Abstract

We consider the fixed boundary flow with canonical interpretability as principal components extended on the non-linear Riemannian manifolds. We aim to find a flow with fixed starting and ending point for multivariate datasets lying on an embedded non-linear Riemannian manifold, differing from the principal flow that starts from the center of the data cloud. Both points are given in advance, using the intrinsic metric on the manifolds. From the perspective of geometry, the fixed boundary flow is defined as an optimal curve that moves in the data cloud. At any point on the flow, it maximizes the inner product of the vector field, which is calculated locally, and the tangent vector of the flow. We call the new flow the fixed boundary flow. The rigorous definition is given by means of an Euler-Lagrange problem, and its solution is reduced to that of a Differential Algebraic Equation (DAE). A high level algorithm is created to numerically compute the fixed boundary. We show that the fixed boundary flow yields a concatenate of three segments, one of which coincides with the usual principal flow when the manifold
is reduced to the Euclidean space. We illustrate how the fixed boundary flow can be used and interpreted, and its application in real data.

Keywords: vector field, manifolds, curve, boundary condition, tangent space

1 Introduction

Most of the existing statistical methods assume a linear dependency between features. As the dimension of the features increases, the representation of the features in a high dimensional space becomes more complex and it becomes more challenging to learn the relation among the features. In many applications, modern data structure is complex and not necessarily linear. It is often the case that there is a lower-dimensional structure, namely a manifold embedded in the high dimensional ambient space, as in the examples of geometric shapes in the shape space [26, 5, 16, 3] and graphs in computer graphics [23, 10, 1].

A series of methods that aim to recover the underlying structure of the lower dimensional manifold have been developed over the past two decades. These methods, usually called manifold learning, are mostly focused on mapping the data in a \(d\)-dimensional space into a set of points close to a \(m\)-dimensional \((m \ll d)\). Among them, the Principal Component Analysis (PCA) is a method commonly used to reduce the dimension of the features in the Euclidean space. To address features lying in a non-linear space (i.e., a manifold), methods such as LLE [24], Isomap [25], MDS [4], and LTSA [29], which find the low-dimensional embedding preserving local properties of the data, may be preferred. See a comprehensive review in [20].

Another line of research relating to statistics on manifolds is centered on the extension of the existing methods defined in the Euclidean space to the manifold space. The manifold space can be the actual physical space that the data lies on or the learnt manifold created through the manifold learning methods. Over the past decades, numerous non-linear approaches have been developed to analyze the data on the manifold directly [14, 9, 12, 17, 8, 15, 13, 7]. Throughout the paper, we focus on the known manifold, based on the assumption that the manifold embedding is known.

Next, we will mainly review the “curve fitting” methods on manifolds. A geodesic is a generalization of the straight lines in the standard Euclidean space to the manifold. The principal geodesic analysis [9], which extends the PCA to the manifold, was proposed to describe the nonlinear variability of data on a manifold. The principal curves, proposed in [11], are flexible one-dimensional curves that pass through the middle of the data points. Having said that, principal curves are able to better capture the non-linear variation of the data in comparison to any of the other regression lines in the Euclidean space. [21] redefined principal curves and surfaces in terms of the gradient and the Hessian of the probability density estimate, based on the consideration that every point on the principal surface should be at the local maximum of the probability density in the local orthogonal subspace, and not the expected value as in [11]. For applications in classification tasks,
[18] proposed a new curve fitting method to find the smooth decision boundary with bounded curvature.

A recent piece of work on principal flows [22] works as an extension of the principal curves on Riemannian manifolds. Therefore, the principal flows are also flexible one-dimensional curves, which pass through the mean of the data points. The principal flows are able to capture the non-geodesic pattern of variation both locally and globally. Instead of handing curves with an explicit parameterization, [19] combine the level set method with the principal flow algorithm to obtain a fully implicit formulation, so that the obtained codimensional-one surface on the manifold fits the dataset well.

![Figure 1](image1.png)  
(a)  
(b)

Figure 1: Earthquake data of the Japan Sea.

We consider flows with boundaries (starting and ending points) instead of passing the middle of the data. When the data comes with multiple paths, we want to isolate one of the paths in particular—the path with a fixed direction. For example, we consider the seismological events of the Japan Sea (year in detail), as shown in Figure 1(a). The principal flow fits the earthquake data marked by the red curve in Figure 1(b), since there are more data points along this path. When we focus on the seismological events caused by the Philippine Sea Plate or the Pacific Plate, the principal flow will not be of interest. In this sense, the blue curve in Figure 1(b) would be more appropriate. Although we could get a curve similar to the blue one by manually excluding all the data for locations that are far away from the Philippine Sea Plate, it is hard to accurately determine which data points to exclude. Hence, it is necessary to consider flows with fixed boundaries.

In this paper, we are motivated to consider the fixed boundary flow that captures the variation of the data in the same way that principal flows work. To achieve this, we redefine vector fields following the direction of maximal variability similar to (but not the same as) the principal flows in Section 2.1. In Section 2.2, we propose an optimization
to capture a smooth curve on the manifold that starts and ends at given points, with the property that, at each point, its derivative is close to the vector field at the same point. We also demonstrate how the optimization can be transformed into an Euler-Lagrange problem and describe an algorithm based on a Differential Algebraic Equation (DAE) in Section 3. The solution of our optimization is further proven to be canonical, in the sense that it will pass through the usual principal component when the manifold is reduced to the Euclidean space (Section 5). In Section 4 and 6, we illustrate that the fixed boundary flow is able to capture patterns of variation in both synthetic and real-world seismological data.

2 Fixed boundary flows

2.1 Review on principal flows

Throughout this paper, we will work within the context of a complete Riemannian manifold $\mathcal{M}$ of dimension $m$, and $\mathcal{M}$ is isometrically embedded into the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$ with $m < d$. The preliminaries used in this section can be found in the Supplementary Materials. Given $\{x_1, \cdots, x_n\} \subset \mathcal{M}$, the methodology for the principal flow seeks to determine a curve on the manifold that passes through the mean of the data, such that the tangent vector along the curve locally follows the first eigenvector of a local tangent space. Specifically, the principal flow of scale $h > 0$ is constructed by starting from the Fréchet mean $\bar{x}$. Assuming that the tangent covariance matrix $\Sigma_h(x)$ has distinct first and second eigenvalues for all $x \in \mathcal{N}(\bar{x})$, the vector field $W = \{W(x) : x \in \mathcal{N}(\bar{x})\}$ defined over the neighborhood of $\bar{x}$ is extended by the first eigenvector $e_1(\bar{x})$ (corresponding with the first eigenvalue $\lambda_1(\bar{x})$) of the tangent covariance matrix $\Sigma_h(x)$, that is,

$$\Sigma_h(x)W(x) = \lambda_1(x)W(x), \quad \text{for } x \in \mathcal{N}(\bar{x}, h). \quad (1)$$

According to [22], the vector field $W : \mathcal{N}(\bar{x}) \to \mathbb{R}^d$ and its extension [28] is a differentiable map and is independent of the local coordinates of $T_x\mathcal{M}$. The principal flow is a union of two parts joining at the Fréchet mean $\bar{x}$, where one part follows the direction of the vector field and the other part follows the opposite direction of the vector field. Each part is required to be at length $r$ and unit speed. Fixing the starting point, total length, tangent direction and speed, the ending point of such a curve is determined.

Mathematically, the principal flow finds a curve $\gamma : [0, r] \to \mathcal{M}$ starting at $\bar{x}$ and maximizing

$$\int_0^r \lambda_1(\gamma(t))\langle \dot{\gamma}(t), W(\gamma(t)) \rangle dt. \quad (2)$$

We remark that $\lambda_1(x)$ and $W(x)$ vary with $x$. Figure 2 shows the distribution of $\lambda_1(x)$ defined in (1) for $x \in \mathcal{N}(\bar{x}, h)$ with respect to its magnitude. Specifically: the red circle represents $\bar{x}$, the black points represent sample points, and the color represents the
magnitude of $\lambda_1(x)$. From Figure 2, one may see that $\lambda_1(x)$ gets larger when the color transitions to yellow, while $\lambda_1(x)$ gets smaller when the color transitions to blue.

![Figure 2: $\lambda_1(x)$ defined in (1) for $x \in N(\bar{x}, h)$. $\lambda_1$ of the red circle is 0.5063 and $\lambda_1$ of the red cross is 2.227.](image)

Figure 2 shows that $\lambda_1(x)$ reaches its trough at $\bar{x}$, which in turn implies that $\lambda_1(x)$ would increase along any direction when it departs from $\bar{x}$. In addition, by

$$d\lambda_1(x) = W(x)^T d\Sigma_h(x) W(x) = \langle W(x)W(x)^T(x - \bar{x}), dx \rangle,$$

the vector field $W(x)$ is along the gradient direction of $\lambda_1(x)$, which means $\lambda_1(x)$ changes at the fastest rate along $W(x)$. Therefore, when $\gamma(t)$ moves along $W(\gamma(t))$, both $\lambda_1(\gamma(t))$ and the inner product $\langle \dot{\gamma}(t), W(\gamma(t)) \rangle$ increases at the fastest rate simultaneously. So, at any $t' \in [0, 1]$, $\dot{\gamma}(t') = W(\gamma(t'))$ maximizes (2) locally.

Things appear to be different for flows starting at other points instead of $\bar{x}$. When we consider flows starting around the boundary of the data cloud, such as the red cross $\bar{x}_1$ in Figure 2, it is seen that $\lambda_1(x)$ will decrease when a curve starts at $\bar{x}_1$ and moves towards the data cloud. By the fact that $\lambda_1(x)$ changes at the fastest rate along $W(x)$, $\gamma(t)$ should always move orthogonally to $W(\gamma(t))$, in order to make $\lambda_1(\gamma(t))$ decrease at the slowest rate. Therefore, in the case of the fixed boundary, maximizing $\lambda_1(\gamma(t))$ and the inner product $\langle \dot{\gamma}(t), W(\gamma(t)) \rangle$ is mutually conflicting, if the same optimization of the principal flow is adopted.

The above discussion shows that finding a flow starting at the boundary of the data cloud is not a simple extension to finding a flow starting at the average of the data cloud. Furthermore, the ending point of a principal flow is determined in the manner that we mentioned previously, and thereby cannot be controlled artificially. When the ending point of a curve is concerned, as in the case of the seismological event, we cannot use the principal flow.
2.2 Definition of fixed boundary flows

Contrasting with the principal flow which has a fixed starting point and an initial direction, the fixed boundary flows are curves with fixed starting points \( \bar{x}_1 \) and ending points \( \bar{x}_2 \). By fixing two boundary points, there are an infinite number of curves. For any such curve, we can find the moving direction and the vector field at every point \( x' \) on it.

The moving directions and vector fields vary with different points and different curves. To achieve the highest variation, we aim to find a curve with a moving direction that matches the direction of maximal variation (i.e., the vector field) as much as possible at any given point on the curve. Using the terminology from classical mechanics, we hope to find a flow, starting at \( \bar{x}_1 \) and ending at \( \bar{x}_2 \), that well approximates the vector field \( W \).

To satisfy the boundary conditions, the fixed boundary flows should be considered globally. Intrinsically local and greedy approaches fail to do this. This section provides a principle to construct such a path with maximal “cumulative variation” motivated by the principal flow. We call the optimal curve the fixed boundary flow. To begin, we describe the class of functions that provide the candidates of a fixed boundary flow.

For the fixed starting point \( \bar{x}_1 \) and ending point \( \bar{x}_2 \), define the class

\[
\Gamma(\bar{x}_1, \bar{x}_2) = \left\{ \gamma : [0, r] \rightarrow M : \gamma \in C^2([0, r]), r \in (0, 1], \gamma(0) = \bar{x}_1, \gamma(r) = \bar{x}_2, \gamma(s) \neq \gamma(s') \text{ for } s \neq s', \ell(\gamma[0, t]) = t, \text{ for all } t \in [0, r] \right\},
\]

where \( \ell(\gamma[0, t]) \) denotes the length of the parametric curve \( \gamma \) from \( \gamma(0) \) to \( \gamma(t) \), for all \( 0 \leq t \leq r \). Since \( t \in [0, 1] \), the length of the curves in the class \( \Gamma(\bar{x}_1, \bar{x}_2) \) is less than 1. The fixed boundary flow is defined as a smooth curve \( \gamma \) on the manifold \( M \), starting and ending at the fixed points, with a derivative vector \( \dot{\gamma} \) that is maximally compatible with the vector field \( W \) of the localized variation.

**Definition 2.1.** (Fixed boundary flow at scale \( h \)) Let \( \bar{x}_1, \bar{x}_2 \in B \), where \( B \) is the neighborhood that contains the data \( \{x_1, \ldots, x_n\} \) on the manifold. Assume that \( \Sigma_h(x) \) have distinct first and second eigenvalues for any \( x \in B \). A fixed boundary flow of \( \{x_1, \ldots, x_n\} \) with fixed boundary \( \bar{x}_1 \) and \( \bar{x}_2 \) is the curve satisfying

\[
\gamma = \arg \sup_{\gamma \in \Gamma(\bar{x}_1, \bar{x}_2)} \int_0^t \langle \dot{\gamma}, W(\gamma(t)) \rangle dt,
\]

where \( W(\gamma(t)) \) is the vector field over the neighborhood of \( \gamma(t) \) for \( 0 \leq t \leq 1 \).

3 Determination of fixed boundary flows via the Euler-Lagrange method

To obtain the fixed boundary flows, we reformulate the variation problem (4) into the Euler-Lagrange formulation with fixed boundary conditions. From (4), the variation
problem is equivalent to finding a critical point of
\[ \mathcal{L}(W, \gamma) = \int_0^{\ell(\gamma)} \langle \dot{\gamma}, W(\gamma(t)) \rangle dt \] (5)
subject to the constraints of the curves over the class \( \Gamma(\bar{x}_1, \bar{x}_2) \) defined in (3), which are elaborated as follows:

1. \( \|\dot{\gamma}(t)\| = 1 \), for \( 0 \leq t \leq r \) and \( \ell(\gamma[0, r]) = \int_0^r \|\dot{\gamma}(t)\| dt = r \).
2. \( \gamma(0) = \bar{x}_1 \) and \( \gamma(r) = \bar{x}_2 \).
3. \( \gamma(s) \neq \gamma(s') \) for \( s \neq s' \).
4. \( F(\gamma(t)) = 0 \) for any \( t \in [0, r] \).

We note that the resulting curve will be a unit speed curve under the first constraint. Based on this constraint, the integral (5) is the case of integration with a variable upper limit \( \ell(\gamma) \), and it makes the implementation awkward. Therefore, we define an equivalent class for \( \Gamma(\bar{x}_1, \bar{x}_2) \), that is,
\[ \Gamma_2(\bar{x}_1, \bar{x}_2) = \{ \gamma : [0, 1] \to M : \gamma \in C^2([0, r]), r \in (0, 1], \gamma(0) = \bar{x}_1, \gamma(1) = \bar{x}_2, \gamma(s) \neq \gamma(s') \text{ for } s \neq s', \int_0^1 \|\dot{\gamma}(t)\|^2 dt = 1 \}. \] (6)

We prove that the sets \( \Gamma(\bar{x}_1, \bar{x}_2) \) and \( \Gamma_2(\bar{x}_1, \bar{x}_2) \) are equivalent in the following proposition.

**Proposition 3.1.** (Reparameterization) The curve sets \( \Gamma_2(\bar{x}_1, \bar{x}_2) \) and \( \Gamma(\bar{x}_1, \bar{x}_2) \) are equal up to curve reparameterization.

In the following, we will focus on finding the curve subject to the curves \( \gamma \in \Gamma_2(\bar{x}_1, \bar{x}_2) \). Define the Lagrangian
\[ \mathcal{L}(t, \gamma, \dot{\gamma}) = \langle \dot{\gamma}, W(\gamma(t)) \rangle + \delta \langle \dot{\gamma}, \dot{\gamma} \rangle + zF(\gamma), \] (7)
where \( \delta \) and \( z \) are Lagrange multipliers. Thus, the original variation problem (4) can be represented as the process of finding critical points of the objective function
\[ \int_0^1 \mathcal{L}(t, \gamma, \dot{\gamma}) dt, \] (8)
with the fixed boundary conditions \( \gamma(0) = \bar{x}_1 \) and \( \gamma(1) = \bar{x}_2 \). Following the standard of the Euler-Lagrange formulation, we obtain the Euler-Lagrange equation
\[ -2\delta \ddot{\gamma} + DF(\gamma)^T z + \left( \dot{W}(\gamma)^T - \dot{W}(\gamma) \right) \dot{\gamma} = 0, \] (9)
subject to the curve $\gamma$ that is constrained on the manifold, where $DF$ is the derivative of $F$. Then, the formulation of the constrained Euler-Lagrange equation can be rewritten into the form

$$Q(\gamma)\ddot{\gamma} + DF(\gamma)^T z = G(\gamma, \dot{\gamma}), \quad (10a)$$

$$F(\gamma) = 0, \quad (10b)$$

where the mass matrix $Q(\gamma) = -2\delta I_d$, and $G(\gamma, \dot{\gamma}) = (\dot{W}(\gamma) - \dot{W}(\gamma)^T) \dot{\gamma}$.

Therefore, the determination of the fixed boundary flow is equivalent to solving the constrained Euler-Lagrange equations (10), which contains a differential equation (10a) and an algebraic constraint in (10). This type of differential equation is called a Differential Algebraic Equation (DAE). Unlike an ordinary differential equation (ODE), the index of a DAE measures its singularity. The index is essentially defined by the number of times that the algebraic constraint (10b) needs to be differentiated with respect to the time variable $t$, so that the resulting differential equations, together with the given differential equations (10a), form an ODE. Therefore, we conclude that the constrained Euler-Lagrange equation (10) is a DAE with an index of at most three.

**Remark 3.1.** DAEs of index two or higher are generally difficult to analyze and solve numerically. To overcome the problems inherent in the solution of the higher index form of DAEs, index reduction methods were proposed to formulate the equations based on differentiation of the algebraic constraint.

**Remark 3.2.** An index-two DAE can be formed by replacing the position level constraint (10b) with its first derivation, which is called the velocity level constraint and is given by,

$$DF(\gamma)\dot{\gamma} = 0. \quad (11)$$

However, we also observe the effects caused by the index reduction. The solution of the index-two system can drift away from the position level constraint (10b). Instead of lying exactly on the manifold $M$, the curve, which we solve from the index-two DAE, could be located on the manifolds parallel to the target manifold $M$. This phenomenon is called algebraic instability. According to [2], this instability will not exist if we replace the single velocity constraint (11) by combining the position level constraint with the velocity level constraint, that is,

$$a_0 F(\gamma) + a_1 DF(\gamma)\dot{\gamma} = 0 \quad (12)$$

with $\sigma(z) = a_0 + a_1 z$ having only negative roots. Therefore, solving the index-three DAE is equivalent to solving the index-two DAE, which is formed by combining (10a) and (12). Although we can reduce the index of the DAE (10) to two, it is still challenging to numerically solve the index-two DAE with boundary conditions. Only one existing R package can solve it. From our experience, the R package bvpSolve was slow and encountered errors when solving the boundary condition index-two DAE.
Remark 3.3. Alternatively, an index-one DAE can be formed using the acceleration level constraint, which is given by

\[ DF(\gamma)\ddot{\gamma} + \dot{\gamma}^T F_{\gamma\gamma}(\gamma)\dot{\gamma} = 0, \]  

where \( F_{\gamma\gamma} \) is the second derivative of \( F \) with respect to \( \gamma \). Throughout this paper, we will focus on solving the index-one DAE with boundary conditions.

3.1 Coordinate projection methods for solving DAE of index one

In this section, we establish the numerical implementation for solving the index-one DAE with boundary conditions. The index-one DAE, which is formed by the differential equation (10a) and the acceleration level constraint (13), can be written into the following matrix form

\[
\begin{pmatrix}
Q(\gamma) & DF(\gamma) \\
DF(\gamma) & 0
\end{pmatrix}
\begin{pmatrix}
\ddot{\gamma} \\
\dot{\gamma}
\end{pmatrix}
= 
\begin{pmatrix}
G(\gamma, \dot{\gamma}) \\
-\dot{\gamma}^T F_{\gamma\gamma}(\gamma)\dot{\gamma}
\end{pmatrix}.
\]  

From the index-one DAE (14), we are able to solve the Lagrange multiplier \( z \),

\[
z = (DF(\gamma)Q(\gamma)^{-1}DF(\gamma)^T)^{-1}(DF(\gamma)Q(\gamma)^{-1}G(\gamma, \dot{\gamma}) + \dot{\gamma}^T F_{\gamma\gamma}(\gamma)\dot{\gamma}).
\]

Substituting the result for \( z \) into the differential equation (10a) leads to the underlying ODE system,

\[
Q(\gamma)\ddot{\gamma} = \tilde{G}(\gamma, \dot{\gamma}),
\]

where

\[
\tilde{G}(\gamma, \dot{\gamma}) = G(\gamma, \dot{\gamma}) - DF(\gamma)^T (DF(\gamma)Q(\gamma)^{-1}DF(\gamma)^T)^{-1}
(DF(\gamma)Q(\gamma)^{-1}G(\gamma, \dot{\gamma}) + \dot{\gamma}^T F_{\gamma\gamma}(\gamma)\dot{\gamma}).
\]

Once the underlying ODE is obtained, we are able to achieve an approximate solution for the index-one DAE by applying the coordinate projection methods to the underlying ODE.

The numerical implementation starts from an initial discrete curve \( \gamma^0 \) starting at \( \bar{x}_1 \) and ending at \( \bar{x}_2 \), with a mesh \( 0 = t_0 < t_1 < ... < t_N = 1 \). The first step of the coordinate projection method is to solve the underlying ODE (16). Let \( \gamma = u \), and rewrite the second-order underlying ODE (16) into a first-order ODE as such:

\[
\begin{pmatrix}
\dot{\gamma} \\
\dot{u}
\end{pmatrix} = 
\begin{pmatrix}
Q(\gamma)^{-1}\tilde{G}(\gamma, u)
\end{pmatrix}.
\]

Then, apply the collocation methods to the first-order ODE (17).

Let \( 0 < c_1 < ... < c_k \leq 1 \). For each sub-interval \([t_{i-1}, t_i] \), there are \( k \) collocation points \( t_j = t_{i-1} + h_i c_j \), where \( j = 1, ..., k \) and \( h_i = t_i - t_{i-1} \). To implement the
collocation methods to the first-order ODE (17), we need to determine the vector field \( W \) in the function \( \tilde{G} \). With the aim of getting the appropriate vector field, a projection modification is adapted into the calculation and the details are elaborated in Figure 3. We first project the point \( \gamma^0(t_i) \) from the initial curve \( \gamma^0 \) to the data cloud and obtain the projected point \( \gamma^0_{\text{proj}}(t_i) \). Then, the vector field \( W(\gamma^0(t_i)) \) is obtained from the local covariance matrix of the projected point \( \gamma^0_{\text{proj}}(t_i) \) in the data cloud, and the vector \( e_1(\gamma^0_{\text{proj}}(t_i)) \) is moved from the projected point \( \gamma^0_{\text{proj}}(t_i) \) to the point \( \gamma^0(t_i) \) along the geodesic. Once the vector field for the ODE (17) is determined, an ODE solution \( \gamma_{\text{ODE}}(t_i) \) is obtained at each mesh point \( t_i \) through several collocation points \( \gamma(t_j), j = 1, \ldots, k \). However, as the ODE solution \( \gamma_{\text{ODE}}(t_i) \) may not necessarily lie on the manifold \( \mathcal{M} \), the ODE solution \( \gamma_{\text{ODE}}(t_i) \) is projected orthogonally back onto the manifold \( \mathcal{M} \). An iterative procedure is carried out to obtain the converged curve \( \tilde{\gamma}(t_i), i = 0, \ldots, N \).

A high-level description of the algorithm optimizing the Lagrangian \( L \) defined in (5) is summarized in Algorithm 1. In [6], the author proved that the coordinate projection modification for the index-one DAE with boundary conditions helps restore all the properties of stability and superconvergence corresponding collocation methods. Then, for \( h := \max_i h_i > 0 \) which is sufficiently small, the global error for the projected collocation solution \( \tilde{\gamma}(t) \) satisfies

\[
|\gamma(t) - \tilde{\gamma}(t)| = O(h^k),
\]

where \( 0 \leq t \leq 1 \).

Figure 3: Calculation of vector field. (a) Projecting \( \gamma^k(t_i) \) into the data cloud at \( \gamma^k_{\text{proj}}(t_i) \). (b) Calculating the local covariance matrix \( \Sigma_h(\gamma^k_{\text{proj}}(t_i)) \) from the data points in the neighborhood \( h \) of the projected point \( \gamma^k_{\text{proj}}(t_i) \); computing the first eigenvector \( e_1(\gamma^k_{\text{proj}}(t_i)) \) from the local covariance matrix. (c) Moving the vector \( e_1(\gamma^k_{\text{proj}}(t_i)) \) from the projected point \( \gamma^k_{\text{proj}}(t_i) \) to \( \gamma^k_{\text{proj}}(t_i) \) along the geodesic. Black cross: data points. Blue dotted line: initial curve \( \gamma^0 \) with starting point \( \tilde{x}_1 \) and ending point \( \tilde{x}_2 \). Green dotted line: an intermediate curve. Red dotted line: converged curve \( \tilde{\gamma} \).
Algorithm 1: Numerical optimization of the Lagrangian $\mathcal{L}$ in (5).

1. Compute the direction vector $v_0 = \bar{x}_2 - \bar{x}_1$;
2. Choose a value $\delta$ for the Lagrange multiplier;
3. Set an initial discrete curve $\gamma^0(t_i), i = 0, \ldots, N$, with $\gamma^0(t_0) = \bar{x}_1$ and $\gamma^0(t_N) = \bar{x}_2$;
4. Take the initial curve $\gamma^0$ as the input curve $\gamma$.
5. Calculate the vector field $W(\gamma(t_i)), i = 0, \ldots, N$ for the points from the input curve $\gamma$. If $\langle W(\gamma(t_i)), v_0 \rangle > 0$, take $W(\gamma(t_i))$ as the vector field; otherwise, take $-W(\gamma(t_i))$ as the vector field.
6. Choose $k$ collocation points in $[t_0, t_N]$ and apply collocation methods to obtain an approximate solution $\gamma_{ODE}(t_i), i = 0, \ldots, N$, for the ODE (17) with the vector field calculated in step (5).
7. Project the ODE solution $\gamma_{ODE}(t_i)$ back onto the manifold and denote the curve lying on the manifold as $\tilde{\gamma}(t_i), i = 0, \ldots, N$.
8. Take the curve $\tilde{\gamma}(t_i)$ as the input curve and repeat step (5) until the curve $\tilde{\gamma}(t_i)$ converges.

Remark 3.4.

1. Step (7) can also be implemented when the ODE solution converges. In our simulations and real data analysis, we do not observe any drift.
2. The algorithm is implemented with a range of choices for the parameter $\delta$. The resulting curve is a discrete curve that can be represented by $\gamma_{\delta}(t_i), i = 0, \ldots, N$.

4 Simulations

To illustrate the fixed boundary flow, we simulated several noisy datasets on the unit sphere and the cone. We remark that the fixed boundary flow is a discrete curve with a derivative that approximately captures the direction of maximum local variation.

In the first part of simulation, we determine a curve on the unit sphere. These true curves are plotted in purple solid lines in Figure 4. The datasets are randomly drawn from the true curves. Specifically, Gaussian noise is added to the points on the true curves with a constraint so that the perturbed points remain on the test manifold. In this manner, we generate three datasets representing different types of variation. The first dataset is concentrated around a “C”-shaped curve on the unit sphere, thus the dataset presents a variation pattern along the geodesic. After that, we consider two datasets obtained from two nonconvex closed curves. In this setting, the simulated dataset presents a local variation pattern along the nonconvex curve. The second dataset is generated from a part (about one sixth) of the six-folded star shape, and the third dataset is concentrated around a part (about half) of the two-folded curve.

We fit the fixed boundary flows using a fixed starting point $\bar{x}_1$ and ending point $\bar{x}_2$, and the initial curves used for the simulations are shown in purple dotted lines in Figure
In our simulation study, the parameter $h$ is fixed and the parameter $\delta$ is set to take different values. For each pair of parameters $(h, \delta)$, we obtained a fixed boundary flow, illustrated by the red dotted line in Figure 4. The numerical results indicate that the performance of the fixed boundary flows captures the local variation of the true curve well, even when the true curve is nonconvex. In contrast to the numerical results obtained from the level set methods based on the variational principal flow method (Figure 9 in [19]), fixed boundary flows for the closed curves are able to follow the local variation at parts of the curve with high curvature. In addition, we have generated a noisy dataset for the whole six-folded star shape. The similar analysis can be found in the Supplementary Materials.
Figure 4: Fixed boundary flows (in red) for three different sets of simulated data on the unit sphere. (a)-(c) noisy “C”-shaped data, fixed boundary flow with $h < \infty$ and initial curve (in blue). (d)-(f) noisy six-folded data, fixed boundary flow with $h < \infty$ and initial curve (in blue). (g)-(i) noisy two-folded data, fixed boundary flow with $h < \infty$ and initial curve (in blue). The kernel function $\kappa_h(x, y) = 1(\|x - y\| \leq h)$ is used to calculate the local covariance.

In the second part of simulation, the testing manifold is extended to a right-circular
unit cone, with apex at \((0, 0, 0)\), height \(H = 1\) and radius \(R = 1\). Three sets of noisy data are generated to investigate the fixed boundary flows on the cone. The first dataset is concentrated around a band on the cone and the fixed boundary flows with two different values for parameter \(\delta\) are shown in red dotted lines in Figure 5(b)-(c). For the second and third dataset, they are generated from a “C”-shaped curve and “S”-shaped curve on the right-circular unit cone correspondingly. In Figure 5(e)-(f) and (h)-(i), the fixed boundary flows fit with two different values for parameter \(\delta\) are illustrated in red dotted lines.
Figure 5: Fixed boundary flows (in red) for three different sets of simulated data on the right-circular unit cone. (a)-(c) noisy band data, fixed boundary flow with $h < \infty$ and initial curve (in blue). (d)-(f) noisy “C”-shaped data, fixed boundary flow with $h < \infty$ and initial curve (in blue). (g)-(i) noisy “S”-shaped data, fixed boundary flow with $h < \infty$ and initial curve (in blue). The kernel function $\kappa_h(x, y) = 1(\|x - y\| \leq h)$ is used to calculate the local covariance.

5 Fixed Boundary Principal Flow in Euclidean Space

The purpose of this section is to prove that fixed boundary flows are canonical, in the sense that they will pass through the usual principal component, in the context of Eu-
clidean spaces. Hereafter, we suppose $\mathcal{M}$ is a linear subspace of $\mathbb{R}^d$, and $h = \infty$, which implies
\[
\Sigma_h(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)(x_i - x)^T.
\]
Under this configuration, we will figure out the supremum of $\mathcal{L}(W;\gamma)$ defined in (5) subjected to the constraint $\gamma \in \Gamma(\bar{x}_1, \bar{x}_2)$ defined in (3).

**Proposition 5.1.** Suppose $\gamma_* : [0, r] \to \mathcal{M}$ such that
\[
\gamma_*(0) = \bar{x}_1 \text{ and } \dot{\gamma}_*(t) = W(\gamma_*(t)) \text{ for } t \in [0, 1].
\]
If $\gamma_*(1) = \bar{x}_2$, then $\gamma_*$ is the unique optima of (4).

**Proof.** Since $W(\gamma(t))$ and $\dot{\gamma}(t)$ are units for any $\gamma \in \Gamma(\bar{x}_1, \bar{x}_2)$, we have
\[
\sup_{\gamma} \mathcal{L}(W;\gamma) = \sup_{\gamma} \int_{0}^{1} \langle \dot{\gamma}(t), W(\gamma(t)) \rangle \leq \int_{0}^{1} \|\dot{\gamma}(t)\| dt = 1,
\]
and the equation holds only if $\dot{\gamma}(t) = W(\gamma(t))$ for any $t \in [0, 1]$. Hence, $\gamma_*$ is the only curve that enables the equation to hold, and is thereby the unique optima of (4). \qed

Proposition 5.1 analyzes the optima of (4) under a strict condition that $\gamma_*(1) = \bar{x}_2$. When this condition is not satisfied, things are more difficult. For further analysis, we suppose the original point of $\mathcal{M}$ to be $\bar{x}_1$ and $[v_1, \cdots, v_d]$ to be the basis with $v_1 = W(\bar{x})$. For convenience, we denote $z_i = v_i^T x$ to be the $i$-th coordinate of any $z \in \mathcal{M}$ and $V_1 = [v_2, \cdots, v_d] \in \mathbb{R}^{d \times (d-1)}$ hereafter.

Before giving our final proposition, we define some important sets and curves first. With $\odot$ representing point-wise multiplication, we denote two subsets of $\Gamma(\bar{x}_1, \bar{x}_2)$ as
\[
\Gamma_+(\bar{x}_1, \bar{x}_2) = \{ \gamma \in \Gamma(\bar{x}_1, \bar{x}_2) : \dot{\gamma}(t) \odot W(\gamma(t)) \geq 0 \text{ for any } t \},
\]
and
\[
\Gamma_+(\bar{x}_1, \bar{x}_2, a) = \{ \gamma \in \Gamma(\bar{x}_1, \bar{x}_2) : v_1^T \gamma(t) = a, \text{ and } \dot{\gamma}(t) \odot W(\gamma(t)) \geq 0 \text{ for any } t \}.
\]
Denote $p_1 = v_1 v_1^T \bar{x}_1$ and $p_2 = v_1 v_1^T \bar{x}_2$ respectively. We set $\gamma^* : [0, \|p_1 - p_2\|] \to \mathcal{M}$ as the straight line between $p_1$ and $p_2$, that is $\gamma^*(t) = p_1 + \frac{t}{\|p_2 - p_1\|}(p_2 - p_1)$, and we also set
\[
\gamma_1^* = \arg\sup_{\gamma \in \Gamma_+(\bar{x}_1, p_1, v_1^T p_1)} \mathcal{L}(W;\gamma), \quad \gamma_2^* = \arg\sup_{\gamma \in \Gamma_+(p_2, \bar{x}_2, v_1^T p_2)} \mathcal{L}(W;\gamma).
\]

Let $\gamma_s$ be the concatenation of $\gamma_1^*$, $\gamma^*$ and $\gamma_2^*$, that is $\gamma_s : [0, \ell(\gamma_1^*) + \ell(\gamma^*) + \ell(\gamma_2^*)] \to \mathcal{M}$ satisfying
\[
\gamma_s(t) = \begin{cases} 
\gamma_1^*(t) & 0 \leq t \leq \ell(\gamma_1^*) \\
\gamma_1^*(\ell(\gamma_1^*)) + \gamma^*(t) & \ell(\gamma_1^*) < t \leq \ell(\gamma_1^*) + \ell(\gamma^*) \\
\gamma_1^*(\ell(\gamma_1^*) + \gamma^*(t)) + \gamma_2^*(t) & \ell(\gamma_1^*) + \ell(\gamma^*) < t \leq \ell(\gamma_1^*) + \ell(\gamma^*) + \ell(\gamma_2^*) \end{cases},
\]
then $\gamma_s$ is continuous and follows Lemma 5.1.
Lemma 5.1. If $\ell(\gamma_1^s) + \ell(\gamma^s) + \ell(\gamma_2^s) \leq 1$, then $\gamma_s$ belongs to the closure of $\bar{\Gamma}_+(\bar{x},\bar{x}_2)$.

Figure 6: Vector field satisfying Assumption 5.1 (blue arrows), flows satisfying $\dot{\gamma}(t) \odot W(\gamma(t)) \geq 0$ (red curves) and $\gamma_s$ (yellow curve).

We demonstrate flows belonging to $\Gamma_+(\bar{x}_1,\bar{x}_2)$ in Figure 6, which displays the cross sectional area of $\mathcal{M}$ along the first and $i$-th axis. The red curves represent three kinds of flows satisfying $\dot{\gamma}(t) \odot W(\gamma(t)) \geq 0$, and thereby belong to some $\Gamma_+(\bar{x}_1,\bar{x}_2)$ with certain $(\bar{x}_1,\bar{x}_2)$’s. The orange curve is the $\gamma_s$. It is continuous but not derivable, and thus it belongs to $\bar{\Gamma}_+(\bar{x}_1,\bar{x}_2)$ instead of $\Gamma_+(\bar{x}_1,\bar{x}_2)$.

In Figure 6, we also use the blue arrows to demonstrate the vector field that we will focus on. Generally speaking, this refers to the arrows at the left half plane point towards $\bar{x}$ and arrows at the right half plane point oppositely. Moreover, the arrows straighten horizontally as they approach the second axis. We summarize the assumptions on the vector field in Assumption 5.1 (b) and (c), and we conduct our discussion under this setting.

Assumption 5.1.
(a) $v^T \bar{x}_1 < 0$ and $v^T \bar{x}_2 > 0$.
(b) For any $x \in \mathcal{M}$, $v_i^T W(x) \geq 0$ and $(v_i^T W(x)) \ast (v_i^T x) \ast (v_i^T x) \geq 0$ for any $i \geq 2$.
(c) Suppose $x$ and $x'$ are in $\mathcal{M}$. If $V_i^T x = V_i^T x'$ and $|v_i^T x| \leq |v_i^T x'| \leq \max\{|v_i^T \bar{x}_1|, |v_i^T \bar{x}_2|\}$, then $|v_i^T W(x)| \leq |v_i^T W(x')|$ for any $i \geq 2$.

Now, we are ready to give the second proposition. This proposition shows that if we restrict $\gamma$ in $\Gamma_+(\bar{x}_1,\bar{x}_2)$, the fixed boundary flow will pass through the usual principal component.

Proposition 5.2. If $\ell(\gamma_s) \leq 1$, then

$$\mathcal{L}(W,\gamma_s) = \sup_{\gamma \in \Gamma_+(\bar{x}_1,\bar{x}_2)} \mathcal{L}(\gamma_s)$$
Proof. For any \( \gamma \in \Gamma_+(\bar{x}_1, \bar{x}_2) \), there exists \( t_0 \) such that \( v_1^T(\gamma(t_0)) = 0 \), since \( v_1^T(\gamma(0)) = v_1^T \bar{x}_1 < 0 \) and \( v_1^T(\gamma(r)) = v_1^T \bar{x}_2 > 0 \). Define \( \gamma_1 : [0, t_0] \to \mathcal{M} \) by \( \gamma_1(t) = \gamma(t) + v_1v_1^T(\gamma(0) - \gamma(t)) \) and \( \gamma_2 : [0, r - t_0] \to \mathcal{M} \) by \( \gamma_2(t) = \gamma(t - t_0) + v_1v_1^T(\gamma(t) - \gamma(t - t_0)) \). It is easy to verify that \( v_1^T \gamma_1(t) = v_1^T \gamma_2(t) = v_1^T \gamma(t) \) and thereby

\[
0 \leq \langle V_1^T \dot{\gamma}(t), V_1^T W(\gamma(t)) \rangle \leq \langle V_1^T \dot{\gamma}(t), V_1^T W(\gamma_1(t)) \rangle = \langle V_1^T \dot{\gamma}(t), V_1^T W(\gamma_1(t)) \rangle
\]

by Assumption 5.1 (b) and (c).

Considering that \( v_1^T \gamma_1(t) \) and \( v_1^T \gamma_2(t) \) are constant, we have \( v_1^T \gamma_1(t) = v_1^T \gamma_2(t) \) and thereby

\[
\int_0^{t_0} \langle V_1^T \dot{\gamma}(t), V_1^T W(\gamma(t)) \rangle dt \leq \int_0^{t_0} \langle V_1^T \dot{\gamma}_1(t), V_1^T W(\gamma_1(t)) \rangle dt = \int_0^{t_0} \langle \dot{\gamma}_1(t), W(\gamma_1(t)) \rangle dt = \mathcal{L}(W, \gamma_1) \leq \mathcal{L}(W, \gamma_0) + \mathcal{L}(W, \gamma_0),
\]

where \( \gamma_0 = \arg \sup_{\gamma \in \Gamma_+(\gamma_0), p_1, v_1^T p_1} \mathcal{L}(W, \gamma) \). Since we can always reparameterize \( \gamma_1 \) to be a unit speed one, say \( \tilde{\gamma}_1 \), the concatenate of \( \tilde{\gamma}_1 \) and \( \gamma_0 \) belongs to \( \Gamma_+(\bar{x}_1, p_1, v_1^T p_1) \). Hence,

\[
\int_0^{t_0} \langle V_1^T \tilde{\gamma}_1(t), V_1^T W(\gamma(t)) \rangle dt \leq \mathcal{L}(W, \gamma_1) + \mathcal{L}(W, \gamma_0) \leq \mathcal{L}(W, \gamma_1^*)
\]

We can similarly verify that

\[
\int_0^{t_0} \langle V_1^T \tilde{\gamma}_2(t), V_1^T W(\gamma(t)) \rangle dt \leq \mathcal{L}(W, \gamma_2).
\]

Moreover, by \( \|W(\gamma(t))\| = 1 \), we have \( v_1^T W(\gamma(t)) \leq 1 \) and thereby

\[
\int_0^{r} v_1^T \dot{\gamma}(t) \cdot v_1^T W(\gamma(t)) dt \leq \int_0^{r} v_1^T \dot{\gamma}(t) dt \leq \mathcal{L}(W, \gamma^*).
\]

Hence,

\[
\int_0^{r} \langle \dot{\gamma}(t), W(\gamma(t)) \rangle dt = \int_0^{r} (v_1^T \dot{\gamma}(t) \cdot v_1^T W(\gamma(t)) + \langle V_1^T \dot{\gamma}(t), V_1^T W(\gamma(t)) \rangle) dt \leq \mathcal{L}(W, \gamma^*) + \mathcal{L}(W, \gamma_1^*) + \mathcal{L}(W, \gamma_2^*) = \mathcal{L}(W, \gamma_s).
\]

Since \( \gamma_s \in \Gamma_+(\bar{x}_1, \bar{x}_2) \), the supremum can be achieved, which completes the proof. \( \square \)

Combining with Proposition 5.2, the inequality

\[
\sup_{\gamma \in \Gamma_+(\bar{x}_1, \bar{x}_2)} \mathcal{L}(W, \gamma) \geq \sup_{\gamma \in \Gamma(\bar{x}_1, \bar{x}_2) / \Gamma_+(\bar{x}_1, \bar{x}_2)} \mathcal{L}(W, \gamma),
\]

concludes that the optimal solution of (4) always passes through the usual principal component.
Next, we will discuss the inequality (19). Actually, if \( \gamma \in \Gamma(\bar{x}_1, \bar{x}_2)/\Gamma_+ (\bar{x}_1, \bar{x}_2) \) satisfies \( \langle v^T_1 \dot{\gamma}(t), v^T_1 W(\gamma(t)) \rangle \geq 0 \), then we define \( \gamma_+ \) by \( \{v^T_1 \gamma_+(t)\}_{i=1}^m \). We specially set \( v^T_1 \gamma_+(t) = v^T_1 \gamma(t) \), and for \( i \geq 2 \) we set

\[
v^T_i \gamma_+(t) = \begin{cases} 
\min \{ \max_{s \leq t} \gamma(t), 0 \} & v^T_i \bar{x}_1 \leq 0, t \leq t_0 \\
\min \{ \max_{s \geq t} \gamma(t), 0 \} & v^T_i \bar{x}_1 \leq 0, t \geq t_0 \\
\max \{ \min_{s \leq t} \gamma(t), 0 \} & v^T_i \bar{x}_1 \geq 0, t \leq t_0 \\
\max \{ \min_{s \geq t} \gamma(t), 0 \} & v^T_i \bar{x}_1 \geq 0, t \geq t_0 
\end{cases}
\] (20)

where \( t_0 \) is defined in the proof of Proposition 5.2. Using Assumption 5.1 (b), we can verify that \( \dot{\gamma}_+(t) \circledcirc W(\gamma_+(t)) \geq 0 \).

Figure 7: Diagram of \( \gamma \) (the blue curve) and \( \gamma_+ \) (the red curve) in \( \mathcal{M} \). The orange and pink curves are segments of \( \gamma \) and the yellow and purple curves are segments of \( \gamma_+ \).

In Figure 7, we display the cross sectional area of \( \mathcal{M} \) along the first and i-th axis for \( i \geq 2 \). In the left panel, the blue curve is \( \gamma \) and the red curve is \( \gamma_+ \). Without loss of generality, we focus on \( v^T_i \bar{x}_1 \leq 0 \) and \( t \leq t_0 \). The other three cases in (20) can be similarly verified.

First, we compare the integrals over the orange curve \( C_1 \) and the yellow curve \( C_2 \). Then the integral on \( C_1 \) denoted by \( I_1 \) is

\[
I_1 = \int_{t_1}^{t_2} v^T_i \dot{\gamma}(t) \cdot v^T_i W(\gamma(t)) \, dt = \int_{t_1}^{t_2} v^T_i W(\gamma(t)) v^T_i d\gamma(t) = \int_{C_1} v^T_i W(z) \, dz,
\]

and \( I_2 = \int_{C_2} v^T_i W(z) \, dz \). Then, \( I_1 - I_2 \) is the integral of \( v^T_i W(z) \) over the closed anti-clockwise curve consisting of \( C_1 \) and the inverse of \( C_2 \). When \( d = 2 \), such integral is equal to an integral over the gray region denoted by \( D \) shown in the right panel of Figure 7 by Green Theorem, that is,

\[
I_1 - I_2 = \iint_D \frac{\partial v^T_i W(z)}{\partial z_1} \, dz_1 d\bar{z}_2 \leq 0,
\]

19
since \( \frac{\partial v^T W(z)}{\partial z_1} \leq 0 \) for \( z_1 \leq 0 \) based on Assumption 5.1(c). For \( i \geq 2 \), if \( \frac{\partial v^T W(z)}{\partial z_1} \geq 0 \) holds for any \( j > i \) and \( \frac{\partial v^T W(z)}{\partial z_1} < 0 \) holds for any \( j < i \), the conclusion can be extended to a higher dimension by the Stokes’ theorem.

Second, we compare the integrals over the purple and pink curve. By Assumption 5.1(b), the integral of \( v^T_i W \) over the purple curve is negative, while the integral over the pink curve is zero. So, the integral of \( v^T_i W \) over the purple curve is less than the pink curve.

The above discussion summarizes
\[
\int_{t_0}^{t} v^T_i \dot{\gamma}(t) \cdot v^T_i W(\gamma(t)) dt \leq \int_{t_0}^{t} v^T_i \dot{\gamma}_+(t) \cdot v^T_i W(\gamma_+(t)) dt,
\]
for any \( i \geq 2 \). Moreover,
\[
\int_{t_0}^{t} \langle V^T_\perp \dot{\gamma}(t), V^T_\perp W(\gamma(t)) \rangle dt \leq L(W, \gamma_+^*),
\]
where the last inequality can be verified by similar proof of Proposition 5.2. Implementing the above discussion for \( t \geq t_0 \) analogically, we also have
\[
\int_{t_0}^{t} \langle V^T_\perp \dot{\gamma}(t), V^T_\perp W(\gamma(t)) \rangle dt \leq L(W, \gamma_+^*).
\]
Along with (18) we conclude
\[
\mathcal{L}(W, \gamma) = \int_{t_0}^{t} v^T_i \dot{\gamma}(t) \cdot v^T_i W(\gamma(t)) dt + \sum_{i=2}^{m} \int_{t_0}^{t} v^T_i \dot{\gamma}_+(t) \cdot v^T_i W(\gamma_+(t)) dt
\]
\[
\leq \mathcal{L}(W, \gamma^*) + \mathcal{L}(W, \gamma^*) + \mathcal{L}(W, \gamma^*) = \sup_{\gamma \in \Gamma_+([\bar{x}_1, \bar{x}_2])} \mathcal{L}(W, \gamma),
\]
which supports the inequality (19).

6 Seismological Data

In this section, we consider a concrete case – an earthquake dataset from the International Seismological Center (ISC). The dataset contains data of large earthquake occurrences (magnitude 5.5 and above, including continental events of magnitude 5.0) from 1904 to 2015. Figure 8 shows the large earthquakes that occurred within the region around the geologic junction where the boundaries of four tectonic plates intersect: the North American Plate, the Eurasian Plate, the Philippine Sea Plate and the Pacific Plate. As expected, earthquakes tend to occur around the tectonic plate boundaries and the shape of the plate boundary carries the global variation (from east to west or north to south) and the localized variation.
As expected, the fixed boundary flows move along the plate boundary given that the starting point and ending point are around a specific plate boundary. For example, we fixed the starting point and ending point near the Philippine Sea Plate. The movement of the fixed boundary flows will reflect the local variation pattern of the data, which is captured by the scale parameter $h$ with the projection modification. As we can see in Figure 8, most of the data points are concentrated at the Japan Trench. For the implement of the projection procedure, one of the things that we noticed was that the projected points for the initial curve (in blue) were likely to stay around the Japan Trench. As shown in Figure 9, the vector field calculated from the local covariance matrix around the Japan Trench moves along the Pacific Plate. This direction of the vector field is not suitable for the region that we are interested in. Thus, we modified the projection procedure with truncated earthquake data instead of the whole dataset. Given the initial curve $\gamma^0$, a truncate parameter $h^*$ is used to obtain the data points in the neighborhood $h^*$ of the initial curve. After that, the points from the initial curve or the intermediate curves are projected to the truncated dataset. By making this modification, we are able to capture the localized variation pattern around the region of interest – the Philippine Sea Plate.

From the numerical results in Figure 10(a) and 10(b), this particular earthquake dataset can serve as an example to illustrate the effect of choosing parameter $h$, which reflects the localized variation pattern that we are able to capture from the dataset.
fit fixed boundary flows with different values for the parameter $h$. When $h$ is relatively large (595 miles), the fixed boundary flow is quite flat, since the vector field does not vary significantly. On the contrary, a small $h$ (396 miles) leads to a fixed boundary flow compatible with the plate boundary of the Philippine Sea Plate, which is shown in Figure 10(a).

In addition, we fit principal flows for the earthquake dataset. As shown in Figure 10 (c)-(d), the principal flows passing through weighted means roughly move along the plate boundary of the Pacific Plate, and they fail to describe the plate boundary of interest for the Philippine Sea Plate. Principal flows passing through the Fréchet mean can be found in the Supplementary Materials. In Figure 11, we present the principal flow and fixed boundary flow on a flat atlas.

Figure 9: Calculation of vector field given initial curve (in blue). Green dots: projected points of the point from the initial curve. Green arrow: vector field calculated from the neighborhood of the projected points.
Figure 10: Analysis for the earthquake data. (a) Fixed boundary flow for $h = 396$ miles (in red). (b) Fixed boundary flow for $h = 595$ miles (in red). (c) Principal flow for $h = 396$ miles (in purple), starting at a weighted mean (in red). (d) Principal flow for $h = 595$ miles (in purple), starting at a weighted mean (in red).
7 Discussion

Compared to the principal flow, the determination of a fixed boundary flow for data points on non-linear manifolds is a very different problem. We have proposed the notion of a fixed boundary flow to define a curve with fixed starting and ending points and a tangent velocity that matches the maximal variation of data in its neighborhood at each point. The local geometry of data variation is represented by the tangent space at the given point, which motivates us to use the local vector fields. Based on this inspiration, we formulate an optimization framework to construct such a smooth curve on the manifold, with a tangent vector that always matches the local vector fields. There is no doubt that the solution to the optimization problem, and equivalently, the fixed boundary flow, depends on how a neighborhood is defined at a certain point, just like the principal flow. When all data points are selected to calculate the local vector fields for any point on the flow (locality parameter $h = \infty$), a segment of the fixed boundary flow would coincide with the principal component under some assumptions, according to Proposition 5.2. With this being said, in practice, one would need to deal with the neighborhood with a finite $h$. In this case, we will numerically compute the fixed boundary flow under a DAE as outlined in Section 3. Experiments on simulations and the seismological data set have demonstrated that the output of our algorithm is roughly similar to our expected fixed boundary flow.

The choice of the neighborhood depending on the scale parameter $h$ determines how local or global covariation features are captured by the fixed boundary flow. If we take $h$ to be infinite like the PCA, the analysis in Section 5 shows that a segment of the fixed boundary flow coincides with the principal component under some assumptions.
Figure 12 illustrates (b) and (c) of Assumption 5.1 with two datasets represented by black points and concentrated around a “C”-shaped curve and a “S”-shaped curve in \( \mathbb{R}^2 \). The two diagrams in the left panel show the vector fields for the two datasets, both of which satisfy Assumption 5.1(b), and the diagrams in the right panel show how \( |v_2^T W(x)| \) varies at different points of \( x \). Specifically, \( |v_2^T W| \) gets larger when the color transitions to yellow, and gets smaller when the color transitions to blue. One can tell from the two diagrams in the right panel that the vector field between the two orange lines satisfies Assumption 5.1(c).

On the other hand, if we take \( h \) to be some finite scale, the concerned problem becomes related to the selection of a suitable \( h \) for each point \( x \). There are potentially several ways to quantify the choice of \( h \). A possible way is to find \( h \) at the point where the scale of the fixed boundary flow picks up the most significant one-dimensional variation, that is, at the point where the first eigenvalue of \( \Sigma_h(x) \), say \( \lambda_1 \), is maximal compared to the other eigenvalues, say \( \lambda_2, \cdots, \lambda_m \). Thus, we conceive a measure of the form

\[ \rho = \frac{\sum_{i=2}^{m} \lambda_i}{\lambda_1}, \quad (21) \]

and choose \( h \) to minimize \( \rho \). The performance of this measure is demonstrated in Figure 13. In this case, ten different \( h \in [0.05, 0.5] \) with step size of 0.05 are chosen. The green circles represent the \( \mathcal{N}(x, h) \) corresponding to different \( h \)'s (right). For each \( h \), the measure \( \rho \) is calculated, and we observe that \( \rho \) reaches its minimum when \( h = 0.25 \). We display the neighborhood of \( h = 0.25 \) using the blue circle. As we can see, the samples
in the blue circle are roughly distributed along the horizontal direction, which coincides with the expected vector field at \(x\). We further plot a red ellipsoid centered at \(x\) with the major semi-axis and minor semi-axis being \(he_1\) and \(\sqrt{\lambda_2}he_2\), respectively. Clearly, taking the vector field \(W(x) = e_1(x)\) is what one would expect. From the Figure 13, we observe that \(\rho\) is not very sensitive to \(h\); that is, any \(h \in [0.15, 0.35]\) gives a suitable neighborhood, within which the samples are distributed around the expected vector field at \(x\). That is the reason why we have used a fixed \(h\) for all points in the fixed boundary for our experiments in Sections 4 and 6.

Figure 13: Left: illustration of how \(\rho\) in (21) varies with different \(h\). Right: \(\mathcal{N}(x, h)\) covered by green circles with these \(h\) values.

The formulation of \(\{\lambda_i\}_{i=1}^m\) can certainly be extended to constructing a “confidence band” for the population fixed boundary flow \(\gamma \in \mathcal{M}\). As we define the confidence band for the flow on the manifold, it should be a confidence ellipsoid. Note that the samples in \(\mathcal{N}(x, h)\) roughly lie within an ellipsoid with principal axes of length \(h\), \(\sqrt{\lambda_2}h\), \(\sqrt{\lambda_3}h\), \ldots, \(\sqrt{\lambda_m}h\), respectively. Thus, for any point \(x = \tilde{\gamma}(t)\) on the computed fixed boundary flow \(\tilde{\gamma}\), we can define an ellipsoid of dimension \((m-1)\) in the intersection of \(T_x\mathcal{M}\) and the normal space at \(x\), which could cover most samples in this intersection. By letting the orthonormal \(U(x) \in \mathbb{R}^{d\times m}\) to be a basis of \(T_x\mathcal{M}\), the confidence ellipsoid is of dimension \((m-1)\) obeying

\[
\left\{ \Pi_{\mathcal{M}}z : (z - \hat{\gamma}(t))^T V \text{ diag } (\frac{\lambda_1}{\lambda_2^2h^2}, \ldots, \frac{\lambda_1}{\lambda_m^2h^2}) V^T (z - \hat{\gamma}(t)) = 1 \right\},
\]

where \(\Pi_{\mathcal{M}}z\) is the projection of \(z\) onto \(\mathcal{M}\) and \(V = (U(\tilde{\gamma}(t))U(\tilde{\gamma}(t))^T - \hat{\gamma}(t)\hat{\gamma}(t)^T)(e_2, \ldots, e_m)\). Note that \(U(x)\) can be numerically estimated with certain theoretical guarantee (see [27]).

We remark that \(\hat{\gamma}(t)\) usually approximates \(W(\tilde{\gamma}(t))\), that is, \(e_1\). This makes \(V\) full column rank, and thereby the dimension of the ellipsoid is \((m-1)\). If \(\hat{\gamma}(t)\) is accidentally orthogonal to \(W(\tilde{\gamma}(t))\), the dimension of the ellipsoid would reduce to \((m-2)\). With certain covering ellipsoid conditions for the samples in the neighborhood, one might con-
sider bounding $\dot{\gamma}$ and $\ddot{\gamma}$ under the current setting. Some of the results in [28] will be helpful in this respect. As this is one of our ongoing works, we will investigate it in the future.

8 Acknowledgements

We thank Professor Victor Panaretos for suggesting this problem. We are grateful for the financial support from the MOE Tier 1 funding (R-155-000-186-114) and Tier 2 funding (R-155-000-184-112) at the National University of Singapore.

APPENDIX

Proofs of Other Formal Statements.

Proof of Proposition 3.1. First, we prove that a curve $\gamma$ in the set $\Gamma$ can be reparameterized to yield a curve in the set $\Gamma_2$. Given any $\gamma \in \Gamma(x_1, x_2)$, let

$$\alpha = \frac{1 + \sqrt{1 - r^2}}{r^2},$$

and define

$$f : [0, 1] \rightarrow \mathcal{M}, \quad f(t) = \gamma(rt), \quad \text{for } 0 \leq t \leq r.$$

Then, we have $\|f'(t)\| = r\|\dot{\gamma}(rt)\| = r$ for $0 \leq t \leq 1$, and $\int_0^1 \|f'(t)\|^2 dt = r^2$. Furthermore, we define

$$\gamma_* : [0, 1] \rightarrow \mathcal{M}, \quad \gamma_*(t) = f(t^\alpha),$$

so that $\|\dot{\gamma}_*(t)\| = r\alpha t^{\alpha - 1}$. It follows that

$$\int_0^1 \|\dot{\gamma}_*(t)\|^2 dt = \int_0^1 r^2 \alpha^2 t^{2\alpha - 2} dt = \frac{r^2 \alpha^2}{2\alpha - 1} = 1,$$

noting that the reparameterization is always possible since $r \leq 1$ and we may choose $\gamma \geq 1$ such that $r^2 \alpha^2 = 2\alpha - 1$ holds. For the boundary condition, we have $\gamma(0) = x_1$ and $\gamma(r) = x_2$. When $t = 0$, we have $rt^\alpha = 0$ and $\gamma_*(0) = \gamma(0) = x_1$. For $t = 1$, we have $\gamma_*(1) = \gamma(r) = x_2$.

Then, we show that a curve in the set $\Gamma_2$ can be reparameterized to yield a curve in the set $\Gamma$. Let $\gamma_2 \in \Gamma_2(x_1, x_2)$, then we have

$$\left(\int_0^1 \|\dot{\gamma}_2\| dt\right) \leq \int_0^1 \|\dot{\gamma}_2\|^2 dt = 1,$$

from the Cauchy-Schwarz inequality. It follows that $\ell(\gamma_2(t)) \leq 1$, and so it can be reparameterized to have unit speed. Unit speeds and curve lengths less than or equal to 1 ensure that $r \in (0, 1]$. \qed
Proof of Lemma 5.1. For simplicity, we denote $\bar{S}$ to be the closure of a set $S$. Let $P$ be the set of polynomials and $C$ be the set of continuous functions. Based on the Stone-Weierstrass Theorem, we have $P \subset C^2 \subset \bar{P} = C$, which implies that the closure of $C^2$ is $C$. Based on this conclusion, we have

\[
\bar{\Gamma}(\bar{x}_1, \bar{x}_2) = \{ \gamma : [0, r] \to \mathcal{M} : \gamma \in C([0, r]), r \in (0, 1],
\gamma(0) = \bar{x}_1, \gamma(r) = \bar{x}_2, \gamma(s) \neq \gamma(s') \text{ for } s \neq s',
\ell(\gamma[0, t]) = t, \text{ for all } t \in [0, r]\},
\]

and

\[
\bar{\Gamma}_+(\bar{x}_1, \bar{x}_2) = \{ \gamma \in \Gamma(\bar{x}_1, \bar{x}_2) : \dot{\gamma}(t) \circ W(\gamma(t)) \geq 0 \text{ for any } t \}.
\]

Since $\gamma_s$ is continuous and satisfies $\dot{\gamma}(t) \circ W(\gamma(t)) \geq 0$, we conclude that $\gamma_s \in \bar{\Gamma}_+(\bar{x}_1, \bar{x}_2)$. □

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 214–223, 2017.

[2] J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. Computer Methods in Applied Mechanics and Engineering, 1(1):1–16, 1972.

[3] D. Bradley, D. Nowrouzezahrai, and P. Beardsley. Image-based reconstruction and synthesis of dense foliage. ACM Transactions on Graphics (TOG), 32(4):74:1–74:10, 2013.

[4] T. F. Cox and M. A. Cox. Multidimensional scaling. Chapman and hall/CRC, 2000.

[5] I. L. Dryden and K. V. Mardia. Statistical Shape Analysis. Wiley, Chichester, 1998.

[6] E. Eich. Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints. SIAM Journal on Numerical Analysis, 30(5):1467–1482, 1993.

[7] B. Eltzner, S. Huckemann, T. Hotz, and K. Mardia. Torus principal component analysis with applications to rna structure, accepted. Annals of Applied Statistics, 2017.

[8] P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing, 87:250–262, 2007.

[9] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Transactions on Medical Imaging, 23(8):995–1005, 2004.

28
[10] R. Gross. Face databases. In *Handbook of Face Recognition*, pages 301–327. Springer, New York, February 2005.

[11] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[12] S. Huckemann and H. Ziezold. Principal component analysis for riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38:299–319, 2006.

[13] S. Jung, I. L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99:551–568, 2012.

[14] P. E. Jupp and J. T. Kent. Fitting smooth paths to spherical data. *Journal of the Royal Statistical Society, Series C*, 36:34–36, 1987.

[15] K. Kenobi, I. L. Dryden, and H. Le. “shape curves and geodesic modelling. *Biometrika*, 97:567–584, 2010.

[16] M. Kilian, N. J. Mitra, and H. Pottmann. Geometric modeling in shape space. *ACM Transactions on Graphics (TOG)*, 26(3):64:1–64:8, 2007.

[17] A. Kume, I. L. Dryden, and H. Le. Shape-space smoothing splines for planar landmark data. *Biometrika*, 94:513–528, 2007.

[18] L. Ladicky and P. Torr. Locally linear support vector machines. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 985–992, 2011.

[19] H. Liu, Z. Yao, S. Leung, and T. F. Chan. A level set based variational principal flow method for nonparametric dimension reduction on riemannian manifolds. *SIAM Journal on Scientific Computing*, 39(4):A1616–A1646, 2017.

[20] Y. Ma and Y. Fu. *Manifold learning theory and applications*. CRC press, 2011.

[21] U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine learning research*, 12(Apr):1249–1286, 2011.

[22] V. M. Panaretos, T. Pham, and Z. Yao. Principal flows. *Journal of the American Statistical Association*, 109(505):424–436, 2014.

[23] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 137–143. IEEE, 1997.

[24] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
[25] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[26] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, pages 311–318. ACM, 1994.

[27] H. Tyagi, E. Vural, and P. Frossard. Tangent space estimation for smooth embeddings of riemannian manifolds. *Information and Inference: A Journal of the IMA*, 2(1):69–114, 2013.

[28] Z. Yao and Z. Zhang. Principal boundary on riemannian manifolds. *Journal of the American Statistical Association*, 2019. To appear.

[29] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004.