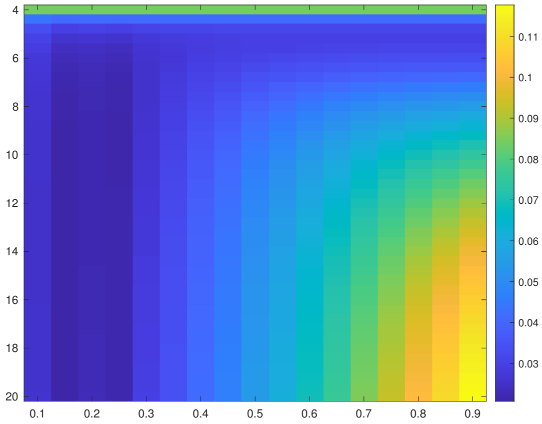

FIG 9. *Parameters reliance of* $\mathrm{Error}(\mathcal{M}_{\mathcal{A}})$ *on the bandwidth h (horizontal axis) and the neighborhood size k(vertical axis).*

In Figure (9), the dark blue color represents a good fitting standard with the corresponding parameters. We can see that setting smaller bandwidth h or restricting within a smaller neighborhood k has a similar effect on $\mathrm{Error}(\mathcal{M}_{\mathcal{A}})$. However, h also cannot be too small. When $h \to 0$, because of initial point and the estimated tangent turning bad with the increasing of the variance in the observations, the fitting criteria will become large indicating the fitted function gradually turns bad.

APPENDIX A

A.1. Asymptotic Property.

PROOF. Recall from (17) that the estimated $\hat{\theta}_k$ has a closed-form:

$$\hat{\theta}_k = (G^T W_h G)^{-1} G^T W_h \ell_k. \tag{37}$$

Note that, ℓ_k is the vector and the i-th elements represents the function value of $\phi_k(\tau_i)/2$. Because the Taylor expansion of $\phi_k(\tau)$ at $\tau = 0$ satisfied

$$\phi_k(0) = 0, \quad \nabla_\tau \phi_k(\tau)|_{\tau=0} = 0.$$

Thus, we have the Taylor expansion starting from the term of order 2 as

$$\frac{1}{2}\phi_k(\tau_i) = g_i^T \theta_k + t_i^T \eta_k + O(\|\tau_i\|_2^4), \tag{38}$$

where $\theta_k = \text{vech}(\nabla\nabla\phi_k(\tau)|_{\tau=0}), \eta_k = \text{vech}(\nabla\nabla\nabla\phi_k(\tau)|_{\tau=0})$, which is a vectorization of the third order derivative of $\phi_k(\tau)$ and t_i^T is the vectorization of the tensor $\tau_i \otimes \tau_i \otimes \tau_i$. Rewritting (38) in vector form, we have

$$\ell_k = G\theta_k + T\eta_k + O(\|\tau_i\|_2^4), \tag{39}$$

where T is a matrix and each row of T is t_i^T. Substitute (39) into (37), we have

$$\begin{aligned}\hat{\theta}_k - \theta_k &= (G^T W_h G)^{-1} G^T W_h \ell_k - \theta_k \\ &= (G^T W_h G)^{-1} G^T W_h (T\eta_k + O(\|\tau_i\|_2^4))\end{aligned}$$

Next, we give an estimation with the order of $G^T W_h G$ and $G^T W_h T$. Firstly, the expectation

$$\begin{aligned}&\text{E}(\frac{1}{n} G^T W_h G) \\ =&\frac{1}{h^D} \int_{y \in \mathcal{M}} \text{vech}(U^T (y-x)(y-x)^T U)\text{vech}(U^T (y-x)(y-x)^T U) K_h(y-x) p(y) dy.\end{aligned}$$

Letting $z = (y-x)/h$, we have $1/h^D dy = dz$. Changing the integrating variable from y to z, we have the expectation $\text{E}(G^T W_h G/n)$ and $\text{E}(G^T W_h T/n)$ becoming

$$\begin{aligned}&\text{E}(G^T W_h G/n) \\ =&h^4 p(x) \int \text{vech}(U^T zz^T U)\text{vech}^T(U^T zz^T U) K(z) p(x+hz) dz \\ =&h^4 p(x) \int \text{vech}(U^T zz^T U)\text{vech}^T(U^T zz^T U) K(z) dz + O(h^5)\end{aligned} \tag{40}$$

Similarly as the expectation derivation procedure in (40) for $\text{E}(G^T W_h G/n)$, we have the expectation

$$\begin{aligned}&\text{E}(G^T W_h T/n) \\ =&h^5 \int \text{vech}(U^T zz^T U)\text{vech}^T(U^T z \otimes U^T z \otimes U^T z) K(z) p(x+hz) dz \\ =&h^5 p(x) \int \text{vech}(U^T zz^T U)\text{vech}^T(U^T z \otimes U^T z \otimes U^T z) K(z) dz + O(h^6)\end{aligned} \tag{41}$$

For the variance, because of the independence of the samples, we have

$$\begin{aligned}&\text{Var}((G^T W_h G)/n) \\ =&\frac{1}{n^2}\text{Var}(\sum_i \text{vech}(\tau_i\tau_i^T)\text{vech}^T(\tau_i\tau_i^T) K_h(y_i - x)) \\ =&\frac{1}{n}\text{Var}(\text{vech}(\tau_i\tau_i^T)\text{vech}^T(\tau_i\tau_i^T) K_h(y_i - x)).\end{aligned} \tag{42}$$

Thus, from (42), for the (s,t)-th element, we have

$$\text{Var}((G^T W_h G)_{st}/n) = \frac{1}{n}\text{Var}(\text{vech}(\tau_i\tau_i^T)_s \text{vech}(\tau_i\tau_i^T)_t K_h(y_i - x)).$$

Using the equality of variance and expectation, to estimate $\text{Var}((G^T W_h G)_{st}/n)$, we just need to estimate the expectations as:

$$\begin{aligned}&\text{E}((G^T W_h G)_{st}^2/n^2) \\ =&\frac{1}{h^{D-8}} p(x) \int (\text{vech}(U^T zz^T U)_s)^2 (\text{vech}^T(U^T zz^T U)_t)^2 K^2(z) p(x+hz) dz\end{aligned} \tag{43}$$

Because of (42) and (43), we have the variance turning into

$$\mathrm{Var}(\frac{1}{n}G^TW_hG) = \frac{1}{nh^{D-8}}, \quad \mathrm{Var}(\frac{1}{n}G^TW_hT) = \frac{1}{nh^{D-10}}$$

As a result, we have

$$\frac{1}{n}G^TW_hG = O(h^4) + O_p(\frac{1}{\sqrt{nh^{D-8}}}), \quad \frac{1}{n}G^TW_hT = O(h^5) + O_p(\frac{1}{\sqrt{nh^{D-10}}})$$

The inverse of $\frac{1}{n}G^TW_hG$ is of order $O(h^{-4}) + O_p(\frac{1}{\sqrt{nh^{D+8}}})$. Taking all those together, we have:

$$\begin{aligned}\hat{\theta}_k - \theta_k =& O(h^{-4} + O_p(\frac{1}{\sqrt{nh^{D+8}}})) * O(h^5 + O_p(\frac{1}{\sqrt{nh^{D-10}}})) \\ =& O(h) + O_p(\frac{1}{\sqrt{nh^{D-2}}})\end{aligned}$$

□

A.2. Distance from x_0 to manifold.

PROOF. Since x_0 is a weighted summation of points $\{x_i, i = 1, 2...\}$ around $\bar{x}$ as

$$x_0 = U\sum_i w_i\tau_i + U_\perp \sum_i w_i\mathcal{A}(\tau_i,\tau_i) + x_0^* + \sigma\sum_i w_i\epsilon_i,$$

where the weight defined as $w_i = K_h(x_i - \bar{x})$. The distance from $\bar{x}$ to $\mathcal{M}_\mathcal{A}$ is upper bound by

$$\min_{y\in\mathcal{M}_\mathcal{A}} \|x_0 - y\|_2^2 \le \|x_0 - x(\tau_w))\|\|_2^2,$$

where $x(\tau_w) \in \mathcal{M}_\mathcal{A}$ and $\tau_w = \sum_w w_i\tau_i$. Next, we bound the distance of $\|x_0 - x(\tau_w))\|\|_2^2$. Substitute τ_w into the function of $\mathcal{M}_\mathcal{A}$, we have

$$\begin{aligned} x_0 - x(\tau_w) &= x_0 - (U\sum_i w_i\tau_i + U_\perp\mathcal{A}(\sum_i w_i\tau_i, \sum_i w_i\tau_i) + x_0^*) \\ &= U_\perp(\sum_i w_i\mathcal{A}(\tau_i,\tau_i) - \mathcal{A}(\sum_i w_i\tau_i, \sum_i w_i\tau_i)) + \sigma\sum_i w_i\epsilon_i \end{aligned} \tag{44}$$

Approximating each of $\mathcal{A}(\tau_i,\tau_i)$ by Taylor's expansion of $\mathcal{A}(\tau,\tau)$ at $\tau = \sum_i w_i\tau_i$, we have

$$\begin{aligned} \mathcal{A}(\tau_i,\tau_i) = \mathcal{A}(\sum_i w_i\tau_i, \sum_i w_i\tau_i) + 2\langle\tau_i - \sum_i w_i\tau_i, \mathcal{A}(\sum_i w_i\tau_i) \\ +\langle(\tau_i - \sum_i w_i\tau_i)(\tau_i - \sum_i w_i\tau_i)^T, \mathcal{A}\rangle\rangle \end{aligned} \tag{45}$$

Substitute (45) into (44), we have

$$x_0 - x(\tau_w) = U_\perp\langle\sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T, \mathcal{A}\rangle + \sigma\sum_i w_i\epsilon_i$$

Because of the independence of ϵ_i and τ_i, the squared 2 norm

$$\|x_0 - x(\tau_w)\|_2^2 = \|\sum_k w_k\mathcal{A}(\tau_k - \sum_i w_i\tau_i, \tau_k - \sum_i w_i\tau_i)\|_2^2 + \sigma^2\sum_i w_i^2\|\epsilon_i\|_2^2$$

Because $\sum_k w_k \mathcal{A}(\tau_k - \sum_i w_i\tau_i, \tau_k - \sum_i w_i\tau_i)$ is linear with respect to $\mathcal{A}$, we have the vector valued operation yields

$$\sum_k w_k \mathcal{A}(\tau_k - \sum_i w_i\tau_i, \tau_k - \sum_i w_i\tau_i) = \sum_{s,t} \mathcal{A}_{s,t} X_{s,t}$$

where the matrix $X = \sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T$. Notice that X is a positive semi-define matrix, yielding:

$$\begin{aligned}
&\sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T \\
=&\sum_k w_k(\tau_k - \tau + \tau - \sum_i w_i\tau_i)(\tau_k - \tau + \tau - \sum_i w_i\tau_i)^T \\
=&\sum_k w_k(\tau_k - \tau)(\tau_k - \tau)^T - (\tau - \sum_i w_i\tau_i)(\tau - \sum_i w_i\tau_i)^T.
\end{aligned} \tag{46}$$

Note that, in (46), the equation is satisfied for any τ. Particularly, letting τ to be the coordinate of $P_{\mathcal{M}_{\mathcal{A}}}(\bar{x})$, similar with (40), when the kernel is in a exponential type, we have:

$$K_h(\bar{x}, x_k) = K_h(P_{\mathcal{M}_{\mathcal{A}}}(\bar{x}), x_k) * K_h(P_{\mathcal{M}_{\mathcal{A}}}(\bar{x}), \bar{x})$$

Because of $K_h(P_{\mathcal{M}_{\mathcal{A}}}(\bar{x}), x_k) = K_h(\tau - \tau_k) * K_h(U_\perp^T(\bar{x} - x_k)))$, where $U_\perp$ is the basis corresponding to the normal space of $\mathcal{M}_{\mathcal{A}}$ at $P_{\mathcal{M}_{\mathcal{A}}}(\bar{x})$. Because $\|U_\perp^T(\bar{x} - x_k))\| = O(\|U^T(\bar{x} - x_k))\|_2^2)$, we have

$$K_h(U_\perp^T(\bar{x} - x_k)))) = O(\exp(-\|U^T(\bar{x} - x_k)\|_2^4/h))$$

Because $K_h(U_\perp^T(\bar{x} - x_k)))) \leq 1$, we have

$$\begin{aligned}
&\frac{1}{n}\mathrm{E}(\sum_k K_h(P_{\mathcal{M}_{\mathcal{A}}}(\bar{x}), x_k)(\tau_k - \tau)(\tau_k - \tau)^T) \\
\leq&\frac{1}{n}\mathrm{E}(\sum_k K_h(\tau_k - \tau)(\tau_k - \tau)(\tau_k - \tau)^T) \\
=&\int K_h(\eta - \tau)(\eta - \tau)(\eta - \tau)^T p(\eta) d\eta = O(h^2)
\end{aligned}$$

For the special case, because of $\sum_k w_k = 1, w_k \geq 0$, when $h \to 0$, there will be only one of $\{w_k, k = 1, 2...\}$ increases to 1 and the remaining parts equal to 0. Therefore, $\sum_i w_i\tau_i \to \tau_r$ as long as x_r is the nearest neighbor of $\bar{x}$. For $k \neq r$, $w_k = 0$. Overall, we have

$$\sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T = 0$$

Because of the modification by the semi-positive matrix $(\tau - \sum_i w_i\tau_i)(\tau - \sum_i w_i\tau_i)^T$, we have the corresponding eigenvalues yields,

$$\lambda_s(\sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T) \leq \lambda_s \sum_k w_k(\tau_k - \tau)(\tau_k - \tau)^T$$

Thus, using $\lambda_{\max}(\sum_k w_k(\tau_k - \tau)(\tau_k - \tau)^T) = O(h^2)$, we know that $\lambda_s(\sum_k w_k(\tau_k - \sum_i w_i\tau_i)(\tau_k - \sum_i w_i\tau_i)^T)$ is of order $O(h^2)$

$$\|\sum_k w_k \mathcal{A}(\tau_k - \sum_i w_i\tau_i, \tau_k - \sum_i w_i\tau_i)\|_2^2 \leq \|\mathcal{A}\|_2^2 O(h^4)$$

□

A.3. Contraction Proof.

PROOF. Denote $\omega(\tau)$, $\xi(\tau, c, s)$ and $\psi(\tau)$ as

$$\omega(\tau) = (2\mathcal{A}(\tau)\mathcal{A}(\tau)^T + I_d)^{-1}, \quad \xi(\tau, c, s) = s + 2\mathcal{A}(\tau, c),$$

$$\psi(\tau_1, \tau_2) = \mathcal{A}(\tau_1)\mathcal{A}^T(\tau_2)$$

Substitute the definition into $\|\tau_{n+1} - \tau_n\|_2$,

$$\|\tau_{n+1} - \tau_n\|_2 = \|\omega(\tau_n)\xi(\tau_n, c, s) - \omega(\tau_{n-1})\xi(\tau_{n-1}, c, s)\|_2 \tag{47}$$

Using the triangle inequality, we have (47) upper bounded by:

$$\begin{aligned}&\|\omega(\tau_n)\xi(\tau_n, c, s) - \omega(\tau_{n-1})\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\omega(\tau_n)\xi(\tau_n, c, s) - \omega(\tau_n)\xi(\tau_{n-1}, c, s)\|_2 + ...\\ +&\|\omega(\tau_n)\xi(\tau_{n-1}, c, s) - \omega(\tau_{n-1})\xi(\tau_{n-1}, c, s)\|_2\end{aligned}$$

Since $\|\omega(\tau_n)\|_2 \leq 1$, using the norm inequality, we know the first term bounded by:

$$\begin{aligned}&\|\omega(\tau_n)\xi(\tau_n, c, s) - \omega(\tau_n)\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\omega(\tau_n)\|_2\|\xi(\tau_n, c, s) - \xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\xi(\tau_n, c, s) - \xi(\tau_{n-1}, c, s)\|_2\\ =&\|\mathcal{A}(c)\|_2\|\tau_n - \tau_{n-1}\|_2\\ \leq&\|\mathcal{A}\|_2\|c\|_2\|\tau_n - \tau_{n-1}\|_2\end{aligned}$$

For the second term, using the norm inequality, we have:

$$\begin{aligned}&\|\omega(\tau_n)\xi(\tau_{n-1}, c, s) - \omega(\tau_{n-1})\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\omega(\tau_n) - \omega(\tau_{n-1})\|_2\|\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\omega(\tau_n)\|_2\|\omega(\tau_{n-1})\|_2\|\psi(\tau_n) - \psi(\tau_{n-1})\|_2\|\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|\psi(\tau_n, \tau_n) - \psi(\tau_{n-1}, \tau_{n-1})\|_2\|\xi(\tau_{n-1}, c, s)\|_2\end{aligned}$$

For $\|\xi(\tau_{n-1}, c, s)\|_2$, it is bounded by

$$\begin{aligned}&\xi(\tau_{n-1}, c, s)\|_2\\ \leq&\|s\|_2 + 2\|\mathcal{A}(\tau_{n-1}, c)\| \leq \|s\|_2 + 2\|\mathcal{A}\|_2\|\tau_{n-1}\|_2\|c\|_2\end{aligned}$$

$$\begin{aligned}&\|\psi(\tau_n, \tau_n) - \psi(\tau_{n-1}, \tau_{n-1})\|_2\\ \leq&\|\psi(\tau_n, \tau_n) - \psi(\tau_n, \tau_{n-1})\|_2 + \|\psi(\tau_n, \tau_{n-1}) - \psi(\tau_{n-1}, \tau_{n-1})\|_2\\ \leq&(\|\mathcal{A}(\tau_n)\|_2 + \|\mathcal{A}(\tau_{n-1})\|_2)\|\mathcal{A}(\tau_n - \tau_{n-1})\|_2\\ \leq&(\|\mathcal{A}(\tau_n)\|_2 + \|\mathcal{A}(\tau_{n-1})\|_2)\|\mathcal{A}\|_2\|\tau_n - \tau_{n-1}\|_2\\ \leq&\|\mathcal{A}\|_2^2(\|\tau_n\|_2 + \|\tau_{n-1}\|_2)\|\tau_n - \tau_{n-1}\|_2\end{aligned}$$

Using the assumption $\|\tau_n\|_2 \leq \beta$, we will have

$$\begin{aligned}&\|\tau_{n+1} - \tau_n\|_2\\ \leq&\|\omega(\tau_n)\xi(\tau_n, c, s) - \omega(\tau_{n-1})\xi(\tau_{n-1}, c, s)\|_2\\ \leq&(2\beta\|\mathcal{A}\|_2^2)(\|s\|_2 + 2\|\mathcal{A}\|_2\|c\|_2\beta)\|\tau_n - \tau_{n-1}\|_2\\ \leq&(4\beta^2\|\mathcal{A}\|_2^3\|c\|_2 + 2\|s\|_2\|\mathcal{A}\|_2^2)\|\tau_n - \tau_{n-1}\|_2\end{aligned}$$

Using the assumption, $(4\beta^2\|\mathcal{A}\|_2^3\|c\|_2 + 2\|s\|_2\|\mathcal{A}\|_2^2) \leq 1$, we will get our result! □

A.4. Optimizing with S^k . Instead of vectorizing S^k, we can also obtain the form of S^k by using the steepest decent algorithm. In the gradient decent algorithm, we do not need the solution property into consideration. By repeatedly iterating with a decent direction, we will get the stationary point.

Optimizing with the S^k directly can be very clear and simple. To find $\mathcal{A}_{\phi_{x_0}}(0)$, we just to need to find the matrix $S^{d+1},...,S^D$, independently. For dimension $k \in \{d+1,...,D\}$, we show the k-th slide of $\mathcal{A}_{\phi_{x_0}}(0)$ can be solved by

$$\hat{\mathcal{A}}^k_{\phi_{x_0}}(0)(\tau_i,\tau_i) = \langle S^k, \tau_i^T\tau_i\rangle + o(\|\tau_i\|_2^2)$$

Meanwhile,

$$\hat{\mathcal{A}}^k_{\phi_{x_0}}(0)(\tau_i,\tau_i) = U_k^T(x_i - x_0)$$

Thus, because our locally fitting intention, we can define a locally weighted loss function by bringing in a kernel $K_h(x_i - x)$ as

$$F(S^k) = \sum_{i=1}^N K_h(x_i - x_0)(\langle S^k, \tau_i\tau_i^T\rangle - U_k^T(x_i - x_0))^2. \tag{48}$$

Take the derivative of $F(S^k)$ in (48) with respect to S^k, we get:

$$\nabla F(S^k) = 2\sum_{i=1}^N K_h(x_i - x_0)(\langle S^k_{n-1}, \tau_i^T\tau_i\rangle - U_k^T(x_i - x_0))\tau_i\tau_i^T.$$

Minimize the loss function we will get the optimum S^k:

$$\hat{S}^k = \min_{S^k} F(S^k). \tag{49}$$

Instead of deriving the closed-form of (49), we update with a gradient decent iteration as

$$S^k_n = S^k_{n-1} - \sum_{i=1}^N K_h(x_i - x_0)(\langle S^k_{n-1}, \tau_i^T\tau_i\rangle - U_k^T(x_i - x_0))\tau_i\tau_i^T.$$

Note that, the convergence point S^k_* will satisfy the first order optimal condition (KKT) because letting $n \to \infty$, we will have $\nabla F(S^k_*) = 0$.

REFERENCES

BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15** 1373–1396.

DAVENPORT, M. A., HEGDE, C., DUARTE, M. F. and BARANIUK, R. G. (2010). Joint manifolds for data fusion. *IEEE Transactions on Image Processing* **19** 2580–2594.

FEFFERMAN, C., IVANOV, S., KURYLEV, Y., LASSAS, M. and NARAYANAN, H. (2018). Fitting a putative manifold to noisy data. In *Conference On Learning Theory* 688–720.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I., WASSERMAN, L. et al. (2014). Nonparametric ridge estimation. *The Annals of Statistics* **42** 1511–1545.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems* (Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE and K. Q. WEINBERGER, eds.) **27** 2672–2680. Curran Associates, Inc.

OZERTEM, U. and ERDOGMUS, D. (2011). Locally defined principal curves and surfaces. *Journal of Machine learning research* **12** 1249–1286.

PANARETOS, V. M., PHAM, T. and YAO, Z. (2014). Principal flows. *Journal of the American Statistical Association* **109** 424–436.

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science* **290** 2323–2326.

SOBER, B. and LEVIN, D. (2019). Manifold approximation by moving least-squares projection (mmls). *Constructive Approximation* 1–46.

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science* **290** 2319–2323.

YAO, Z. and XIA, Y. (2019). Manifold Fitting under Unbounded Noise. *arXiv preprint arXiv:1909.10228*.

ZHANG, Z. and ZHA, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing* **26** 313–338.