이학박사 학위논문

Nonparametric dimension reductions on Riemannian manifolds

리만다양체 상의 비모수적 차원축소방법론

2022년 8월

서울대학교 대학원

통계학과

이 종 민

Nonparametric dimension reductions on Riemannian manifolds

지도교수 오희석

이 논문을 이학박사 학위논문으로 제출함

2022년 4월

서울대학교 대학원

통계학과

이 종 민

이종민의 이학박사 학위논문을 인준함

2022년 5월

위 원 장 이 재 용 (인)

부위원장 오 희 석 (인)

위 원 임 채 영 (인)

위 원 이 우 주 (인)

위 원 임 예 지 (인)

Nonparametric dimension reductions on Riemannian manifolds

by

Jongmin Lee

A Thesis
submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
August, 2022

ABSTRACT

Nonparametric dimension reductions on Riemannian manifolds

Jongmin Lee
The Department of Statistics
The Graduate School
Seoul National University

Over the decades, parametric dimension reduction methods have been actively developed for non-Euclidean data analysis. Examples include Fletcher et al. (2004); Huckemann et al. (2010); Jung et al. (2011, 2012); Zhang et al. (2013). Sometimes the methods are not enough to capture the structure of data. This dissertation presents newly developed *nonparametric* dimension reductions for data observed on manifold, resulting in more flexible fits. More precisely, the main focus is on the generalizations of principal curves into Riemannian manifold. The principal curve is considered as a nonlinear generalization of principal component analysis (PCA). The dissertation consists of four main parts as follows.

First, the approach given in Chapter 3 lie in the same lines of Hastie (1984); Hastie and Stuetzle (1989) that introduced the definition of original principal curve on Euclidean space. The main contributions of this study can be summarized as follows: (a) We propose both extrinsic and intrinsic approaches to form principal curves on D-sphere S^D, $D \geq 2$. (b) We establish the stationarity of the proposed principal curves on S^D. (c) In extensive numerical studies, we show the usefulness of the proposed method through real seismological data and real Human motion capture data as well as simulated data on 2-sphere, 4-sphere.

Secondly, As one of further work in the previous approach, a robust nonparametric dimension reduction is proposed. To this ends, L_1- and Huber loss are used rather than L_2 loss. The contributions of this study can be summarized as follows: (a) We study robust principal curves on spheres that are resistant to outliers. Specifically, we propose absolute-type and Huber-type principal curves, which go through the median of data, to robustify the principal curves for a set of data which may contain outliers. (b) For a theoretical aspect, the stationarity of the robust principal curves is investigated. (c) We provide practical algorithms for implementing the proposed robust principal curves, which are computationally feasible and more convenient to implement.

Thirdly, An R package **spherepc** (Lee et al., 2022b) comprehensively providing dimension reduction methods on a sphere is introduced with details for possible reproducible research. To the best of our knowledge, no available R packages offer the methods of dimension reduction and principal curves on a sphere. The existing R packages providing principal curves, such as **princurve** (Hastie and Weingessel, 2015) and **LPCM** (Einbeck et al., 2015), are available only on Euclidean space. In addition, existing nonparametric dimension reduction methods on manifold involve somewhat complex intrinsic optimizations (Panaretos et al., 2014; Liu et al., 2017; Yao et al., 2019). The proposed R package **spherepc** provides the state-of-the-art principal curve technique on the sphere (Lee et al., 2021a) and comprehensively collects and implements the existing techniques (Fletcher et al., 2004; Jung et al., 2011; Hauberg, 2016).

Lastly, for an effective initial estimate of complex structured data on manifold, local principal geodesics are first provided and the method is applied to various simulated and real seismological data. For variance stabilization and theoretical investigations for the procedure, nextly, the focus is on the generalization of Kégl (1999); Kégl et al. (2000), which provided the new definition of principal curve on Euclidean space, into generic Riemannian manifolds. Theories including consistency and convergence rate of the procedure by means of empirical risk minimization principle, are further established on generic Riemannian manifolds. The consequences on the real data analysis and simulation study show the promising characteristics of the proposed approaches.

Keywords: Curve fitting, Dimensionality reduction, Principal curves, Riemannian manifolds, Reproducibility, Statistical learning theory
Student Number: 2018–25763

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Dimension reduction methods are widely used in various fields, including statistics and machine learning, by efficiently compressing data and removing noise (Benner et al., 2005). A variety of dimension reduction techniques have been developed to represent and analyze data on Euclidean space. Over the decades, there have been growing interests in the analysis of non-Euclidean data with extensive applications: directional data (Mardia and Gadsden, 1977; Gray et al., 1980; Jung et al., 2011; Mardia, 2014; Liu et al., 2017; Mallasto and Feragen, 2018), shape data (Kendall, 1984; Jung et al., 2012; Fletcher et al., 2004; Huckemann and Ziezold, 2006; Fletcher et al., 2009; Huckemann et al., 2010; Zhang and Fletcher, 2013; Mallasto and Feragen, 2018; Shin and Oh, 2022), symmetric positive definite matrix-valued data (diffusion tensor imaging) (Pennec et al., 2006; Fletcher and Joshi, 2007; Mallasto and Feragen, 2018), torus-valued data (Eltzner et al., 2018; Jung et al., 2021), and Lie group-valued data (Human kinetic study) (Hauberg, 2016; Mallasto and Feragen, 2018; Telschow et al., 2019). For example, Siddiqi and Pizer (2008) and Cippitelli et al. (2016) introduced a Cartesian product of sphere S^2 and $\mathbb{R}$ for medial representation and skeleton data, respectively. For these representations, the conventional dimension reduction methods on Euclidean space have been modified by considering geodesics on non-Euclidean space as in Fletcher et al. (2004); Huckemann and Ziezold (2006); Huckemann et al. (2010).

The meaning of *nonparametric* that is contained in the title of this dissertation is now given. To explain intuitively, we consider regression problem. In the regression problem, the general model is written as

$$Y = f(X) + \epsilon,$$

where X, Y, ϵ denote explanatory, response variables, and random noise respectively. In most cases, the aim is to find the mean function $m = \mathbb{E}[Y \mid X = \cdot]$ and a procedure for estimating the f is called as *nonparametric* when the functional form on f is not imposed. (Rigorously, the dimension of space of f is *infinite*). In our problem, for an M-valued random variable X and a continuous function $f : [0,\, 1] \to M$, $\lambda_f(X) \in [0,\, 1]$ is defined by the minimum parameter such that $f(\lambda_f(X))$ is the closest point in f from X; formally,

$$\lambda_f(X) = \min_{\lambda \in [0,\, 1]} \left\{ d(X,\, f(\lambda)) = \min_{\mu \in [0,\, 1]} d(X,\, f(\mu)) \right\}$$

where it is well defined due to the compactness of $[0,\, 1]$. In analogy, the model considered in the thesis can be written as

$$X = f(\lambda_f(X)) + \epsilon,$$

where ϵ, $+$ denote a random noise and vector addition respectively when $M = \mathbb{R}^D$; otherwise they need some care to be rigorously defined because M generally has no vector operation $+$, owing to its inherent nonlinearity. For a given M-valued random variable X, the aim is to find the principal curve f that goes through the middle of data, as

$$f = \mathbb{E}_X\big[X \mid \lambda_f(X) = \cdot\big],$$

where $\mathbb{E}_X$ is the expectation with respect to X. It is the definition of principal curve given later in Chapter 2. The model can be regarded as nonparamteric since the functional form on f is not assumed and hence the dimension of the collection of such f is infinite.

The rest of the dissertation is organized as follows. In Chapters 3 and 4, the interest lies in dimension reduction (curve fitting) for data on spheres. As a *nonparametric* and flexible way, Chapter 3 directly extended principal curves suggested by Hastie and Stuetzle (1989) onto spheres in both extrinsic and intrinsic ways with corresponding theoretical properties. As a study closely related to the chapters, Hauberg (2016) developed principal curves on Riemannian manifold. Hauberg (2016), however, uses an approximate method by projecting data onto a *finite* set of points, unlike the original principal curve in Hastie and Stuetzle (1989) which projects data onto a *continuous* curve. This approximate projection causes a problem that may project different data points onto a single point mistakenly. Chapter 3 proposes a new principal curve for sphere-valued data by projecting the data onto a continuous curve without any approximations, improving the performance of dimension reduction. The curve is called as *spherical principal curves* (SPC). The proposed approach in Chapter 3 is two-fold: One is an extrinsic approach that requires the setting of additional embedding space for a given manifold. The other is an intrinsic approach that does not need an embedding space. The chapter 3 investigates the stationarity of the principal curves on spheres in both extrinsic and intrinsic ways.

In Chapter 4, spherical principal curves given in Chapter 3 are robustified by using absolute loss and Huber loss. In literature, Hastie (1984) removed distant observations in estimating a robust curve; thus, it may lose data information. Banfield and Raftery (1992) modified the ordinary algorithm suggested by Hastie (1984); Hastie and Stuetzle (1989) to reduce the estimation bias when the curvature of the underlying curve highly varies. Tibshirani (1992) suggested a probabilistic definition of principal curves based on a Gaussian mixture model and applied an EM algorithm for estimation to alleviate bias. Since real data sometimes contain large noises, Stanford and Raftery (2000) proposed another probabilistic method for identifying the curvilinear features from data with background spatial noises. In addition, some studies in the field of directional statistics have provided robust point estimates for sphere-valued data, e.g. He and Simpson (1992). On the other hand, few studies have been conducted to study robust curve fitting methods for sphere-valued data. As a notable extension of principal component analysis on manifolds, Panaretos et al. (2014) proposed a smooth curve on the manifolds, termed as principal flow, which goes through the center of data and is estimated by using a nonparametric way with preserving a canonical interpretation of PCA. There are several related follow-up studies. For example, Liu et al. (2017) applied a level set-based approach to estimate flexible and robust curves. Yao et al. (2019) relaxed the constraint of boundary conditions on principal flows, and Yao and Zhang (2020) used a principal flow method to deal with a classification problem on manifold. However, these methods used variational approaches like the Euler-Lagrange equation involved with differential equations on manifold, making it rather difficult to reproduce the methodologies. As shown in Figure 1.1, the specific problem and objective considered in this study

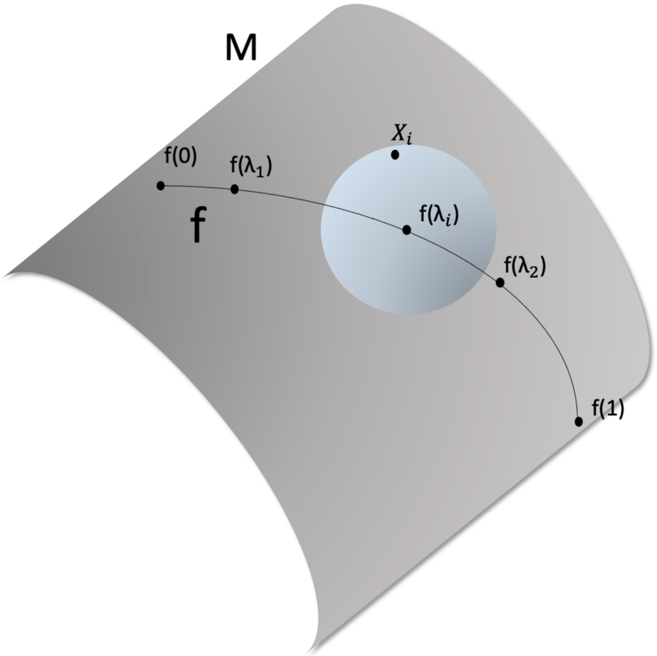


Figure 1.1: Process of data generating is illustrated. The population curve (ground-truth) $f : [0,\, 1] \to M$ is colored in black. A data point X_i is generated by adding a random noise (colored in blue) to $f(\lambda_i)$.

are as follows. Suppose that we observe $\{x_i\}_{i=1}^{n} \subset M$ generated from the population curve $f : [0,\, 1] \to M$. Specifically, $\{\lambda_i\}_{i=1}^{n}$ independently follow a uniform distribution $U[0,\, 1]$, and x_i are obtained by adding independent random noises to $f(\lambda_i)$,

$i = 1, 2, \ldots, n$. Chapter 4 further assumes that observations are generated from a single population curve f that is smooth and non-intersecting. The objective is to recover the population curve f from the observations. In most cases, however, the random noises are generally assumed to be Gaussian distributed, which is sometimes impractical in applications (Petrus, 1999; Huber, 2004). Chapter 4 deals with cases where the random noises follow from heavy-tailed distributions. To the best of our knowledge, this setting is not fully discussed in conventional curve fitting methods on manifolds, including Fletcher et al. (2004); Huckemann et al. (2010); Jung et al. (2012); Panaretos et al. (2014); Eltzner et al. (2018); Lee et al. (2021a). The contributions of Chapter 4 can be summarized as follows: (a) On spheres S^D for $D \geq 2$, study robust principal curves that are resistant to outliers. Specifically L_1-type and Huber-type principal curves, which go through the median of data, is introduced to robustify the principal curves for dataset which may contain outliers. (b) For a theoretical aspect, the stationarity of the robust principal curves is investigated. (c) Provide practical algorithms for implementing the proposed L_1-type and Huber-type principal curves, which are computationally feasible and convenient to implement.

Chapter 5 introduces an R package **spherepc** (Lee et al., 2022b) that considers several dimension reduction techniques on a sphere, which encompass recently developed approaches such as SPC and LPG as well as some existing methods, and discuss how to implement these methods through **spherepc**. The examples of existing methods are principal geodesic analysis (Fletcher et al., 2004), exact principal circle (Lee et al., 2021a), and principal curves proposed by Hauberg (2016). Hauberg (2016) proposed an algorithm to find the principal curves on manifolds. However, the principal curves proposed by Hauberg (2016) represent the data continuously because of the approximation of the projection step required to fit the curves. Recently, Lee et al. (2021a) proposed a new method, termed spherical principal curves (SPC), that constructs principal curves, ensuring a stationary property on spheres. SPC is useful for representing circular or waveform data with smaller reconstruction errors than conventional methods. In some cases, however, SPC has the disadvantage of being sensitive to initial estimate. As a result, there are some data structures where SPC is not efficient, for example, spiral, zigzag, tree-shaped data. To cope with such problem, a localized version of principal curve, called local principal geodesics (LPG), is developed. A function for LPG is also provided in the package **spherepc**.

To the best of our knowledge, no available R packages offer the methods of dimension reduction and principal curves on a sphere. The existing R packages providing principal curves, such as **princurve** (Hastie and Weingessel, 2015) and **LPCM** (Einbeck et al., 2015), are available only on Euclidean space, not on a sphere or a Riemannian manifold. The proposed package **spherepc** for R provides the state-of-the-art principal curve technique on the sphere (Lee et al., 2021a) and comprehensively collects and implements the existing techniques (Fletcher et al., 2004; Hauberg, 2016; Lee et al., 2021b).

Chapter 6 will present the newly developed method capable of identifying the structure of data that have complicated underlying structures and lie on Riemannian manifold. Specifically, principal curve method proposed by Kégl (1999); Kégl et al. (2000) is generalized to Riemannian manifold. Kégl (1999); Kégl et al. (2000) proposed nonparametric dimension reduction method. The motivating example is given in Figure 1.2, which illustrates the several procedures for data description when $M = \mathbb{R}^D$. The top panels of Figure 1.2 show the zero (mean) and one dimensional (first principal component) descriptions for the data, respectively. Note that the mean of data points is the point which minimizes the sum of squares of distances from the data points to itself and that (first) principal component is the line which minimizes the sum of squares of (orthogonal) distances from data points to itself. In this respect, Kégl (1999); Kégl et al. (2000) proposed a nonlinear description for data, length-constrained principal curve, which minimizes the distances from data points to the curve under a predetermined length constraint, as illustrated in the bottom panel of Figure 1.2. Along this line, Chapter 6 will develop the method on Riemannian manifold. The methodology is termed as *local principal curves on Riemannian manifold* (LPCRM). The procedure is also applied to various simulated

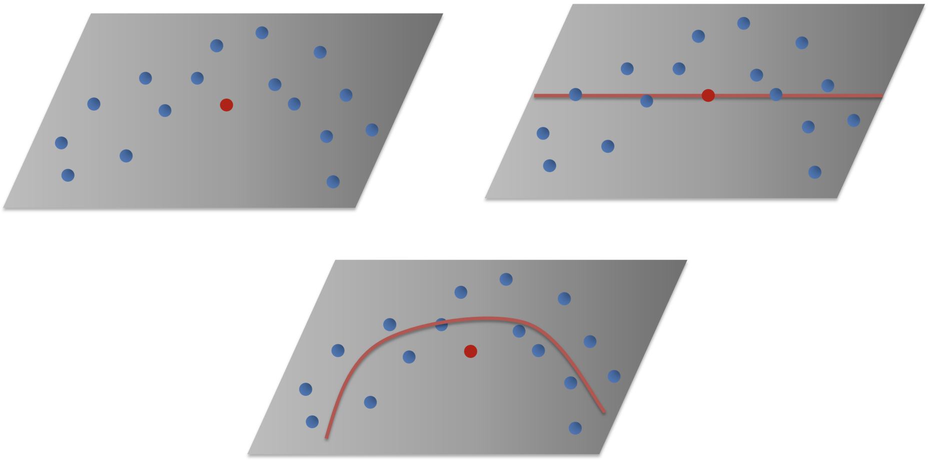


Figure 1.2: Data (blue) are distributed on $M = \mathbb{R}^D$ and three procedures are illustrated for data descriptions (red) based on minimization of least squares. Top left: Data are represented as their mean (red). Top right: Data are compressed as a linear line (first principal component). Bottom: Data are represented as a principal curve (red) which minimizes the sum of squares of distances from the data to curve under a predetermined length constraint.

data. Although this work is in progress, the theoretical characteristics including consistency, cubic-convergence rate, and non-asymptotic concentration inequality of the procedure are established by means of statistical learning theory.

Finally, Chapter 7 summarizes and emphasizes the contributions of the thesis again and discusses the future topics.

Chapter 2

Preliminaries

2.1 Principal curves

The principal curve firstly proposed by Hastie (1984); Hastie and Stuetzle (1989) can be considered as a nonlinear generalization of PCA that finds an affine subspace maximizing the variance of the projections of data. A curve is a function from one-dimensional closed interval to a given space, that is, $f : I \to \mathbb{R}^D$. A curve f is called self-consistent or a *principal curve* of a $\mathbb{R}^D$-valued random variable X if the curve satisfies

$$f(\lambda) = \mathbb{E}_X[X \mid \lambda_f(X) = \lambda], \tag{2.1}$$

where $\mathbb{E}_X$ is taken over with respect to X and

$$\lambda_f(x) := \min \left\{ \lambda \in I \mid \|x - f(\lambda)\| = \min_{\mu \in I} \|x - f(\mu)\| \right\}$$

is the projection index of a point x onto the curve f. The definition (2.1) means that $f(\lambda)$ is the average of all data points projected onto $f(\lambda)$ itself.

One of the most important consequences of the self-consistency is that the principal curve is a critical (stationary) point with respect to reconstruction error for small perturbations (Hastie, 1984; Hastie and Stuetzle, 1989). However, it is difficult to directly formulate a principal curve by solving the self-consistency equation of (2.1). Thus, Hastie and Stuetzle (1989) represented a curve as the first order spline, connected by T points. Then, they iteratively updated the curve to achieve the self-consistency condition using the following two steps, *projection* and *expectation*: (a) In the projection step, the given data are projected onto the curve. (b) In the expectation step, T points of the curve are updated to satisfy the self-consistency.

Before closing this section, we remark the meaning of *self-consistency.* The proposed methods do not use an intrinsic optimization that involves complex computational algorithms. Instead, the principal curves are based on a concept of *self-consistency* (Efron, 1967; Hastie, 1984; Hastie and Stuetzle, 1989; Flury and Tarpey, 1996), which is a fundamental concept in statistics covering EM-algorithm (Dempster et al., 1977), K-means clustering, and self-organizing maps (Kohonen, 1990), as noted in Flury and Tarpey (1996). As for now, we deeply explain the definition and estimation algorithm of principal curves, based on self-consistency. Equation (2.1) means that the curve f goes through the "middle" of data. As one can see in the definition of (2.1), the expression of f contains a term for f, $\lambda_{`f'}(X) = \lambda$ in the right-hand side. This is the exact reason why it is called "self"-consistency. The *essence* of this principle is to estimate a fixed point of (2.1). Our algorithm iterates projection step and median step for a candidate curve f to satisfy equation (2.1). Specifically, for the i-th curve f^i, we obtain $f^{i+1} := \mathbb{E}[X \mid \lambda_{f^i}(X) = \lambda]$ and have f^{i+2} by plugging f^{i+1}. Then repeat the process for $i = 1, 2, \ldots$ until the change is below a specific threshold (e.g., 0.01). In summary, the proposed principal curves cannot be obtained in a way that minimizes a cost function. Instead, the curves are estimated based on principle of self-consistency. The estimation method based on self-consistency has two advantages: (1) computationally fast and (2) reproducible, compared to an optimization framework. These advantages facilitates the production of R package **spherepc** that will be concretely explained in Chapter 5.

2.2 Riemannian manifolds and centrality on manifold

Manifold is a second-countable and Hausdorff topological space which locally resembles a Euclidean space. A smooth manifold (e.g., the unit 2-sphere in Figure 2.1) is a manifold equipped with a differentiable structure (or called atlas). Riemannian manifold M is a smooth manifold equipped with smoothly varying inner product $<,>_p$ on tangent bundle TM $(= \bigsqcup_{p \in M} T_pM)$ (or called as Riemannian metric) where the Riemannian metric can *measures* the magnitudes and angle of two tangent vectors on a tangent space. A (minimal) *geodesic* is the shortest smooth curve joining two points on M. The distance between the points along the curve is termed as *geodesic distance*, $d(\cdot, \cdot)$, where the distance is different with Riemannian metric but relies on the choice of Riemannian metric on M. For more details, see Boothby (1986); Lee (2006).

Nextly, exponential and logarithm maps will be defined. For each $p \in M$, *exponential map* is a differentiable map from a neighborhood of p in T_pM to M. For a vector v in the neighborhood, the geodesic at p with direction v, $\gamma : [0, 1] \to M$, uniquely exists so that γ satisfies that $\gamma(0) = p$, $\gamma^{'}(0) = v$, and $\|\gamma^{'}(t)\| = \|v\|$ for any $t \in [0, 1]$. The exponential map at p is defined as

$$\exp_p(v) := \gamma(1) \in M. \tag{2.2}$$

If (M, d) is connected and complete as a metric space, then the geodesic continues as much as we want from the Hopf-Rinow theorem (e.g. Theorem 6.13. of Lee (2006)). In other words, the exponential map at p, $\exp_p : T_pM \to M$, is defined on the entire T_pM. For the simplest case, $M = \mathbb{R}^D$, since $T_p\mathbb{R}^D \simeq \mathbb{R}^D$ for any $p \in \mathbb{R}^D$, the exponential and logarithm maps are both identity. In particular, when $M = S^D := \left\{(x_1, x_2, \ldots x_{D+1}) \in \mathbb{R}^{D+1} \mid \sum_{i=1}^{D+1} x_i^2 = 1\right\}$, naturally embedded into the ambient space $\mathbb{R}^{D+1}$, the exponential map at $p = (0, 0, \ldots, 0, 1) \in \mathbb{R}^{D+1}$ can be written as

$$\exp_p(v) = (v_1 \frac{\sin\|v\|}{\|v\|}, v_2 \frac{\sin\|v\|}{\|v\|}, \ldots, v_D \frac{\sin\|v\|}{\|v\|}, \cos\|v\|),$$

for any $v \in T_pS^D \simeq \mathbb{R}^D$ with $\|v\| \leq \pi$ in which $\|\cdot\|$ denotes the standard norm in $\mathbb{R}^D$. *logmap* is the inverse map of exponential map. The logmap at p, $\log_p : S^D \to T_pS^D$, is written by

$$\log_p(w) = (w_1 \frac{\theta}{\sin\theta},\, w_2 \frac{\theta}{\sin\theta},\, \ldots,\, w_D \frac{\theta}{\sin\theta})$$

for any $w = (w_1, w_2, \ldots, w_{D+1}) \in S^D \setminus (0, 0, \ldots, 0, -1) \subset \mathbb{R}^{D+1}$, where $\theta = \arccos(w_{D+1})$. See Buss and Fillmore (2001) for details.

We now consider a probability distribution μ on a complete and connected Riemannian manifold M with its geodesic distance $d(\cdot, \cdot)$. Before explaining the notion of centrality on the manifold, a simple motivating example is given in Figure 2.1. A typical example for manifold is sphere and the unit 2-sphere, in particular, is considered, as $S^2 = \{(x, y, z) \in \mathbb{R}^3 \,|\, x^2 + y^2 + z^2 = 1\} \subset \mathbb{R}^3$. In the left panel of Figure 2.1, the Euclidean mean of data points (blue) in $\mathbb{R}^3$, $(1/3, 1/3, 1/3)$, is not lying on the unit 2-sphere. the conventional (Euclidean) mean is not available on manifolds. It stems from the fact that S^2 is *not* vector space, inherently. Due to the lack of linearity on S^D or generic manifold, the centrality for data on manifold

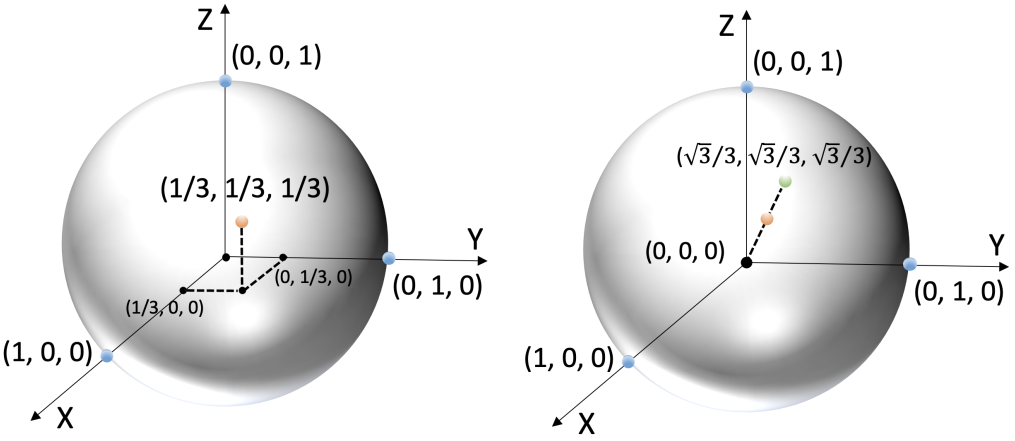


Figure 2.1: Left: The Euclidean mean (orange) of three points (blue) is not lying on the unit 2-sphere. Right: The extrinsic and intrinsic means (green) of the three points (blue).

should be defined.

There are several ways to defining the centrality on manifolds. Note that Euclidean space has the property that the arithmetic mean is the center of gravity, which minimizes the sum of squares of the distances from each point to itself. As an analog of this property, *intrinsic mean* (Fréchet mean or barycenter)(Fréchet, 1948) is

$$\operatorname*{arg\,min}_{m \in M} \int d^2(m, x)\mu(dx). \tag{2.3}$$

Generally, it is not easy to treat the geodesic distance. Thus, regarding the manifold M as an embedded subspace in Euclidean space $\mathbb{R}^D$ for some $D \geq 2$, the *extrinsic mean* (Huckemann et al., 2010; Bhattacharya and Patrangenaru, 2003, 2005) can be defined by replacing the geodesic distance in (2.3) with Euclidean distance in $\mathbb{R}^D$. Formally, the extrinsic mean with respect to ξ is defined as

$$\operatorname*{arg\,min}_{m \in M} \int \|\xi(m) - \xi(x)\|^2 \mu(dx), \tag{2.4}$$

where $\|\cdot\|$ is a Euclidean norm in $\mathbb{R}^D$ and ξ denotes an embedding from M to $\mathbb{R}^D$. In the real line, the median of data is a point that minimizes the sum of the distances from each data point to itself. The analog to multi-dimensional Euclidean space or manifold case is termed as *geometric median* or spatial median (Fletcher et al., 2009; Yang, 2010), which is defined as

$$\operatorname*{arg\,min}_{m \in M} \int d(m, x)\mu(dx). \tag{2.5}$$

The geometric median is an alternative measure for central tendency and is more robust to outliers than the barycenter from empirical and theoretic perspectives (Fletcher et al., 2009). Under a mild condition for μ, the geometric median uniquely exists (Yang, 2010). For a set of data $\{x_1, x_2, \ldots, x_n\} \subset M$, the sample version of the geometric median is

$$\operatorname*{arg\,min}_{m \in M} \sum_{i=1}^{n} w_i d(m, x_i), \tag{2.6}$$

where w_i denote nonnegative weights of x_i with $\sum_{i=1}^{n} w_i > 0$. It is known that geometric median has no closed-form (Fletcher et al., 2009). Thus, it should be obtained in an iterative way. For this purpose, we define a L_1-type loss function $g : M \to \mathbb{R}$ as $g(x) = \sum_{i=1}^n w_i d(x, x_i)$. By using the same arguments in Fletcher et al. (2009), the derivative of g can be obtained as

$$\nabla g(x) = -\sum_{i=1}^{n} w_i \log_x(x_i)/\|\log_x(x_i)\| \in T_x M,$$

for $x \notin \{x_i, x_2, \ldots, x_n\}$. Based on the gradient method, Fletcher et al. (2009) suggested an (Weiszfeld) algorithm for estimating the geometric median on Riemannian manifold as follows. The algorithm is similar to the method of iteratively reweighted least squares.

Algorithm 1: Geometric median on manifold

```
For a dataset $\{x_i\}_{i=1}^n \subset M$ and their nonnegative weights $\{w_i\}_{i=1}^n$ with
  $\sum_{i=1}^n w_i > 0$, set an initial value as $m_1 = x_1$. ;
while $(\Delta m \geq$ threshold$)$ do
    - $w_i' = w_i/\|\log_{m_k}(x_i)\|,\ 1 \leq i \leq n$ ;
    - $\Delta m = \sum_{1 \leq i \leq n,\, x_i \neq m_k} \frac{w_i'}{\sum_{i=1}^n w_i'} \log_{m_k}(x_i)$ ;
    - $m_{k+1} = \exp_{m_k}(\Delta m)$ ;
end
```

Before closing this section, we remark that if a dataset is not collinear and well-localized, e.g. on unit spheres $d(x, y) < \frac{\pi}{2}$ for any x, y of the dataset, the corresponding geometric median uniquely exists. Specifically, consider two cases: (i) Sectional curvatures of M are bounded above from $\kappa > 0$ and $\mathrm{diam}(M) := \sup\{d(x, y) \,|\, x,\ y \in M\} < \pi/(2\sqrt{\kappa})$. (ii) M has non-positive sectional curvatures. If M satisfies either (i) or (ii), then the weighted geometric median uniquely exists (Theorem 1 in Fletcher et al. (2009)). Algorithm 1 moreover converges to the geometric median. For more details, see Fletcher et al. (2009).

2.3 Principal curves on Riemannian manifolds

Hauberg (2016) proposed principal curves on Riemannian manifolds by expressing a curve as a set of T points, $f = \{C_1, C_2, \ldots, C_T\}$, joined by geodesics. The estimation algorithm of the curve follows that of Hastie and Stuetzle (1989) with an approximation. Specifically, the mean operation in the expectation step is performed by intrinsic mean, and the projection is conducted by finding the nearest point in f as

$$\mathrm{proj}(x) = \arg\min_{C_i \in f} d(x, C_i),$$

which is not an *exact* projection onto the continuous curve.

Chapter 3

Spherical principal curves

This chapter is based on Lee et al. (2020) (archive preprint) and Lee et al. (2021a) which has been published in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 2165-2171. The main contributions of this chapter can be summarized as follows: (a) We propose both extrinsic and intrinsic approaches to form principal curves on D-sphere, $D \geq 2$. (b) We verify the stationarity of the proposed principal curves on S^D. (c) We show the usefulness of the proposed method through real data analysis and simulation studies.

The main contributions of this chapter can be summarized as follows: (a) We propose both extrinsic and intrinsic approaches to form principal curves on D-sphere S^D, $D \geq 2$. (b) We verify the stationarity of the proposed principal curves on S^D. (c) We show the usefulness of the proposed method through real data analysis and simulation study. The detailed proofs of the theoretical properties of the proposed principal curves method are given in Section 3.4 and Appendix A.1.

The chapter is organized as follows. In Section 3.1, a newly developed exact principal circle on spheres is studied, which is used for the initialization of the proposed principal curves. Section 3.2 presents the proposed principal curves with a practical algorithm and investigates the stationarity of them theoretically. In Section 3.3, the experimental results of the proposed method are provided through real earthquake data from the U.S. Geological Survey, real motion capture data, and simulation studies on S^2 and S^4. Section 3.4 discusses justifications of *exact* projection step and theoretical properties of the proposed principal curves. Finally, concluding remarks are given in Section 3.5.

3.1 Enhancement of principal circle for initialization

Methods for fitting circles to data on S^2 are actively used in many applications, especially in astronomy and geology, to recognize undisclosed patterns of data (Mardia and Gadsden, 1977; Gray et al., 1980). This section improves the principal circle to be used as an initialization of the principal curves proposed in Section 3.2.

3.1.1 Principal geodesic and principal circle

The principal curve algorithm of Hastie and Stuetzle (1989) uses the first principal component as the initial curve, which is easily calculated by singular value decomposition (SVD) of the data matrix in Euclidean space. Along with this line, the proposed principal curve algorithm in Section 3.2 requires an initial curve. The principal geodesic analysis (PGA) by Fletcher et al. (2004) can be considered as a generalization of PCA that performs dimension reduction of data on the Cartesian product of simple manifolds, such as $\mathbb{R}^3$, S^2, and $\mathbb{R}_+$. To this end, Fletcher et al. (2004) projected each manifold component of the data into a tangent space at the intrinsic mean of each component. As a result of the tangent space approximation of each component, data are approximated by points in Euclidean space, so applying PCA allows dimension reduction to be performed through the inverse process of the tangent projection, i.e. *exponential map* that preserves a distance and angle at a base point. For spherical cases, they mainly perform tangent space projection using an inverse exponential map, called *log map*. The explicit forms of exponential and log maps of S^2 are described in Fletcher et al. (2004); Jung et al. (2011) and Jung et al. (2012).

However, PGA always results in a great circle going through the intrinsic mean on the sphere, as shown in Figure 3.1, and the class of great circles on a sphere is

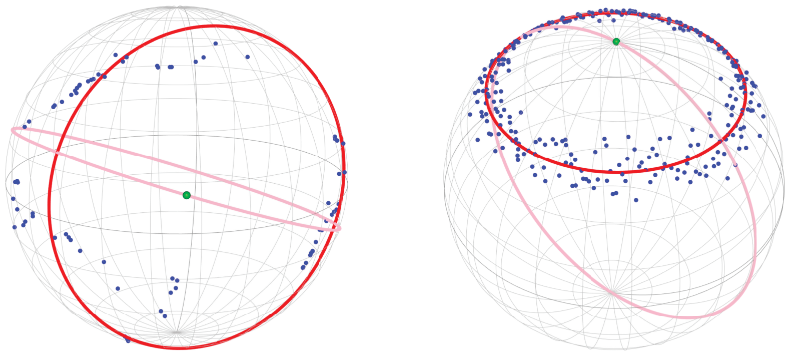


Figure 3.1: Left: Spherical distribution of significant earthquakes (blue) with its intrinsic mean (green), the result (pink) by PGA, and the result (red) by our proposed principal circle. Right: Circular simulated data (blue) with its intrinsic mean (green), the result (pink) by PGA, and the result (red) by our proposed principal circle.

sometimes limited to suitably fit a dataset on the sphere (Jung et al., 2011; Hauberg, 2016). For example, the left panel of Figure 3.1 shows earthquake data from the U.S. Geological Survey showing the location (blue dot) of significant earthquakes with Mb magnitude 8 or higher around the Pacific since 1900. The data will be analyzed in detail in Section 3.3. In Figure 3.1, while the result (pink) by PGA does not fit the data correctly, our principal circle (red), presented later in Section 3.1.2, improves the representation of the data. Further, in the right panel of Figure 3.1, our principal circle suitably fits the circular simulated data, whereas the result (pink) by PGA does not capture the variation of the data. The failure of PGA stems from the fact that the above two data sets are far from their intrinsic means, as noted in Jung et al. (2011), Jung et al. (2012), and Hauberg (2016).

In the literature, there is an attempt by Jung et al. (2011) that generalizes the PGA to a circle on S^2. The circle on S^2 that minimizes a reconstruction error is called *principal circle*, where the reconstruction error is defined as the total sum of squares of the geodesic distance between the curve and the data. Jung et al. (2011) used a double iteration algorithm that uses the log map to project the data into the tangent space and then finds the principal circle. However, this approach has two problems. First, using the tangent approximation when minimizing the distance may causes numerical errors. If the data points are located away from the mean, the numerical errors may increase because there is no local isometry between the sphere and its tangent plane according to the *Gauss's Theorema Egregium* (see p. 363-370 of Boothby (1986)). Second, due to the topological difference between the sphere and the plane, the existence of principal circles in the tangent plane is not guaranteed. For example, Figure 3.2 shows simulated data, where the underlying structure is a great circle, and the intrinsic mean is the North Pole (0, 0, 1), where the data points are mostly concentrated around the North Pole. From the compactness of the sphere, the least-squares circle always exists regardless of the data structure. It is an advantage of the intrinsic approaches. On the other hand, the least-squares circle does not exist if the data points projected onto the tangent space at their intrinsic mean are collinear, as shown in the middle and right panels of Figure 3.2. It coincides that several circle fitting procedures in a plane, such as Kåsa (1976) and Coope (1993), fail when the data points are collinear, as noted in Umbach and Jones (2003). Moreover in this case, the (tangent) plane cannot consider the periodicity of the data, as opposed to the left panel of Figure 3.2. Ignoring the periodic structure of data, as noted in Eltzner et al. (2018), may reduces the efficiency of a method. This study proposes a new principal circle that does not rely on tangent projection for the effective initialization of the proposed principal curve presented in Section 3.2. We obtain the constraint-free optimization problem by expressing the center of the circle using the spherical coordinate system in Section 3.1.2 and Section 3.1.3.

3.1.2 Exact principal circle

For our principal circle, we consider an intrinsic optimization algorithm that does not use any approximations. Let $d(x, y)$ be the geodesic distance between $x, y \in S^2$. For a given dataset $\mathcal{D}$ and a circle C on S^2, let $\delta(\mathcal{D}, C)$ be the sum of squares of

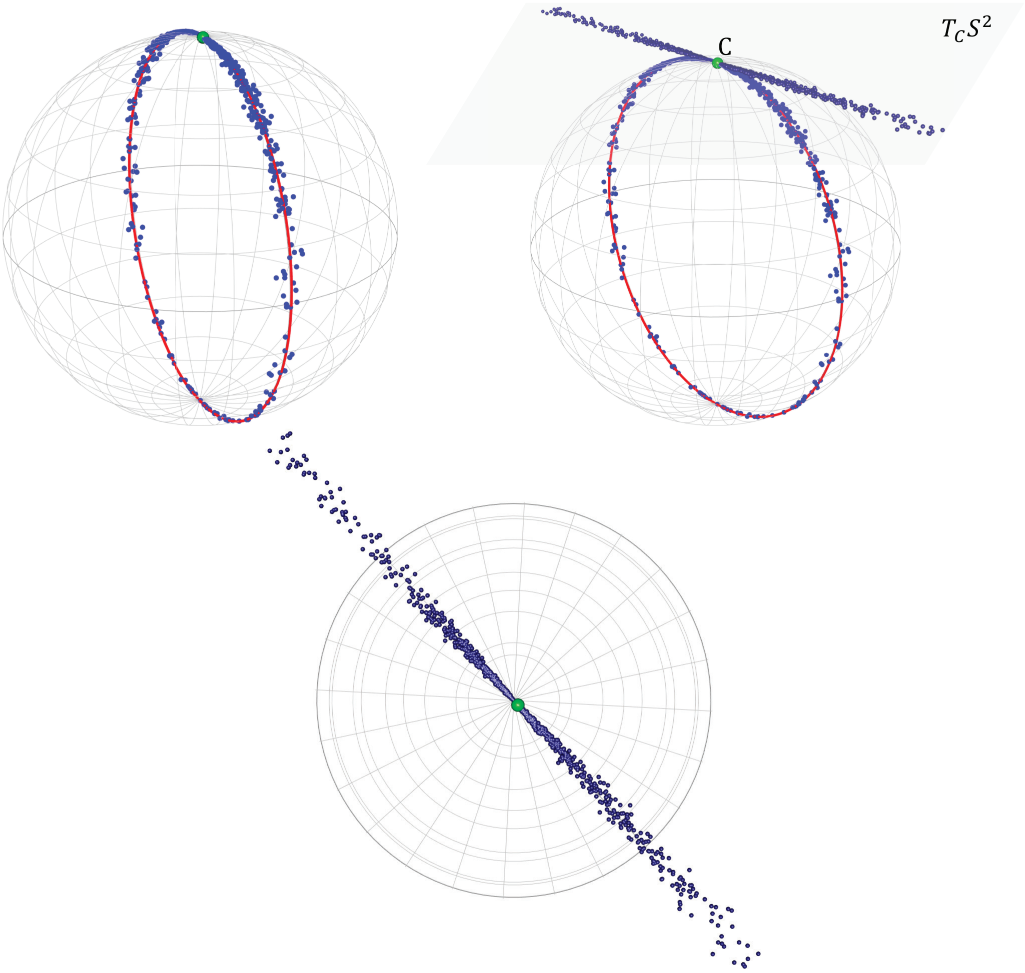


Figure 3.2: Top Left: Simulated data points (blue) with the intrinsic mean (0, 0, 1) (green) and the result of the proposed principal circle (red); Top Right: The projected points from the sphere onto the tangent plane at $C = (0, 0, 1)$; Bottom: The projected points viewed from above the Northern Hemisphere.

distances between circle and data, defined as

$$\delta(\mathcal{D},\, C) = \sum_{x \in \mathcal{D}} d\big(x,\, \mathrm{proj}_C(x)\big)^2,$$

where $\mathrm{proj}_C(x)$ denotes a projection of x on C. The goal is to find a circle C on S^2 that minimizes $\delta(\mathcal{D},\, C)$. To solve this optimization problem, we represent a circle C by a center c of the circle and a radius $r \in [0,\, \pi]$, the geodesic distance between the center c and the circle C. This representation is not unique Jung et al. (2011). For example, let $c' \in S^2$ be the antipodal point of c that is diametrically opposite to c on S^2, then $(c,\, r)$ and $(c',\, \pi - r)$ represent the same circle C. Nevertheless, it is not crucial to the optimization problem because we simply find a representation of the least square circle. By using a spherical coordinate system, it is able to parameterize c as $(\theta,\, \rho)$, where θ denotes the azimuthal angle and ρ is the polar angle. By symmetry of the circle, $d\big(x,\, \mathrm{proj}_C(x)\big)$ can be easily calculated by

$$d\big(x,\, \mathrm{proj}_C(x)\big) = d(x,\, c) - r.$$

Thus, we have

$$\delta(\mathcal{D},\, C) = \sum_{x \in \mathcal{D}} \big(d(x,\, c) - r\big)^2. \tag{3.1}$$

With letting $c = (\theta_c,\, \rho_c)$ and $x = (\theta_x,\, \rho_x)$ in the spherical coordinate system, the geodesic distance $d(x,\, c)$ is given by the spherical law of cosines with three points c, x, and the polar point (see Lemma 3 in Section 3.4. below for details)

$$d(x,\, c) = \arccos\big(\cos\rho_c \cos\rho_x + \sin\rho_c \sin\rho_x \cos(\theta_c - \theta_x)\big). \tag{3.2}$$

By putting (3.2) into (3.1), it follows that $\delta(\mathcal{D},\, C)$ is represented as a three-parameter differentiable function $\delta_{\mathcal{D}}(\theta_c,\, \rho_c,\, r)$ in domain $[0,\, 2\pi] \times [0,\, \pi] \times [0,\, \pi]$ as follows,

$$\delta_{\mathcal{D}}(\theta_c,\, \rho_c,\, r) = \sum_{x \in \mathcal{D}} \big(\arccos\big(\cos\rho_c \cos\rho_x + \sin\rho_c \cdot \sin\rho_x \cos(\theta_c - \theta_x)\big) - r\big)^2. \tag{3.3}$$

Since $[0,\, 2\pi] \times [0,\, \pi] \times [0,\, \pi]$ is compact, the function $\delta_{\mathcal{D}}(\theta_c,\, \rho_c,\, r)$ holds a global minimum value. Thus, it can apply the gradient descent method to find the solution. Here is the algorithm to find a principal circle from the above description.

Algorithm 2: Exact principal circle by gradient descent

```
Initialize $(\theta_c,\ \rho_c,\ r)$ as $(\overline{\theta},\ \overline{\rho},\ \pi/2)$. ;
**while** *($\Delta\delta(\mathcal{D},\ C) \geq$ threshold)* **do**
    - $(\theta_c,\ \rho_c,\ r) \leftarrow (\theta_c, \rho_c,\ r) - \beta\nabla\delta_{\mathcal{D}}(\theta_c,\ \rho_c,\ r)$ ;
**end**
```

As in many nonlinear least-square algorithms, such as Gauss-Newton algorithm and Levenberg-Marquardt algorithm (see Chapter 4 of Scales (1985) for details), the above Algorithm 2 may converge to a local minimum or a saddle point instead of the global minimum, since $\delta_{\mathcal{D}}(\theta_c,\ \rho_c,\ r)$ is non-convex. Thus, initial values should be selected carefully. If the data points in $\mathcal{D}$ are not too apart and localized, then it is reasonable to choose $(\theta_x,\ \rho_x,\ \pi/2)$ for some $x \in \mathcal{D}$ as an initial. The spherical coordinates of the intrinsic mean of $\mathcal{D}$ with radius $r = \pi/2$, denoted by $(\overline{\theta},\ \overline{\rho},\ \pi/2)$, if necessary with varying $r \in [0,\ \pi]$, is also recommended as initial values. In the case of a non-localized data set, one can implement the algorithm with various initial settings as much as one wants, compare the consequences of δ, and finally choose the circle with the lowest δ as the principal circle. Note that, in existing methods for fitting circles to data on spheres, such as Gray et al. (1980); Jung et al. (2011, 2012), there are no assurances that their algorithms finally achieve the circle minimizing (3.1). Although δ is not convex globally, it is convex on a neighborhood of a global minimum point. Hence, it is reasonably expected that if an initial value is suitably close to an optimum point, then Algorithm 1 converges to the optimum. A specification about the neighborhood for which δ is convex, and rigorous proof for convergence of Algorithm 2 on that neighborhood remains a challenge. In the real data analysis and the simulated studies later in Section 2.5, however, implementations of Algorithm 1 with several initial values result in almost the same principal circles and converge rapidly. Thus, there are no practical difficulties in our experiments. In addition, β is the step size of Algorithm 2, and it relies on the dataset $\mathcal{D}$. The algorithm may diverge when β is large (e.g., greater than 0.01). In simulated examples and real data on Section 2.5., we use 0.001. Since too small β causes computational time to be high, an appropriate β should be selected properly throughout experiments from a relatively larger value of β to the lower one.

3.1.3 Extension to hyperspheres

In the case of high-dimensional spheres, to find a one-dimensional circle that attempts to represent a given data closely, we provide both extrinsic and intrinsic ways. The former is easy to implement and more computationally feasible because it uses an extrinsic approach and is not exactly found. The latter directly extends the exact principal circle in the previous section into higher-dimensional spheres using the framework of principal nested spheres (Jung et al., 2012); however, it takes time to compute compared to the former approach.

Circle as an initialization

Later in Section 3.3.2, we will use the following extrinsic method as an initial estimate of the spherical principal curves for waveform simulated data on S^4. Specifically, we consider $S^D = \{y = (y_1, y_2, \ldots, y_{D+1}) \in \mathbb{R}^{D+1} \mid \sum_{i=1}^{D+1} y_i^2 = 1\}$ for $D \geq 2$, as an embedded surface in the ambient space $\mathbb{R}^{D+1}$. That is, $\{x_i\}_{i=1}^n \subset S^D \hookrightarrow \mathbb{R}^{D+1}$ are regarded as elements in $\mathbb{R}^{D+1}$, not taking into account a nonlinear dependence of the data; though, ensuring lower computational complexity. Note that any one-dimensional circle on S^D is an intersection of a two-dimensional plane and S^D. Hence, the strategy is to find the 2-plane $P \subset \mathbb{R}^{D+1}$ that closely represents the data $\{x_i\}_{i=1}^n$ with respect to the standard distance in $\mathbb{R}^{D+1}$, rather than geodesic distance in S^D. That is, the plane P is the two-dimensional vector subspace of $\mathbb{R}^{D+1}$ spanned by first two principal components of the data, and then $P \cap S^D$ is a one-dimensional circle to find. Although the extrinsic circle is capable of approximating the meaningful data, there may be some instances that need more precise initial estimate for the data.

Exact principal circle

For a better initial guess of the proposed principal curves, we provide an exact principal circle on $S^D = \big\{y = (y_1,\, y_2,\, \ldots,\, y_{D+1}) \in \mathbb{R}^{D+1} \mid \sum_{i=1}^{D+1} y_i^2 = 1\big\}$ for $D \geq 3$. The arguments in Section 3.1.2 can be applied to higher-dimensional spheres S^D for $D \geq 3$ if the geodesic distance of (3.1) can be precisely calculated. To this end, let $\mathcal{D} = \{x_i\}_{i=1}^n$ be a dataset on S^D, and denote a $(D-1)$-dimensional subsphere on S^D as C. Using a spherical coordinates for S^D, $x = (x_1,\, x_2,\, \ldots,\, x_D,\, x_{D+1}) \in S^D \subset \mathbb{R}^{D+1}$ can be parametrized as

$$\begin{aligned}
x_1 &= \cos(\varphi_1)\\
x_2 &= \sin(\varphi_1)\cos(\varphi_2),\\
x_3 &= \sin(\varphi_1)\sin(\varphi_2)\cos(\varphi_3)\\
&\vdots\\
x_D &= \sin(\varphi_1)\cdots\sin(\varphi_{D-1})\cdot\cos(\varphi_D)\\
x_{D+1} &= \sin(\varphi_1)\cdots\sin(\varphi_{D-1})\cdot\sin(\varphi_D),
\end{aligned}$$

where $\varphi_1,\, \varphi_2,\, \cdots,\varphi_{D-1},\, \varphi_D$ are angular coordinates with $\varphi_D \in [0,\, 2\pi)$ and the others ranging over $[0,\, \pi)$. Note that $d(x,\, c) = \arccos(x \cdot c)$, where $\cdot$ denotes the (standard) inner product in $\mathbb{R}^{D+1}$. Thus,

$$\begin{aligned}
d(x,\, c) = \arccos\big(\cos(\varphi_{1c})\cos(\varphi_{1x}) + \sum_{k=1}^{D-2}\big[\prod_{i=1}^{k}\sin(\varphi_{ic})\sin(\varphi_{ix})\big]\cdot\cos(\varphi_{(k+1)c})\cos(\varphi_{(k+1)x})\\
+\big[\prod_{i=1}^{D-1}\sin(\varphi_{Dc})\sin(\varphi_{Dx})\big]\cdot\cos(\varphi_{Dc}-\varphi_{Dx})\big),
\end{aligned} \tag{3.4}$$

where $\{\varphi_{ic}\}_{i=1}^D$ and $\{\varphi_{ix}\}_{i=1}^D$ are the corresponding angular coordinates of c and x, respectively. By putting (3.4) into (3.1), $\delta(\mathcal{D},\, C)$ is represented as the $(n+1)$-parameter differentiable function $\delta_D(\varphi_{1c},\, \ldots,\, \varphi_{Dc},\, r)$ in domain $[0,\, \pi]^{D-1}\times[0,\, 2\pi]\times$

$[0,\ \pi]$ as follows:

$$\begin{aligned}\delta_{\mathcal{D}}(\varphi_{1c},\ \ldots,\ \varphi_{Dc},\ r) = \sum_{x \in \mathcal{D}} \Bigg(& \arccos\Big(\cos(\varphi_{1c})\cos(\varphi_{1x}) \\ & + \sum_{k=1}^{D-2} \big[\prod_{i=1}^{k} \sin(\varphi_{ic})\sin(\varphi_{ix}) \big] \cdot \cos(\varphi_{(k+1)c})\cos(\varphi_{(k+1)x}) \\ & + \big[\prod_{i=1}^{D-1} \sin(\varphi_{ic})\sin(\varphi_{ix}) \big] \cdot \cos(\varphi_{Dc} - \varphi_{Dx}) \Big) - r \Bigg)^2 .\end{aligned} \tag{3.5}$$

Note that, in the case of $D = 2$, the above equation (3.5) becomes (3.3). $\delta_{\mathcal{D}}$ holds a global minimum value due to the compactness of the domain $[0,\ \pi]^{D-1} \times [0,\ 2\pi] \times [0,\ \pi]$. Therefore, an exact principal circle on S^D can be obtained by gradient descent, the same way in Algorithm 2, except that the number of parameters is $D+1$. Let $(\overline{\varphi_1},\ \overline{\varphi_2},\ \ldots,\ \overline{\varphi_D})$ denote the spherical coordinates of the intrinsic mean of $\mathcal{D}$. Here is the algorithm to find a principal circle on S^D.

```
Algorithm 3: Exact principal nested sphere on hypersphere $S^D$
Initialize $(\varphi_{1c}, \varphi_{2c}, \ldots, \varphi_{Dc}, r)$ as $(\overline{\varphi_1}, \overline{\varphi_2}, \ldots, \overline{\varphi_D}, \pi/2)$.;
while ($\Delta\delta(\mathcal{D}, C) \geq$ threshold) do
    $(\varphi_{1c}, \varphi_{2c}, \ldots, \varphi_{Dc}, r) \leftarrow$
      $(\varphi_{1c}, \varphi_{2c}, \ldots, \varphi_{Dc}, r) - \beta\nabla\delta_{\mathcal{D}}(\varphi_{1c}, \varphi_{2c}, \ldots, \varphi_{Dc}, r)$ ;
end
```

It is possible that Algorithm 3 converges to a local minimum or a saddle point of $\delta_{\mathcal{D}}$, owing to its non-convexity. Therefore, an initial value should be carefully chosen, for instance, a data point in $\mathcal{D}$ and the intrinsic mean of $\mathcal{D}$. The discussions about initial values and step size β are the same as those of Algorithm 2.

By applying Algorithm 3 to a given data iteratively, we can obtain a one-dimensional sphere, i.e., an exact principal circle on S^D that can be the initialization of the spherical principal curves. For more details about the procedure, see Jung et al. (2012). It is noteworthy that from the perspective of the principal nested spheres, our method can be applied to find nested spheres in an *exact* way.

3.2 Proposed principal curves

This section presents our new *exact* principal curves on D-sphere for $D \geq 2$ from both intrinsic and extrinsic perspectives. We further investigate the stationarity of the proposed principal curves.

3.2.1 Exact projection step on S^D

As mentioned in Section 2.3, the approach of Hauberg (2016) does not perform the exact projections onto curves. On the other hand, the exact projections on S^D are carried out in our method, which results in more elaborated principal curves. To this end, we parametrize the curve as a set of T points joined by geodesics as in Hauberg (2016). Specifically, we first project the data point to each geodesic segment of the curve and then obtain the exact projection on the curve by choosing the closest geodesic segment. Let $\lambda_f(x)$ be the projection index of a point x to the curve $f(\lambda)$ for $\lambda \in [0,\, 1]$,

$$\lambda_f(x) = \min_{\lambda \in [0,\, 1]} \{\lambda \,|\, d(x,\, f(\lambda)) = \min_{\gamma \in [0,\, 1]} d(x,\, f(\gamma))\}, \tag{3.6}$$

The projection of x onto the curve can be obtained as $f\big(\lambda_f(x)\big)$.

The following subsections describe a procedure for projecting a point onto a geodesic segment on S^D. Given $A,\ B,\ C \in S^D \subset \mathbb{R}^{D+1}$, we find the closest point to C on the geodesic segment joining A and B. When $A = B$, the process is obvious, and in the case of $A = -B \in \mathbb{R}^{D+1}$, there is no unique geodesic connecting A and B. Hence, we only consider the case that A and B are linearly independent, i.e., $(A \cdot B)^2 \neq 1$, where $\cdot$ denotes the dot product in $\mathbb{R}^{D+1}$. We first deal with the projection on S^2 and then extend it into hyperspherical cases.

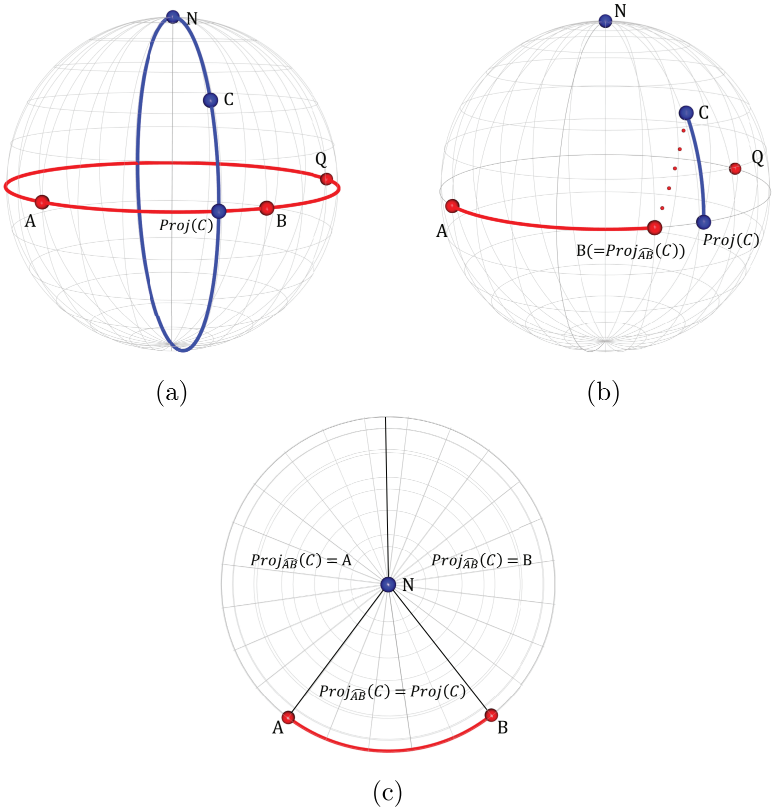

N
C
Q
A
$Proj(C)$
B
(a)
N
C
Q
A
B($=Proj_{\widehat{AB}}(C)$)
$Proj(C)$
(b)
$Proj_{\widehat{AB}}(C) = A$
$Proj_{\widehat{AB}}(C) = B$
N
$Proj_{\widehat{AB}}(C) = Proj(C)$
A
B
(c)

Figure 3.3: Illustration of the projection procedure on S^2: (a) The case that C is projected inside $\widehat{AB}$, i.e., $\text{proj}_{\widehat{AB}}(C) = \text{proj}(C)$ and $I \geq 0$. The projection of C is an intersection point of two great circles. (b) The case that C is projected onto B in a non-orthogonal way (red dotted line), i.e., $\text{proj}(C) \neq \text{proj}_{\widehat{AB}}(C) = B$ and $I < 0$. (c) An image of the sphere viewed from above the Northern Hemisphere in the projection of C.

Projection on S^2

Before describing the projection procedure on S^2, it is important to notice that $(A \cdot B)^2 \neq 1$ is equivalent to $A \times B \neq 0$, where $\times$ denotes the cross product in $\mathbb{R}^3$. In addition, if $A \times B / \|A \times B\| = \pm C$, then any points on geodesic through A and B have the same distance from C. From now on, we assume $A \times B / \|A \times B\| \neq \pm C$.

Figure 3.3 shows the projection procedure. We define the North Pole N concerning A and B as $N = \frac{A \times B}{\|A \times B\|} \in S^2$ and a center Q of the great circle through N and C as $Q = \frac{N \times C}{\|N \times C\|} \in S^2$ that is contained in the great circle through A and B. Then, the projection of C onto the great circle through A and B, $\text{proj}(C)$, becomes an intersection point of two great circles, as shown in Figure 3.3a,

$$\text{proj}(C) = Q \times N = \frac{(A \times B) \times C}{\|(A \times B) \times C\|} \times \frac{(A \times B)}{\|A \times B\|} \in S^2.$$

Note that $\text{proj}(C)$ is not always included in the geodesic segment $\widehat{AB}$ joining A and B as Figure 3.3b. For this reason, we define an indicator $I = -\big(A - \text{proj}(C)\big) \cdot \big(B - \text{proj}(C)\big)$, indicating whether $\text{proj}(C)$ is inside $\widehat{AB}$ or not, i.e., orthogonally projected onto $\widehat{AB}$ or not. Finally, the projection of C onto $\widehat{AB}$, $\text{proj}_{\widehat{AB}}(C)$, is

$$\text{proj}_{\widehat{AB}}(C) = \begin{cases} \text{proj}(C), & \text{if } I \geq 0 \\ \arg\min_{E \in \{A,\, B\}} d(C,\, E), & \text{if } I < 0. \end{cases}$$

Projection on hypersphere

For $A,\, B,\, C \in S^D \subset \mathbb{R}^{D+1}$, if $B \cdot C = C \cdot A = 0$, then all points on $\widehat{AB}$ have the same geodesic distance of $\pi/2$ from C, which is verified in Section 3.4.1; hence, assume that A, B, and C do not satisfy $B \cdot C = C \cdot A = 0$. Let V be a two-dimensional vector space in $\mathbb{R}^{D+1}$ spanned by A and B.

As shown in Figure 3.4, we aim to find the projection of C onto $V \cap S^D$, $\text{proj}(C)$, by following two steps: (Step 1) Locate the projection of C onto V, C'. (Step 2) Find the projection of C' onto $V \cap S^D$. Note that the resulting projection is equivalent to the projection of C onto $V \cap S^D$, $\text{proj}(C)$. The rigorous justification of the above procedure is provided in Section 3.4.1.

(Step 1): We find the closest point $C' \in V$ from C. Let $C' = \mu A + \lambda B$ for $\mu,\, \lambda \in \mathbb{R}$. Then C' should satisfy the orthogonal condition, $(C - C') \cdot A = (C - C') \cdot B = 0$. By plugging the equation $C' = \mu A + \lambda B$ into the above condition and solving the systems of linear equations with respect to μ and λ, it follows that

$$C' = \frac{C \cdot A - (A \cdot B)(B \cdot C)}{1 - (A \cdot B)^2} A + \frac{B \cdot C - (A \cdot B)(C \cdot A)}{1 - (A \cdot B)^2} B,$$

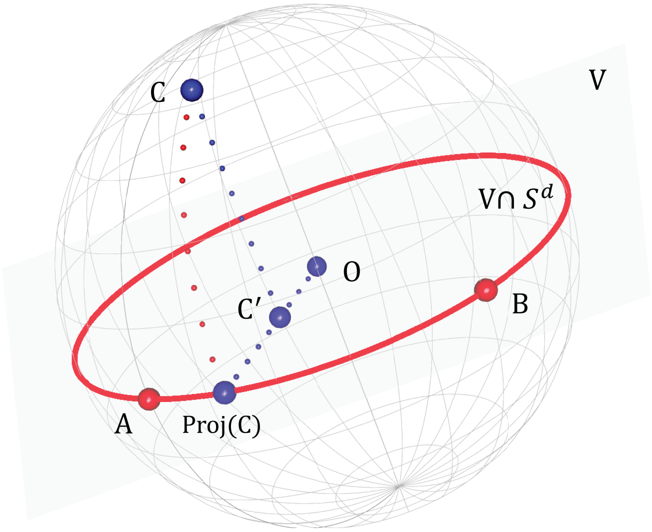


Figure 3.4: Illustration of the projection procedure on S^D

where the denominator is non-zero and $C' \neq 0 \in \mathbb{R}^{D+1}$ because of the assumptions; $(A \cdot B)^2 \neq 1$, and A, B, and C do not satisfy $B \cdot C = C \cdot A = 0$.

(Step 2): The projection of C' onto $V \cap S^D$, $\mathrm{proj}(C)$, is obtained by just normalizing C' so that it is in S^D. Therefore, we have

$$\mathrm{proj}(C) = \frac{C'}{\|C'\|} = \frac{\big(C \cdot A - (A \cdot B)(B \cdot C)\big)A + \big(B \cdot C - (A \cdot B)(C \cdot A)\big)B}{\Big\|\big(C \cdot A - (A \cdot B)(B \cdot C)\big)A + \big(B \cdot C - (A \cdot B)(C \cdot A)\big)B\Big\|}.$$

Similarly, we define the indicator $I = -\big(A - \mathrm{proj}(C)\big) \cdot \big(B - \mathrm{proj}(C)\big)$ to find the projection of C onto $\widehat{AB}$, $\mathrm{proj}_{\widehat{AB}}(C)$. Due to the fact that A, B, and $\mathrm{proj}(C)$ are in the one-dimensional unit circle $V \cap S^D$, we obtain $I \neq 0$ unless $\mathrm{proj}(C) = A$ or B. Since I is continuous with respect to $\mathrm{proj}(C) \in V \cap S^D$, it indicates that whether $\mathrm{proj}(C)$ is in $\widehat{AB}$ or not. We finally obtain $\mathrm{proj}_{\widehat{AB}}(C)$ as

$$\mathrm{proj}_{\widehat{AB}}(C) = \begin{cases} \mathrm{proj}(C), & \text{if } I \geq 0 \\ \arg\min_{E \in \{A,B\}} d(C,\, E), & \text{if } I < 0. \end{cases}$$

Note that the distance between C and $\widehat{AB}$ is the geodesic distance from C to $\mathrm{proj}_{\widehat{AB}}(C)$, which can be calculated as

$$d(C,\, \mathrm{proj}_{\widehat{AB}}(C)) = \arccos(C \cdot \mathrm{proj}_{\widehat{AB}}(C)). \tag{3.7}$$

3.2.2 Expectation step on S^D

The expectation step follows the principal curve of Hauberg (2016), i.e., updates the weighted average with smoothing that makes the curve closer to the self-consistency condition. Suppose that we have n data points $\mathcal{D} = \{x_i\}_{i=1}^n$ and the corresponding projection indices $\{\lambda_i\}_{i=1}^n$, where $\lambda_i = \lambda_f(x_i)$ for $i = 1, \ldots, n$. Let T denote the number of points of an initial curve. Then, the local weighted smoother iteratively updates the t^{th} point of the principal curve, C_t, with the weighted mean of data points. In this study, we use a quadratic kernel $k(\lambda) = (1 - \lambda^2)^2 \cdot \delta_{|\lambda| \leq 1}$, as Hauberg (2016), and the weight of each data point is given by $w_{t,\,i} = k(|\lambda_f(C_t) - \lambda_i|/\sigma)$ where $\sigma = q \cdot (\text{length of } f)$.

Extrinsic approach

The extrinsic mean on S^D can be calculated by considering the canonical embedding $S^D \hookrightarrow \mathbb{R}^{D+1}$. Specifically, for a curve $f = \{C_1, \ldots, C_T\}$ and each point C_t, the extrinsic mean is obtained by averaging the data points represented in Euclidean coordinates as

$$m_t(D,\, f) = \sum_{i=1}^n w_{t,\,i} x_i / \| \sum_{i=1}^n w_{t,i} x_i \|, \quad t = 1, 2, \ldots, T \tag{3.8}$$

where $\|\cdot\|$ is the standard norm in $\mathbb{R}^{D+1}$. Then C_t is updated by $m_t(\mathcal{D},\, f)$. The extrinsic approach is advantageous in terms of the computational complexity compared to the intrinsic approach. Furthermore, the extrinsic way ensures the stationarity of the principal curves on hyperspheres S^D for $D \geq 2$, which will be discussed in Section 3.2.4.

Intrinsic approach

From the intrinsic perspective, the weighted mean of data points can be obtained by the optimization

$$m_t(\mathcal{D},\, f) = \arg\min_x \sum_{i=1}^n w_{t,\,i} d^2(x,\, x_i), \quad t = 1, 2, \ldots, T, \tag{3.9}$$

and then each C_t is updated by $m_t(\mathcal{D},\, f)$. The intrinsic mean exists uniquely if the points are in an open hemisphere of S^D, i.e., $\exists p \in S^D$ s.t. $d(x_i,\, p) < \frac{\pi}{2}$ for $1 \leq i \leq n$ Buss and Fillmore (2001). Since the intrinsic mean cannot be obtained in a closed form, to solve (3.9), algorithms based on tangent space approximation, such as Buss and Fillmore (2001); Fletcher et al. (2004), can be used.

Before closing this section, as an alternative measure of the centrality of data, the geometric median can be considered to robustify the principal curves for a dataset that might contain outliers instead of the extrinsic or intrinsic mean. *Median*-based principal curves and their associated characteristics can be developed along with the same line of our procedure, which is the main topic in Chapter 4.

3.2.3 Algorithm

Initialization

For a better estimation of principal curves, we initialize a principal curve as an exact principal circle on D-sphere S^D. The detailed descriptions of the circle and its algorithm were previously provided in Section 3.1.

Spherical principal curves

The proposed spherical principal curves on S^D can be obtained by algorithm 4 below.

Note that $d^2\big(x_i,\, f\big(\lambda_f(x_i)\big)\big)$ is calculated by (3.2). As far as Euclidean space is concerned as embedding space, the extrinsic approach is advantageous for computational efficiency (Bhattacharya et al., 2012). However, if the data points are not contained within local regions at the expectation step, the intrinsic method may have better performances than the extrinsic one. Furthermore, the intrinsic approach can be attractive because of its inherent metric.

Algorithm 4: Spherical principal curves

```
Initialize curve $f = \{C_1, \ldots, C_T\}$. ;
Parameterize the curve as $f(\lambda)$ by some constant speed. Calculate $\lambda_f(x_i)$ in
  (3.6) for $i = 1, 2, \ldots, n$. ;
Calculate errors $\delta(\mathcal{D}, f) = \sum_{i=1}^{n} d^2\big(x_i, f\big(\lambda_f(x_i)\big)\big)$. ;
while ($\Delta\delta(\mathcal{D}, f) \geq$ threshold) do
    (Expectation) $C_t \leftarrow m_t(\mathcal{D}, f)$ for $t = 1, 2, \ldots, T$. ;
    Reparameterize the curve by some constant speed. ;
    (Projection) Calculate $\lambda_f(x_i)$ for $i = 1, 2, \ldots, n$. ;
    Calculate $\delta(\mathcal{D}, f) = \sum_{i=1}^{n} d^2\big(x_i, f\big(\lambda_f(x_i)\big)\big)$. ;
end
```

3.2.4 Stationarity of principal curves

For a random vector X in $\mathbb{R}^D$, $D \in \mathbb{N}$, the stationarity of the principal curve of X is given by Hastie and Stuetzle (1989) as

$$\left.\frac{\partial \mathbb{E}_X[d^2(X, f + \epsilon g)]}{\partial \epsilon}\right|_{\epsilon=0} = 0, \tag{3.10}$$

where f and g are smooth curves in $\mathbb{R}^D$ satisfying $\|g\| \leq 1$ and $\|g'\| \leq 1$, and $d(X, f)$ denotes the (Euclidean) distance from X to the curve f.

However, since spheres are not vector spaces such as $\mathbb{R}^D$, additions are not directly defined on spheres. Thus, it is necessary to redefine some concepts, such as addition and perturbation, in order to extend the properties of the principal curves in Euclidean space to spheres. To this end, we conversely consider $f + g$ instead of g. Specifically, let f and $f + g$ be smooth curves on D-sphere parameterized with $\lambda \in [0, 1]$. Then, we define $f + \epsilon g$ in a pointwise sense as follows.

Definition 1. *For $a, b \in S^D$ and $\epsilon \in [0, 1]$, div(a, b, ϵ) is a set of points on geodesics between a and b satisfying $\forall c \in$ div(a, b, ϵ), $d(a, c) = \epsilon d(a, b)$ and c is on a geodesic between a and b.*

Note that if $d(a, b) < \pi$, then the geodesic between a and b on S^D is unique. In this case, $div(a, b, \epsilon)$ is a single point set and $div(a, b, -\epsilon)$ can be defined as a reflection of $div(a, b, \epsilon)$ with respect to a.

Definition 2. *Let f and $f + g$ be smooth curves on S^D parameterized with $\lambda \in [0, 1]$ satisfying $\|g\| < \pi$, where $\|g\| := \max_{\lambda\in[0, 1]} d\big(f(\lambda), (f + g)(\lambda)\big)$. Then, for $\epsilon \in [-1, 1]$, $f + \epsilon g$ is a curve on S^D, where $(f + \epsilon g)(\lambda) = div(f(\lambda), (f + g)(\lambda), \epsilon)$, $\forall\lambda \in [0, 1]$.*

Note that $f + \epsilon g$ is a smooth curve on S^D. For a detailed proof, refer to the proposition 1 in Section 3.4.2. Let X be a S^D-valued random variable that doesn't have a point mass. We call f as an extrinsic-type principal curve of X or self-consistent if f satisfies

$$\pi\big(\mathbb{E}[\xi(X) \,|\, \lambda_f(X) = \lambda]\big) = f(\lambda) \text{ for a.e. } \lambda, \tag{3.11}$$

where $\xi : S^D \to \mathbb{R}^{D+1}$ is the canonical embedding and $\pi : \mathbb{R}^{D+1} \setminus \{0\} \to S^D$ by $X \to \frac{X}{\|X\|}$ is the standard projection (retraction) from $\mathbb{R}^{D+1}$ to S^D. In analogy to (3.10), we provide the following theorem on spheres. Note that $\cdot$ represents the standard inner product on $\mathbb{R}^{D+1}$ and $d(X, f + \epsilon g)$ denotes the geodesic distance from X to the curve $f + \epsilon g$. The definition of ϵ-perturbation, $f + \epsilon g$, coincides with that of Euclidean case because geodesics on Euclidean space are straight lines. $\|g'\|$ is also defined by mimicking the Euclidean case as follows.

Definition 3. $\|g'\| = \max_{\lambda\in[0, 1]} \|g'(\lambda)\|$, *where* $\|g'(\lambda)\| = \max_{\epsilon\in[0, 1]} \left\|\frac{\partial^2(f+\epsilon g)(\lambda)}{\partial\lambda\partial\epsilon}\right\|$.

The mild conditions for a f and a S^D-valued random variable X are assumed as follows:

(A1) $f : [0, 1] \to S^D$ is smooth (C^3), not self-intersecting (i.e. $\lambda_1 \neq \lambda_2 \in [0, 1) \Rightarrow f(\lambda_1) \neq f(\lambda_2)$), and parameterized by some constant speed, i.e. $|f'(\lambda)| = s > 0$ for any $\lambda \in [0, 1]$.

(A2) A S^D-valued random variable X has no point mass and is supported on $B(\zeta)$ where $B(\zeta) := \{x \in S^D \,|\, |f''(\lambda_f(x)) \cdot x| > \zeta\}$ for some constant $\zeta > 0$.

(A3) $\lambda_f(X) \in (0,\, 1)$ for a.e. X.

The reason for assumption (A1) is two-fold: One is for simplicity of discussion and notation compared to parameterization by *arc-length*. The other is that this setting covers only the class of finite-length curves on a sphere, i.e. $L(f) = \int_0^1 \sqrt{1 + f''^2(t)}dt < \infty$, where $L(f)$ denotes the length of f and f'' has a maximum on the compact interval $[0,\, 1]$ from (A1). Therefore, we can exclude the pathological examples such as space-filling curves and self-similar curves (e.g. fractal curves) on the sphere, which may have infinite length. The support condition of X in (A2) is mild since it can be shown that the area of $S^D \setminus B(\zeta) \to 0$ as $\zeta \to 0$ from Lemma 4 in Section 3.4. In other words, the condition is almost negligible by letting arbitrarily small $\zeta > 0$. The detailed explanations for the assumptions are given in Section 3.4.

Theorem 1. *Under* $(A1) - (A3)$*,* f *is an extrinsic principal curve of* X *if and only if*

$$\left.\frac{\partial \mathbb{E}_X[\cos\big(d(X,\, f + \epsilon g)\big)]}{\partial \epsilon}\right|_{\epsilon=0} = 0. \tag{3.12}$$

Proof. See Section 3.4.2. □

Note that (3.12) can be interpreted as an analogy of (3.10) because $2 - 2\cos x \approx x^2$ for small x. We further consider the intrinsic perspective of the stationarity. We define a curve f as an intrinsic-type principal curve of X if the intrinsic mean of X conditioned on $\lambda_f(X) = \lambda$ is equal to $f(\lambda)$ for a.e. λ,

$$\mathbb{E}_{int}[X \mid \lambda_f(X) = \lambda] = f(\lambda) \;\text{ for a.e. } \lambda, \tag{3.13}$$

where $\mathbb{E}_{int}[\cdot]$ denotes an intrinsic mean of a random variable on S^D.

Note that the intrinsic mean of a S^D-valued random variable Y is unique if $d(Y,\, p) < \frac{\pi}{2}$ a.s. for $\exists p \in S^D$, i.e., the support of Y is contained in an open hemisphere (Pennec et al., 2006). We verify that the intrinsic principal curves on S^2 satisfy the stationarity.

Theorem 2. *Under* $(A1) - (A3)$, f *is an intrinsic principal curve of* X *if and only if*

$$\frac{\partial \mathbb{E}_X[d^2(X,\, f + \epsilon g)]}{\partial \epsilon}\bigg|_{\epsilon=0} = 0. \tag{3.14}$$

Proof. See Section 3.4.2. □

The constraints $B(\zeta)$ in Theorems 1 and 2 are required to ensure the differentiation of the projection index $\lambda_{f+\epsilon g}(X)$ with respect to ϵ. Note that the constraints are almost negligible by letting ζ infinitesimally small; see Lemma 4 in Section 3.4.2 for details.

We finally remark that the stationarity of the principal curves in Euclidean space provides a rationale for the principal curves of Hastie and Stuetzle (1989) that is a nonlinear generalization of the linear principal component. Following the same line, the above stationarity results provide a theoretical justification that the proposed approaches directly generalize the principal curves of Hastie and Stuetzle (1989) from Euclidean space to spheres. In the intrinsic approach, the case of S^D with $D \geq 3$ remains a challenge.

3.3 Numerical experiments

This section conducts numerical experiments with real data analysis and simulated examples to assess the practical performance of the proposed methods. The experiments can be reproducible by R package, **spherepc** at `https://cran.r-project.org/package=spherepc` (Lee et al., 2022a), which implements the spherical principal curves for a variety of datasets lying on S^2. For more details, see Chapter 5 or Lee et al. (2022a).

3.3.1 Real data analysis

Earthquake data on S^2

We consider earthquake data from the U.S. Geological Survey (`https://earthquake.usgs.gov/earthquakes/map/`) in Figure 3.5 that represent the distribution of sig-

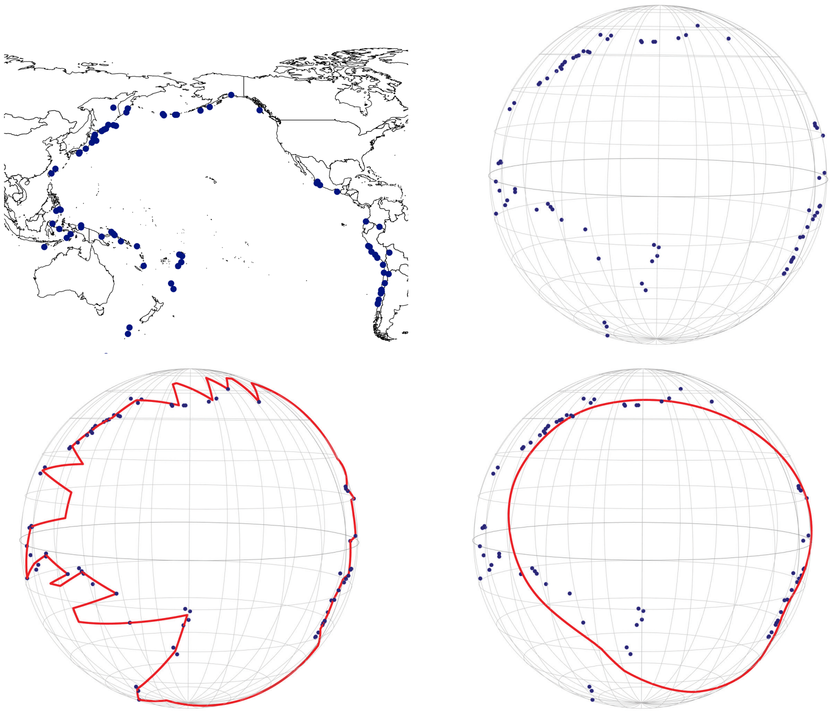


Figure 3.5: Top: Earthquake data is distributed in globally and they are visualized by two-dimensional and three-dimensional view, from left to right. Bottom: The proposed extrinsic principal curves of $T = 500$ with $q = 0.01$ (left) and $q = 0.2$ (right) are shown, respectively. Blue points represent the observations and red lines are the fitted curves.

nificant earthquakes (8+ Mb magnitude) around the Pacific Ocean since 1900. As shown in the figure, 77 observations are distributed in the vicinity of the borders between the Pacific, Eurasian, and Nazca plates. Since the plates are gradually moving towards different directions, recognizing the unrevealed patterns of borders provides essential information about seismological events such as earthquakes and volcanoes (Mardia and Gadsden, 1977; Biau and Fischer, 2011). In the following experiment, we utilize the spherical principal curves to recover the plates' borders by extracting curvilinear features of the observations.

We have implemented the proposed principal curves connected by $T = 500$, with various values of hyperparameter q that is the bandwidth of kernel in the expectation step. Figure 3.5 shows the results with $q = 0.01, 0.2$. We observe that a small q produces a wiggly and overfitted curve. It is noteworthy that the choice of q affects the quality of the fitted curve. Duchamp and Stuetzle (1996) proved that principal curves are always the saddle point of the expectation of the squared distance from a particular random variable, pointing out that cross-validation is not reliable for the model selection of principal curves, i.e., determination of q. Kégl et al. (2000) defined principal curves that minimize reconstruction errors in the constraint of the curve length, but used a heuristic way to determine the corresponding hyperparameter, the length of the curves. In the current study, the value of q is selected by visual inspection through all our experiments. An objective way to select q is left for future research.

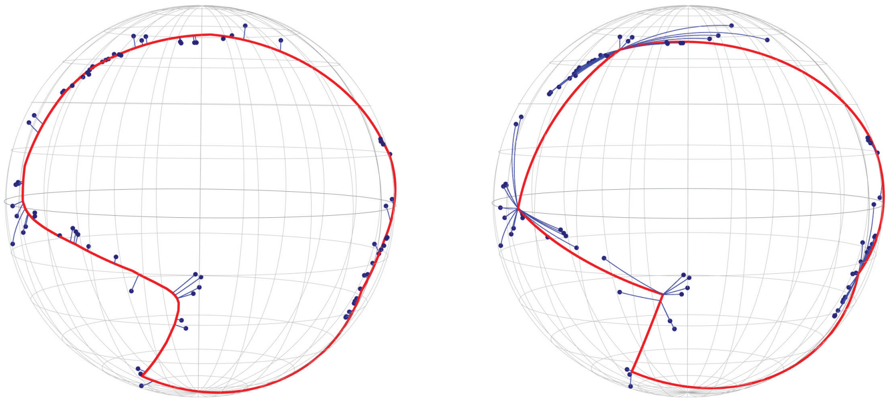


Figure 3.6: Projection results by the proposed extrinsic method (left) and Hauberg's method (right) with $T = 77$ and $q = 0.1$.

As one can see, the proposed extrinsic curve represents a given data as a continuous curve, while Hauberg's method projects several local data at one point.

We further compare the proposed extrinsic principal curves with the method of Hauberg (2016). Figure 3.6 shows both results with $q = 0.1$, where the purple lines represent the fitted curves, and the blue lines represent the projections from the data to the curve. The proposed extrinsic principal curve continuously represents the given data on the curve, while the method of Hauberg (2016) projects several local points to a single location. The comparison is further summarized in Table 1. As a result, the number of distinct projections (# proj) by our method is much larger than that of Hauberg's method. It implies that the proposed principal curve continuously represents the data, whereas the method of Hauberg tends to cluster the data. We also measure a reconstruction error (RE) defined as $\sum_{i=1}^{n} d^2\big(x_i, \hat{f}(\lambda_{\hat{f}}(x_i))\big)$ with observations $\{x_i\}_{i=1}^{n}$ and fitted values $\{\hat{f}(\lambda_{\hat{f}}(x_i))\}_{i=1}^{n}$. As listed in Table 1, our method outperforms Hauberg's method in terms of the reconstruction error.

Table 3.1: The values of RE and # proj by the proposed methods and Hauberg's method on the earthquake data

			Extrinsic	Intrinsic	Hauberg
$T = 77$	$q = 0.2$	RE	**2.662**	4.391	12.067
		# proj	**74/77**	72/77	22/77
	$q = 0.1$	RE	**0.463**	0.467	4.920
		# proj	**76/77**	**76/77**	9/77
	$q = 0.05$	RE	**0.359**	**0.359**	1.313
		# proj	**74/77**	73/77	16/77
	$q = 0.01$	RE	**0.061**	**0.061**	0.227
		# proj	**75/77**	**75/77**	27/77
$T = 500$	$q = 0.2$	RE	**2.193**	3.460	11.300
		# proj	**75/77**	72/77	30/77
	$q = 0.1$	RE	**0.715**	0.732	3.903
		# proj	**75/77**	74/77	18/77
	$q = 0.05$	RE	0.298	**0.200**	0.963
		# proj	**75/77**	**75/77**	27/77
	$q = 0.01$	RE	**0.036**	**0.036**	0.121
		# proj	**75/77**	**75/77**	37/77

Motion capture data on S^2

We now consider a benchmark data on S^2, motion capture data of a person walking in a circular pattern (Ionescu et al., 2011, 2014; Hauberg, 2016; Mallasto and Feragen, 2018). The data represent the orientation of the person's left *thigh bone* and naturally lie on S^2. There are 338 data points in the data set that are periodic.

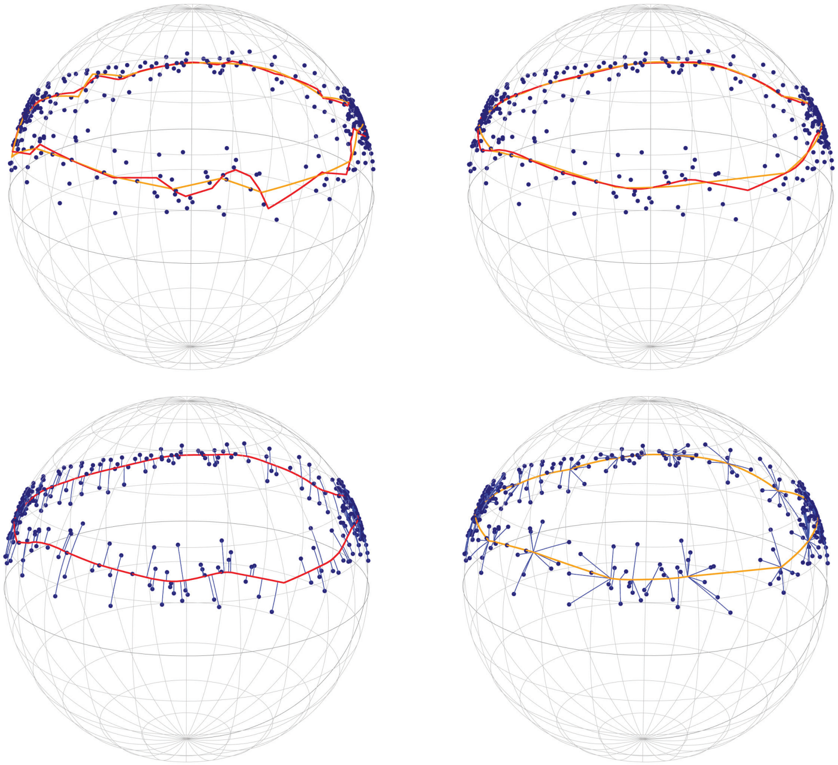


Figure 3.7: The results of the proposed extrinsic method (red) and Hauberg's method (yellow) with $T = 100$ are presented. The results with $q = 0.03$, $q = 0.05$ (top) and projection results (bottom) by the two methods with $q = 0.05$ are represented.

Figure 3.7 shows both results with $q = 0.03, 0.05$, where the red and yellow lines represent the fitted curves, and the blue lines represent the projections from the data to the curves. The proposed extrinsic principal curve continuously represents the given data on the curve, while the method of Hauberg (2016) projects several local points to a single location. Furthermore, Table 3.2 lists the quantitative results of the proposed methods and the method of Hauberg (2016). As listed, the proposed methods outperform Hauberg's method in terms of the reconstruction error and represent the data more precisely.

Table 3.2: The values of RE and # proj by the proposed methods and Hauberg's method on the motion capture data

			Extrinsic	Intrinsic	Hauberg
$T = 500$	$q = 0.05$	RE	**2.502**	2.504	2.534
		# proj	336/338	**337/338**	223/338
	$q = 0.03$	RE	**1.741**	**1.741**	2.637
		# proj	332/338	**333/338**	119/338
	$q = 0.01$	RE	**0.669**	**0.669**	1.253
		# proj	315/338	**317/338**	92/338

3.3.2 Simulation study

Simulation on S^2

We consider two types of functions on the unit sphere with spherical coordinates $(r = 1,\, \theta,\, \phi)$, where θ is the azimuthal angle and ϕ is the polar angle: (Circle) it is formed of $(r = 1,\, \theta,\, \phi)$ with $0 \leq \theta < 2\pi$ and $\phi = \pi/4$. (Wave) it is defined as $(r = 1,\, \theta,\, \phi)$ with $0 \leq \theta < 2\pi$ and $\phi = \alpha \sin(\theta f) + \pi/2$, where the frequency $f = 4$ and the amplitude $\alpha = 1/3$.

For each type of functions, we generate $n = 100$ data points by sampling θ uniformly in $[0,\, 2\pi)$ and adding Gaussian noises sampled from $N(0,\, \sigma^2)$ to ϕ. Figure 3.8 shows the results on the waveform data with $T = 500$ and $q = 0.05$. Both extrinsic and intrinsic principal curves extract the true waveform effectively, while Hauberg's approach yields a rather sharp curve. In Section 3.3.2, we provide additional visual

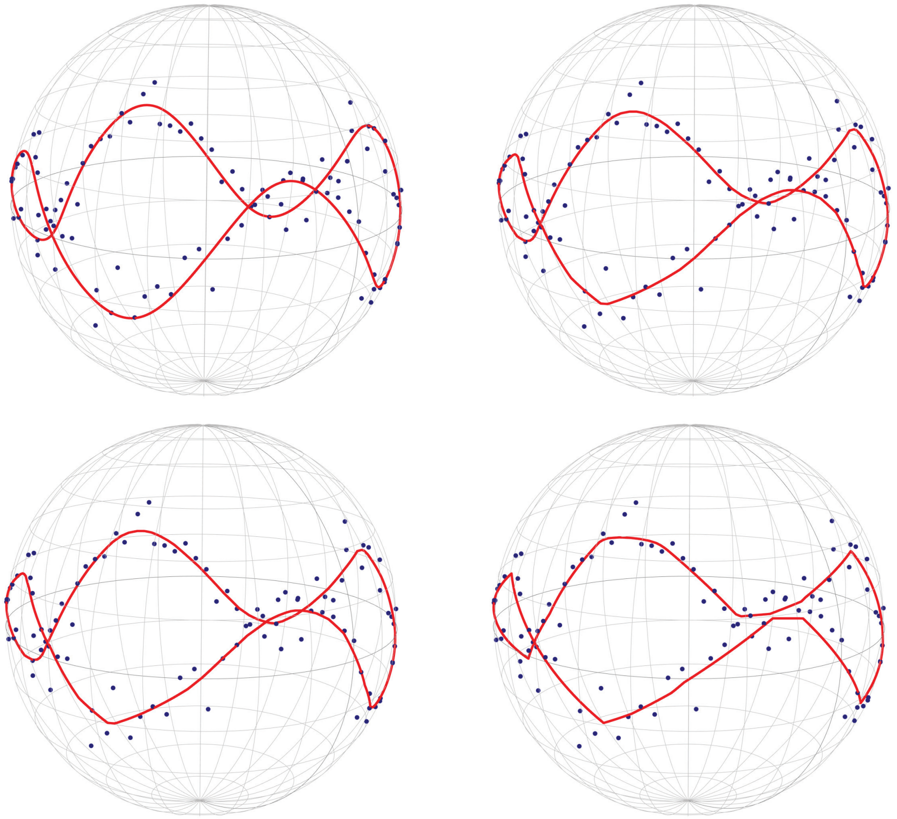


Figure 3.8: From top left to bottom right: True waveform and noisy data (blue dots), the extrinsic principal curve, the intrinsic principal curve, and the curve by Hauberg's method with $T = 100$.

results with various parameter settings.

We next quantify the performance of the proposed methods by measuring a reconstruction error between the fitted and true curves to measure the reconstruction ability of the methods. For the fitted curve $\hat{f}$, the reconstruction error is defined as $\sum_{i=1}^{n} d^2\big(x_i,\, \hat{f}\big(\lambda_{\hat{f}}(\tilde{x}_i)\big)\big)$, where $\{x_i\}_{i=1}^{n}$ denote the true values of the generating curves (ground-truth) and $\{\tilde{x}_i\}_{i=1}^{n}$ denote noisy sample values. We also count the number of distinct projection points to evaluate the continuity of resulting curves of the methods. Moreover, we compare the proposed spherical principal curves with existing method Hauberg (2016) over various settings $T = 100$, 500, $q = 0.05$, 0.03, 0.01,

Table 3.3: Averages of reconstruction errors and their standard deviations in the parentheses by each method

	True form	Method	Noise level					
			$\sigma = 0.07$			$\sigma = 0.1$		
			$q = 0.05$	$q = 0.03$	$q = 0.01$	$q = 0.05$	$q = 0.03$	$q = 0.01$
$T = 100$	Circle	Proposed	0.093 (0.026)	0.12 (0.027)	0.095 (0.013)	0.201 (0.048)	0.216 (0.046)	0.137 (0.025)
		Hauberg	0.117 (0.073)	0.408 (0.149)	0.298 (0.038)	0.370 (0.205)	0.74 (0.208)	0.494 (0.063)
	Wave	Proposed	0.71 (0.114)	0.329 (0.097)	0.084 (0.023)	0.673 (0.150)	0.346 (0.113)	0.124 (0.038)
		Hauberg	2.444 (0.059)	2.158 (0.155)	0.568 (0.055)	2.544 (0.118)	2.103 (0.563)	0.796 (0.094)
$T = 500$	Circle	Proposed	0.088 (0.026)	0.118 (0.023)	0.091 (0.018)	0.21 (0.050)	0.207 (0.043)	0.129 (0.018)
		Hauberg	0.089 (0.027)	0.205 (0.079)	0.269 (0.034)	0.233 (0.087)	0.453 (0.177)	0.397 (0.079)
	Wave	Proposed	0.535 (0.065)	0.239 (0.056)	0.072 (0.020)	0.574 (0.094)	0.237 (0.082)	0.110 (0.031)
		Hauberg	2.006 (0.697)	1.831 (0.146)	0.529 (0.043)	1.906 (0.847)	1.756 (0.696)	0.688 (0.073)

and $\sigma = 0.07$, 0.1.

Table 3.3 lists the average values of reconstruction errors and their standard deviations over 50 simulation sets. As listed, the proposed (intrinsic) principal curves outperform Hauberg's method, recovering the true curves accurately. Table 3.4 provides the average values of distinct projection points and their standard deviations. The proposed method provides a very large number of distinct projection points compared to that of Hauberg's method. Overall, as listed in Tables 3.3 and 3.4, our methods perform better than that of Hauberg (2016), including the case that the number of points of the curves ($T = 500$) is much larger than the number of data points ($n = 100$). In addition, the results of the intrinsic and extrinsic principal curves are similar in terms of both reconstruction error and the number of distinct projection points, which appear with the fact that the intrinsic and extrinsic means are almost identical for localized data, as noted in Bhattacharya and Patrangenaru (2005). The results of the extrinsic approach are almost identical to those of the intrinsic one, and hence are omitted.

Influence of T and q

Here we discuss the influence of the hyperparameters T and q. To this end, we consider the waveform simulated data used in Section 3.3.2. Figure 3.9 visualizes the fitted curves by the proposed extrinsic method for various q's in the range of [0.01, 0.1] at intervals of 0.01 with a fixed $T = 500$. As shown in the top panels of Figure 3.9, the resulting curve with $q = 0.01$ is wiggly, and the curve with $q = 0.1$ is almost flat. In general, the curves tend to overfit data when the q value is small, whereas the curves tend to underfit data when the q value is large. On the other hand, the bottom panels of Figure 3.9 show the fitted curves by the same method for a fixed $q = 0.06$ and varying T in $\{10, 20, 50, 100, 200, 500\}$. The curve of the bottom left panel implemented by a small T value, such as $T = 10$, does not represent the data well. For appropriate T values, the spherical principal curves of the right panel successfully recover the underlying structure of the data.

Table 3.4: Averages of distinct projection points and their standard deviations in the parentheses

	True form	Method	Noise level					
			$\sigma = 0.07$			$\sigma = 0.1$		
			$q = 0.05$	$q = 0.03$	$q = 0.01$	$q = 0.05$	$q = 0.03$	$q = 0.01$
$T = 100$	Circle	Proposed	99.02 (0.32)	98.92 (0.34)	98.84 (0.47)	99.08 (0.34)	98.72 (1.11)	98.20 (1.12)
		Hauberg	87.70 (7.95)	56.68 (17.99)	64.70 (3.22)	69.80 (12.28)	47.04 (15.44)	60.42 (2.83)
	Wave	Proposed	93.36 (4.47)	97.28 (2.13)	99.32 (0.51)	95.82 (3.77)	96.72 (2.22)	99.10 (0.65)
		Hauberg	22.72 (2.77)	25.94 (2.65)	62.14 (2.49)	24.5 (3.63)	32.16 (16.72)	58.84 (3.04)
$T = 500$	Circle	Proposed	99.08 (0.27)	99.02 (0.25)	98.76 (0.69)	99.1 (0.30)	99.04 (0.49)	99.30 (1.09)
		Hauberg	97.8 (1.47)	89 (8.63)	79.28 (4.60)	93.64 (5.29)	78.72 (13.08)	73.86 (7.28)
	Wave	Proposed	99.18 (0.39)	98.5 (1.27)	99.26 (0.56)	99.22 (0.42)	98.84 (1.20)	99.18 (0.66)
		Hauberg	45.2 (24.8)	43.38 (3.72)	73.20 (3.42)	52.04 (26.81)	50.64 (19.99)	71.52 (4.38)

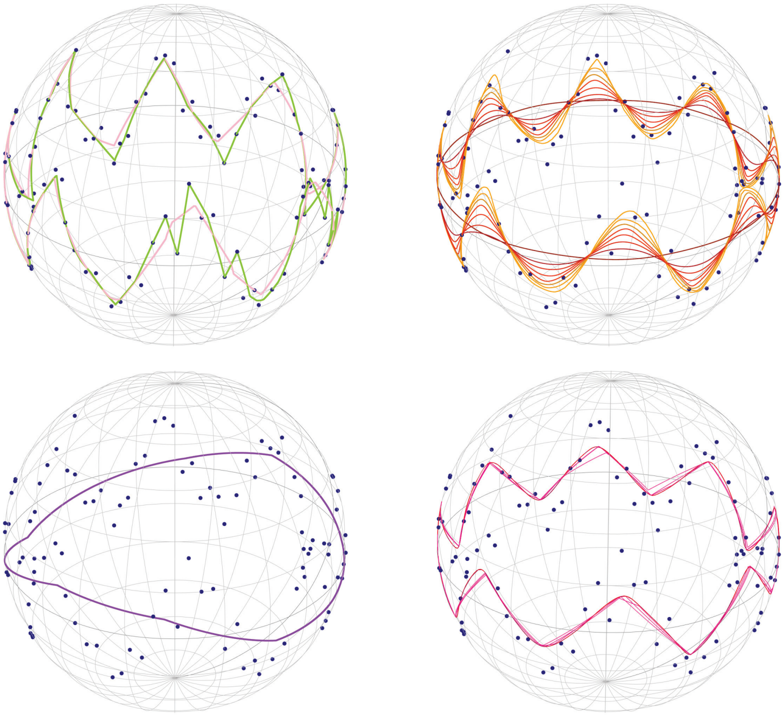


Figure 3.9: Noisy waveform simulated data are colored in blue. Top left: Extrinsic-type principal curves with $q = 0.01$ (green) and 0.02 (pink) for fixed $T = 500$; Top right: Influence of varying q over [0.03, 0.1] with a step size 0.01 (from yellow to brown) for fixed $T = 500$; Bottom left: Extrinsic-type principal curves (purple) with $T = 10$ and $q = 0.06$; Bottom right: Influence of varying T in {20, 50, 100, 200, 500} (from violet to red) for a fixed $q = 0.06$

Simulation on hypersphere

Table 3.5: A simulation result of waveform data on S^4

Method	$q = 0.03$	$q = 0.005$	$q = 0.002$
Proposed (extrinsic)	**0.211 (0.230)**	**0.179 (0.162)**	0.199 (0.235)
Proposed (intrinsic)	0.729 (0.493)	0.267 (0.264)	**0.150 (0.232)**
Hauberg	1.990 (0.815)	0.481 (0.215)	0.357 (0.251)

We conduct a simulation study on a hypersphere. To this end, we consider a waveform simulated data on S^4 represented by four angular parameters $\varphi_1,\ \varphi_2,\ \varphi_3 \in [0,\ \pi)$, and $\varphi_4 \in [0,\ 2\pi)$. The explicit representation on S^D, $D \geq 3$ is given in Section 3.1.3. Mimicking a waveform dataset on S^2 in Section 3.3.2, we craft simulation sets $(r = 1,\ \varphi_1,\ \varphi_2,\ \varphi_3,\ \varphi_4)$ with $\varphi_1 = \varphi_2 = \varphi_3 = \alpha \sin(\varphi_4 f) + \pi/2$ and $0 \leq \varphi_4 < 2\pi$, frequency $f = 2$, and amplitude $\alpha = 1/2$. Data points of $n = 200$ are generated by sampling φ_4 uniformly in $[0,\ 2\pi)$ and adding the random noises sampled from $N(0,\ \sigma^2)$ to φ_1 with $\sigma = 0.05$. Table 3.5 lists the average values of reconstruction errors defined on Section 3.3.2 and their standard deviations over 50 simulation sets for each method with $T = 300$. As listed, the proposed principal curves outperform Hauberg's method, recovering the true curves more closely.

3.4 Proofs

3.4.1 Justification of the projection steps on S^D

Let $A = (a_1,\ a_2,\ \ldots,\ a_{D+1}),\ B = (b_1,\ b_2,\ \ldots,\ b_{D+1}),\ C = (c_1,\ c_2,\ \ldots,\ c_{D+1}) \in S^D \subset \mathbb{R}^{D+1}$ with $(A \cdot B)^2 \neq 1$. Any point P on $\widehat{AB}$ is denoted by $P = \mu A + \lambda B$ for $\mu,\ \lambda \in \mathbb{R}_+$ with $\mu^2 + \lambda^2 = 1$. If $B \cdot C = C \cdot A = 0$, then we have

$$d(C,\ P) = \arccos\big(C \cdot (\mu A + \lambda B)\big) = \pi/2.$$

Hence, any points on $\widehat{AB}$ have the same geodesic distance of $\pi/2$ from C. We may assume that A, B, and C do not satisfy $B \cdot C = C \cdot A = 0$.

The orthogonal complement of V in $\mathbb{R}^{D+1}$, $V^{\perp}$, has a dimension of $D-1$, owing to the fact that $\mathbb{R}^{D+1} = V \oplus V^{\perp}$ with $\oplus$ denoting the direct sum. As a column vector notation, we choose an orthonormal basis for V as R_1, $R_2 \in \mathbb{R}^{D+1}$ and an orthonormal basis for $V^{\perp}$ as $R_3, R_4, \ldots, R_{D+1} \in \mathbb{R}^{D+1}$. Define a $(D+1) \times (D+1)$ matrix $R = [R_1, R_2, R_3, \ldots, R_D, R_{D+1}]^T$. Clearly, R is a rotation (orthogonal) matrix, i.e. $R \in O(n) = \big\{X \in M_{D+1,\,D+1}(\mathbb{R}) \,|\, X^T X = I\big\}$ and satisfies that $RA = (\tilde{a}_1, \tilde{a}_2, 0, 0, \ldots, 0)$ and $RB = (\tilde{b}_1, \tilde{b}_2, 0, 0, \ldots, 0)$. Let $\tilde{A} = RA$, $\tilde{B} = RB$, and $\tilde{C} = RC = (\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_D, \tilde{c}_{D+1})$. Let $\tilde{V}$ be a two-dimensional vector space spanned by $\tilde{A}$ and $\tilde{B}$, as shown in the right panel of Figure 3.10. It follows that $\tilde{V} = \{x = (x_1, x_2, x_3, \ldots, x_{D+1}) \,|\, x_3 = x_4 = \cdots = x_{D+1} = 0\}$. We denote the projection of $\tilde{C}$ onto $\tilde{V}$ as $\tilde{C}' = (\tilde{c}_1, \tilde{c}_2, 0, \ldots, 0) \in \mathbb{R}^{D+1}$ with $\tilde{c}_1^2 + \tilde{c}_2^2 \neq 0$. For any $\tilde{P} = (\tilde{p}_1, \tilde{p}_2, 0, \ldots, 0) \in \tilde{V} \cap S^D$, it follows that

$$d(\tilde{C}, \tilde{P}) = \arccos(\tilde{c}_1 \tilde{p}_1 + \tilde{c}_2 \tilde{p}_2) \geq \arccos(\sqrt{\tilde{c}_1^2 + \tilde{c}_2^2}), \tag{3.15}$$

where the last inequality holds due to the Cauchy-Schwarz inequality $(\tilde{c}_1\tilde{p}_1 + \tilde{c}_2\tilde{p}_2)^2 \leq (\tilde{c}_1^2 + \tilde{c}_2^2)(\tilde{p}_1^2 + \tilde{p}_2^2) = (\tilde{c}_1^2 + \tilde{c}_2^2)$. The equality of (3.15) holds when $(\tilde{p}_1, \tilde{p}_2) = t(\tilde{c}_1, \tilde{c}_2)$ for some $t \in \mathbb{R}_+$. It means that the closest point $\tilde{P}$ on $\tilde{V} \cap S^D$ from $\tilde{C}$ is found by normalizing $\tilde{C}'$ so that it is in S^D. Since R is an orthogonal matrix, for any $P \in V \cap S^D$ and $\tilde{P} = RP \in \tilde{V} \cap S^D$, it follows, as a column vector notation, that

$$d(\tilde{C}, \tilde{P}) = \arccos(\tilde{C}^T \tilde{P}) = \arccos(C^T R^T R P) = \arccos(C^T P) = d(C, P).$$

Accordingly, $\mathrm{proj}(C)$ is obtained by applying R^{-1} to $\mathrm{proj}(\tilde{C})$ that is the projection of $\tilde{C}$ onto $\tilde{V} \cap S^D$. Since the rotation is a rigid motion, it completes the proof.

3.4.2 Stationarity of principal curves

Here we cover a smooth ($\mathcal{C}^3$) curve that does not cross on $\big(i.e.,\ \lambda_1 \neq \lambda_2 \in [0,\,1) \Rightarrow f(\lambda_1) \neq f(\lambda_2)\big)$, including curves with end points and closed curves, which can be both parameterized over interval $[0,\,1]$ by a constant speed, i.e. $f'(\lambda) = s > 0$ for any $\lambda \in [0,\,1]$. In the latter case, a boundary condition is needed; any order partial derivatives of f at end points are the same, i.e., $f^{(k)}(0) = f^{(k)}(1)$ for all $k \geq 0$. For a

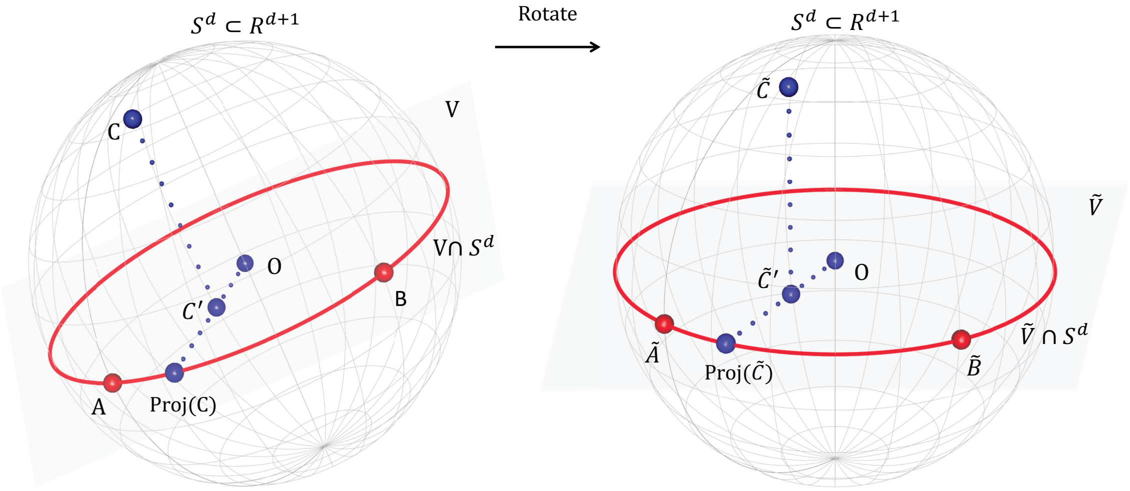


Figure 3.10: (Left) The projection process of C onto the one-dimensional great circle $V \cap S^D$ (red) in a hypersphere $S^D \subset \mathbb{R}^{D+1}$: (i) find the projection of C onto V, C' and (ii) obtain the projection of C' onto $V \cap S^D$, proj(C). (Right) The rotated configuration of the objects.

sphere-valued random variable X, we further assume that the curve f are not short enough to cover the support of X well, i.e., $\lambda_f(X) \neq 0,\ 1$ for a.e. X. For example, any closed curve satisfies the condition $\lambda_f(X) \neq 0,\ 1$ for a.e. X, meaning that almost all X is orthogonally projected onto the curve f. Note that f is smooth on $[0,\ 1]$, i.e. f is smoothly extended on $[0,\ 1]$; thus any order its derivatives are continuous on $[0,\ 1]$. Our main purpose is to prove the stationarity of extrinsic, intrinsic principal curves $f : [0,\ 1] \to S^D$ for $D \geq 2$ that satisfy the (3.12) and (3.14) in Theorems 1 and 2. We first consider the 2-sphere, and then extend D-spheres, $D \geq 2$.

When moving from Euclidean space to spherical surfaces, topological properties such as measurability and continuity are preserved, while the formula using specific distance should be modified. This modification could be obtained by embedding a spherical surface S^D into a $(D+1)$-dimensional Euclidean space. Specifically, we embed a spherical surface as a unit sphere centered at the origin, i.e. $S^D \hookrightarrow \mathbb{R}^{D+1}$, and investigate further derivations. For a smooth curve $f : [0,\ 1] \to S^D \subset \mathbb{R}^{D+1}$, suppose that f is parameterized by a constant speed with respect to λ. The lemma is then held.

Lemma 1. *$f'(\lambda) \cdot f(\lambda) = 0$ and $f''(\lambda) \cdot f'(\lambda) = 0$, $\forall \lambda \in [0, 1]$, where $\cdot$ denotes the standard inner product in $\mathbb{R}^{D+1}$.*

Proof of Lemma 1. It is directly obtained by differentiating $f(\lambda) \cdot f(\lambda) = 1$ and $f'(\lambda) \cdot f'(\lambda) = constant$ by λ. □

When $D = 2$, we assume that f is expressed as three-dimensional coordinates $(f(\lambda)_1,\, f(\lambda)_2,\, f(\lambda)_3)$. Then the following lemmas are held.

Lemma 2. *Suppose that $f(\lambda)$ and x are expressed as three-dimensional vectors. Then, it follows that $d(f(\lambda),\, x) = \arccos(f(\lambda) \cdot x)$, where $\arccos(f(\lambda) \cdot x)$ is the angle between $f(\lambda)$ and x. Then, $\frac{df}{d\lambda}(\lambda_f(0,\, 0,\, 1))_3 = 0$. Thus, it follows that $\frac{df}{d\lambda}(\lambda_f(0,\, 0,\, 1)) = a\big(-f(\lambda_f(0,\, 0,\, 1))_2, f(\lambda_f(0,\, 0,\, 1))_1,\, 0\big)$ for some $a \in \mathbb{R}$. Note that $\lambda_f(x)$ denotes the projection index of point x to the curve f.*

Proof of Lemma 2. For $p = (0,\, 0,\, 1)$, it follows that $d(f(\lambda),\, p) = \arccos(f(\lambda)_3)$. From the assumption that $f(\lambda)$ is a smooth curve and the fact that $d(f(\lambda),\, p)$ has the minimum at $\lambda_f(0,\, 0,\, 1)$, the remaining part of the lemma follows by differentiation with respect to λ. □

Lemma 3. *(Spherical law of cosines) Let $u,\, v,\, w$ be points on a sphere, and a, b and c denote $d(w,\, u)$, $d(w,\, v)$ and $d(u,\, v)$, respectively. If C is the angle between a and b, i.e., the angle of the corner opposite c, then, $\cos c = \cos a \cos b + \sin a \sin b \cos C$. Further, with three-dimensional vectors $u,\, v,\, w$, it follows that $\sin a \sin b \cos C = (w \times u) \cdot (w \times v)$, where $\times$ denotes cross product in $\mathbb{R}^3$.*

For arbitrary $D \leq 2$, the following property can be obtained from Definition 2.

Proposition 1. *Under the same conditions in Definition 2, $f(\epsilon,\, \lambda) := (f + \epsilon g)(\lambda)$ is smooth on $[-1,\, 1] \times [0,\, 1]$. Hence, for any $\epsilon \in [-1,\, 1]$, $f + \epsilon g : [0,\, 1] \to S^D$ is a smooth curve on S^D and $\lim_{\epsilon \to 0}(f + \epsilon g)(\lambda) = f(\lambda)$ for $\lambda \in [0,\, 1]$.*

Proof of Proposition 1. For simplicity, we denote $f + g$ by h. Let $R_{a,b}(\theta)$ be a rotation matrix that rotates points on S^D by θ in the direction along the geodesic from a to b with $a, b \in S^D \subset \mathbb{R}^{D+1}$ satisfying $a \neq -b \in \mathbb{R}^{D+1}$ and θ ranging over $[0, \pi)$. The rotation matrix has a closed form. Specifically, for $a \neq b$ $R_{a,b}(\theta) = I_{D+1} + \sin(\theta)B + (\cos(\theta) - 1)(bb^T + cc^T)$ where T denotes the transpose of a matrix, $c = (a - b(b^T a))/\left\|a - b(b^T a)\right\|$, and $B = bc^T - cb^T$ as a column vector notation. (see Section 8.1 in Jung et al. (2012) for more details). In this respect we obtain $f(\epsilon, \lambda) := (f + \epsilon g)(\lambda) = R_{f(\lambda),h(\lambda)}(\theta) \cdot f(\lambda)$, where $\theta = \epsilon \arccos(f(\lambda) \cdot h(\lambda))$. Thus, $(f + \epsilon g)(\lambda)$ $(= f(\epsilon, \lambda))$ is smooth on $[-1, 1] \times [0, 1]$ since all functions f, h, R, and θ are smooth. Therefore, for a fixed $\epsilon \in [0, 1]$, the smoothness of $(f + \epsilon g)(\lambda)$ for $\lambda \in [0, 1]$ also follows. Moreover, the last equality is a direct consequence of the definition of $f + \epsilon g$. □

From now on, we present the detailed explanation for Definitions 2 and 3 stated in Section 3.2.4. For a given $f + g$ $(= h)$, the $\|g\|$ and $\|g'\|$ is defined by mimicking the Euclidean case. For examples, in Euclidean space $(f + g)(\lambda) = f(\lambda) + g(\lambda)$. We thus obtain that $g(\lambda) = \frac{\partial}{\partial \epsilon} f_\epsilon(\lambda)$ where $f_\epsilon(\lambda) := f + \epsilon g(\lambda)$. From this fact, on spheres, the magnitude of perturbation $\|h - f\|$ is defined by $g(\epsilon_0, \lambda) := \frac{\partial}{\partial \epsilon}\big|_{\epsilon=\epsilon_0} f_\epsilon(\lambda)$, $\|g(\lambda)\| = d(f(\lambda), g(\lambda)) = \left|g(\epsilon_0, \lambda)\right|$, and finally $\|h - f\| = \|g\| = \max_\lambda \|g(\lambda)\| \neq \pi$. The boundedness of $\|g\|$ guarantees that the ϵ-internal division between f and h, f_ϵ, uniformly converges to f on $\lambda \in [0, 1]$ as $\epsilon \to 0$. Notice that from the compactness of the unit sphere, $\|h - f\|$ is inherently not greater than π; thus, $\|h - f\| \neq \pi$ implies that $\|h - f\| < \pi$. Moreover, the norm of derivative of perturbation $\|(h - f)'\|$ is defined by $g'(\epsilon, \lambda_0) = \frac{\partial}{\partial \lambda}\big|_{\lambda=\lambda_0} g(\epsilon, \lambda)$, $\|g'(\lambda_0)\| = \max_\epsilon \|g'(\epsilon, \lambda_0)\|$, and finally $\|(h - f)'\| = \|g'\| := \max_{\lambda_0} \|g'(\lambda_0)\|$. Let x be a point on a sphere. By the continuity of f and the compactness of the domain $[0, 1]$, $\min_{\lambda \in [0, 1]} d(x, f(\lambda))$ can be attained. Let $d(x, f)$ denote the geodesic distance from x to f, i.e., $d(x, f) := \min_{\lambda \in [0, 1]} d(x, f(\lambda))$. By the continuity of f again, $\{\lambda \in [0, 1] \,|\, d(x, f(\lambda)) = d(x, f)\}$ is closed and therefore compact. Thus, the projection indices $\lambda_f(x) = \min\{\lambda \,|\, d(x, f(\lambda)) = d(x, f)\}$ and $\lambda_{f+\epsilon g}$ are well-defined. The latter holds due to the fact that $f + \epsilon g$ is a continuous curve by Proposition 1. If the set $\{\lambda \,|\, d(x,\, f(\lambda)) = d(x,\, f)\}$ not a singleton, then the point x is called an *ambiguity point* of f. The set of ambiguity points of the smooth curve has spherical measure 0; thus, the ambiguity points are negligible when calculating the expected value.

Proposition 2. *Spherical measure of the set of ambiguity points of smooth curve f is 0.*

Detailed steps for a proof of Proposition 2 are similar with those of Hastie and Stuetzle (1989); thus, we omit the proof. Meanwhile, Hastie (1984) proved that the index function $\lambda_f : x \mapsto \lambda_f(x)$ is measurable, while the proof is not perfectly correct. The modified proof is shown as follows.

Proposition 3. *(Measurability of index function) For a continuous curve f on S^D, the index function $\lambda_f : S^D \to [0,\, 1]$ by $x \mapsto \lambda_f(x)$ is measurable.*

Proof of Proposition 3. See Appendix A.1. □

According to the Proposition 3, $\lambda_f(X)$ is a random variable, provided that X is a S^D-valued random variable for $D \geq 2$. Thus, a conditional expectation on $\lambda_f(X)$ is feasible.

Proposition 4. *(Continuity of projection index under perturbation) If x is not an ambiguity point for continuous curve f, then $\lim_{\epsilon \to 0} \lambda_{f+\epsilon g}(x) = \lambda_f(x)$.*

Proof of Proposition 4. See Appendix A.1. □

In the proof of Proposition 4, it is possible to apply the triangle inequality on a sphere because the sphere is a metric space equipped with its geodesic distance. The following proposition is an useful tool for proving the Theorems 1 and 2.

Proposition 5. *(Uniform continuity of projection index under perturbation) $\lim_{\epsilon \to 0} \lambda_{f+\epsilon g}(x) = \lambda_f(x)$ uniformly on the set of non-ambiguity points of f. That is, for every $\eta > 0$, there exists a $\delta > 0$ such that for any non-ambiguity points x, if $|\epsilon| < \delta$, then $|\lambda_{f_\epsilon}(x) - \lambda_f(x)| < \eta$.*

To show the uniform continuity of projection index, it is required that $|f''|$ is bounded. It is a direct result from the smoothness of f ($\mathcal{C}^3$) and compactness of $[0, 1]$. A proof is similar to that of Proposition 4; thus, we omit the proof. Meanwhile, to prove Theorems, it is needed to show that λ_{f_ϵ} is differentiable for ϵ and its derivative is uniformly bounded. To this end, it is necessary to define a subset of S^D as $B(\zeta) = \{x \in S^D \mid |f''(\lambda_f(x)) \cdot x| > \zeta\}$ for $\zeta \geq 0$. Obviously, the collection of sets $\{B(\zeta)\}_{\zeta \geq 0}$ is decreasing as $\zeta \to 0$. Moreover, the following lemma implies that as $\zeta \to 0$ $B(\zeta)$ covers S^D almost everywhere.

Lemma 4. *The image of smooth function from* $[0, 1]$ *to* S^D *has measure 0. Moreover,*

$$S^D \setminus B(0) = \{x \in S^D \,|\, |f''(\lambda_f(x)) \cdot x| = 0\}$$

is a union of images of two smooth functions from $[0, 1]$ *to* S^D*, which implies that* $S^D \setminus B(0)$ *has spherical (D-dimensional Hausdorff) measure 0. Therefore, the measure of* $S^D \setminus B(\zeta) \to 0$ *as* $\zeta \to 0$.

Proof. See Appendix A.1. □

Note that Lemma 4 implies the constraints of a random variable X in Theorems 1 and 2 are almost negligible by letting ζ *arbitrarily* close to zero. Denote the set of ambiguity points of smooth curve f on a sphere as A, which has measure zero by Proposition 2.

Lemma 5. *Let* A *be the set of ambiguity points of smooth curve* f *on a sphere. Suppose that, for any* $x \in S^2$, $\lambda_f(x) \in (0, 1)$, *and* $x \in A^c \cap B(\zeta)$ *for arbitrarily small* $\zeta > 0$. *Then* $\lambda(\epsilon) := \lambda_{f_\epsilon}(x)$ *is a smooth function for* ϵ *on an open interval containing 0. Moreover,* $\frac{\partial \lambda(\epsilon)}{\partial \epsilon}$ *is uniformly bounded on* $A^c \cap B(\zeta)$. *That is, there are constants* $C > 0$ *and* $\delta > 0$ *such that if* $|\epsilon_0| < \delta$ *and* $x \in A^c \cap B(\zeta)$, *then* $\left| \frac{\partial \lambda_{f_\epsilon}(x)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} \right| < C$.

Proof. See Appendix A.1. □

Proof of Theorem 1

Proof. First of all, we prove the theorem on S^2. If $f = h$, then nothing to prove. Thus, we assume that the curves f and $f + g$ $(= h)$ are not identical and further both are parameterized by $\lambda \in [0,\, 1]$. To prove the result, we need to show that the conditional expectation is zero after exchanging the order of the derivative and expectation.

First, for order exchange, it is necessary to show that the following random variable

$$\begin{aligned} Z_\epsilon(X) &= \frac{\cos\big(d(X,\, f+\epsilon g)\big) - \cos\big(d(X,\, f)\big)}{\epsilon} \\ &= \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_{f+\epsilon g}(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_f(X))\big)\big)}{\epsilon} \end{aligned} \quad (3.16)$$

is uniformly bounded for any sufficiently small $|\epsilon| > 0$. Then we apply bounded convergence theorem. Since the projection index of X represents the closest point in the curve,

$$Z_\epsilon(X) \leq \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_{f+\epsilon g}(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_{f+\epsilon g}(X))\big)\big)}{\epsilon}. \quad (3.17)$$

For simplicity, Denote $f_g(\lambda_\epsilon) = (f+g)(\lambda_{f+\epsilon g}(X))$, $f_\epsilon(\lambda_\epsilon) = (f+\epsilon g)(\lambda_{f+\epsilon g}(X))$, and $f(\lambda_\epsilon) = f(\lambda_{f+\epsilon g}(X))$. By applying Lemma 3 to $\cos\big(d(X,\, f_\epsilon(\lambda_\epsilon))\big)$, the inequality of (3.17) becomes

$$\begin{aligned} Z_\epsilon(X) &\leq \frac{\cos\big(d\big(X,\, f_\epsilon(\lambda_\epsilon)\big)\big) - \cos\big(d\big(X,\, f(\lambda_\epsilon)\big)\big)}{\epsilon} \\ &= \frac{\cos(d(X,\, f(\lambda_\epsilon)))(\cos(d(f_\epsilon(\lambda_\epsilon),\, f(\lambda_\epsilon))) - 1) + (f(\lambda_\epsilon) \times f_\epsilon(\lambda_\epsilon)) \cdot (f(\lambda_\epsilon) \times X)}{\epsilon}. \\ &= \frac{\cos(d(X,\, f(\lambda_\epsilon)))(\cos(\epsilon d(f_g(\lambda_\epsilon),\, f(\lambda_\epsilon))) - 1) + A_\epsilon(f(\lambda_\epsilon) \times f_g(\lambda_\epsilon)) \cdot (f(\lambda_\epsilon) \times X)}{\epsilon}, \end{aligned} \quad (3.18)$$

where

$$A_\epsilon = |f(\lambda_\epsilon) \times f_\epsilon(\lambda_\epsilon)|/|f(\lambda_\epsilon) \times f_g(\lambda_\epsilon)| = \sin(\epsilon d(f(\lambda_\epsilon), f_g(\lambda_\epsilon)))/|f(\lambda_\epsilon) \times f_g(\lambda_\epsilon)|.$$

The last equality is done by Definition 1. To get the upper bound of $Z_\epsilon(X)$, we further use the following fact, $|\frac{\sin \epsilon C}{\epsilon}| \leq |C|$ and $|\frac{1-\cos \epsilon C}{\epsilon}| \leq \frac{|\epsilon| C^2}{2}$ for $C \in \mathbb{R}$ and

$\epsilon \in \mathbb{R}$. Then, we have

$$Z_\epsilon(X) \;\; \leq \;\; \big|\cos\big(d\big(X,\, f(\lambda_\epsilon)\big)\big)\big|\frac{|\epsilon|B^2}{2} + \frac{B}{|f(\lambda_\epsilon)\times f_g(\lambda_\epsilon)|}\big|(f(\lambda_\epsilon)\times f_g(\lambda_\epsilon))\cdot(f(\lambda_\epsilon)\times X)\big|,$$

where $B = d\big(f(\lambda_\epsilon),\, f_g(\lambda_\epsilon)\big) \leq \|g\| < \pi$. Note that any smallest geodesic distance on a unit sphere is smaller than π. In addition, we can assume that ϵ is less than $1/\pi$ because we are only interested in ϵ near 0. Thus, we obtain the upper bound of $Z_\epsilon(X)$ in (3.18)

$$Z_\epsilon(X) \leq \frac{\pi}{2} + \pi = \frac{3\pi}{2}.$$

A lower bound of $Z_\epsilon(X)$ can be similarly obtained. Let $f_g(\lambda) := (f+g)(\lambda_f(X))$, $f_\epsilon(\lambda) := (f+\epsilon g)(\lambda_f(X))$ and $f(\lambda) := f(\lambda_f(X))$. By following the same path, we have

$$\begin{aligned} Z_\epsilon(X) \;\; &\geq \;\; \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_f(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_f(X))\big)\big)}{\epsilon} \\ &= \;\; \frac{\cos\big(d\big(X,\, f(\lambda)\big)\big)\big(\cos\big(d\big(f_\epsilon(\lambda),\, f(\lambda)\big)\big) - 1\big) + (f(\lambda)\times f_\epsilon(\lambda))\cdot(f(\lambda)\times X)}{\epsilon} \\ &= \;\; \frac{\cos\big(d\big(X,\, f(\lambda)\big)\big)\big(\cos\big(\epsilon d\big(f_g(\lambda),\, f(\lambda)\big)\big) - 1\big) + B_\epsilon(f(\lambda)\times f_g(\lambda))\cdot(f(\lambda)\times X)}{\epsilon}, \end{aligned} \tag{3.19}$$

where

$$B_\epsilon = |f(\lambda)\times f_\epsilon(\lambda)|/|f(\lambda)\times f_g(\lambda)| = \sin(\epsilon d(f(\lambda),\, f_g(\lambda)))/|f(\lambda)\times f_g(\lambda)|.$$

By the same way, it can be shown that

$$Z_\epsilon(X) \geq -\frac{3\pi}{2}.$$

Hence, we show that

$$|Z_\epsilon(X)| \leq \frac{3\pi}{2},$$

which is bounded for any $0 \neq |\epsilon| \leq 1/\pi$. Then, by Lebesgue's dominated convergence theorem, it follows that

$$\frac{\partial \mathbb{E}_X \cos(d\big(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0} = \mathbb{E}_X \frac{\partial \cos(d\big(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0}.$$

Thus, the proof is completed provided that the following equation holds

$$\mathbb{E}\Big[\frac{\partial \cos\big(d(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0} \;\Big|\; \lambda_f(X)=\lambda\Big]=0, \text{ for a.e. } \lambda.$$

By the definition of derivative,

$$\frac{\partial \cos\big(d(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0} = \lim_{\epsilon\to 0} Z_\epsilon(X),$$

and as shown, $Z_\epsilon(X)$ is bounded. Since f and $f+g$ are continuous, by Proposition 1, if X is not an ambiguity point of f and $f+g$, then

$$\lim_{\epsilon\to 0} f_g(\lambda_\epsilon)=f_g(\lambda), \quad \lim_{\epsilon\to 0} f(\lambda_\epsilon)=f(\lambda).$$

Next, to show the limit of Z_ϵ, we use the fact that $\lim_{\epsilon\to 0}\frac{\sin \epsilon C}{\epsilon}=C$ and $\lim_{\epsilon\to 0}\frac{1-\cos\epsilon C}{\epsilon}=0$ for $C\in\mathbb{R}$ and $\epsilon\in\mathbb{R}$. When $f_g(\lambda)\neq f(\lambda)$, it follows that

$$\lim_{\epsilon\to 0} \text{ RHS of (3.18)} = d\big(f(\lambda),\, f_g(\lambda)\big)\frac{f(\lambda)\times f_g(\lambda)}{|f(\lambda)\times f_g(\lambda)|}\cdot(f(\lambda)\times X)=\mu(\lambda)\cdot(f(\lambda)\times X),$$

where $\mu(\lambda)=d(f(\lambda),\, f_g(\lambda))(f(\lambda)\times f_g(\lambda))/|f(\lambda)\times f_g(\lambda)|$ if $f(\lambda)\neq(f+g)(\lambda)$ and $\mu(\lambda)=0$ otherwise. Similarly, we obtain that

$$\lim_{\epsilon\to 0} \text{ RHS of (3.19)} = \mu(\lambda)\cdot(f(\lambda)\times X).$$

In summary, if X is not an ambiguity point of f and $f+g$, and $f(\lambda_f(X))\neq(f+g)(\lambda_f(X))$, then we have

$$\frac{\partial \cos\big(d(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0} = \mu(\lambda_f(X))\cdot\big(f(\lambda_f(X))\times X\big). \tag{3.20}$$

In the case of $f(\lambda_f(X))=(f+g)(\lambda_f(X))$, the equation (3.20) also hold because its left- and right-hand sides are 0. From Proposition 2, the limit of (3.20) is established for a.e. X. Note that, since X is a random variable and $\lambda_f(X)$ is measurable with respect to X according to Proposition 3, $\lambda_f(X)$ is a random variable. It implies that conditional expectation on $\lambda_f(X)$ is feasible. Hence, the following equality holds

$$\mathbb{E}_X\Big[\frac{\partial \cos\big(d(X,\, f+\epsilon g)\big)}{\partial \epsilon}\Big|_{\epsilon=0}\Big]=\mathbb{E}_X\Big[\mu(\lambda_f(X))\cdot\big(f(\lambda_f(X))\times X\big)\Big]. \tag{3.21}$$

Finally, if f is an extrinsic principal curve, then

$$\mathbb{E}_X\big[X \mid \lambda_f(X) = \lambda\big] = cf(\lambda)$$

for $\exists c \in \mathbb{R}$. Hence, it follows that

$$\begin{aligned}\mathbb{E}\big[\mu(\lambda_f(X)) \cdot \big(f(\lambda_f(X)) \times X\big) \mid \lambda_f(X) = \lambda\big] &= \mathbb{E}\big[\mu(\lambda) \cdot (f(\lambda) \times X) \mid \lambda_f(X) = \lambda\big] \\ &= \mu(\lambda) \cdot (f(\lambda) \times cf(\lambda)) = 0.\end{aligned}$$

Hence, we have

$$\begin{aligned}\text{LHS of (3.21)} &= \mathbb{E}_X\big[\mu(\lambda) \cdot (f(\lambda) \times X)\big] \\ &= \mathbb{E}_\lambda\big[\mathbb{E}\big[\mu(\lambda) \cdot (f(\lambda) \times X) \mid \lambda_f(X) = \lambda\big]\big] = 0.\end{aligned}$$

To prove the converse, we assume that

$$\begin{aligned}&\mathbb{E}_\lambda\big[\mathbb{E}\big(\mu(\lambda) \cdot (f(\lambda) \times X) \mid \lambda_f(X) = \lambda\big)\big] \\ &\qquad = \mathbb{E}_\lambda\big[\mu(\lambda) \cdot \mathbb{E}\big[f(\lambda) \times X \mid \lambda_f(X) = \lambda\big]\big] = 0,\end{aligned}$$

for all smooth $f + g$ satisfying $\|g\| < \pi$ and $\|g'\| \leq 1$. Since $f + g$ is only concerned with $\mu(\lambda)$, it follows that

$$\mathbb{E}\big[f(\lambda) \times X \mid \lambda_f(X) = \lambda\big] = f(\lambda) \times \mathbb{E}\big[X \mid \lambda_f(X) = \lambda\big] = 0, \quad \text{for a.e. } \lambda.$$

Therefore, we have

$$\mathbb{E}\big[X \mid \lambda_f(X) = \lambda\big] = cf(\lambda)$$

for $\exists c \geq 0$, which completes the proof.

Next, we consider the hypersphere case S^D for $D \geq 3$. For given smooth curves f and h $(= f + g)$ parametrized by $\lambda \in [0,\, 1]$, if $f = h$, the result is obvious. Thus, we assume that f and $f+g(= h)$ are not identical. Suppose that $X \in A^c \cap B(\zeta)$ for some small $\zeta > 0$ and $\lambda_f(X) \in (0,\, 1)$ for a.e. X, where A denotes the set of ambiguity points of f. As the proof of the case of S^2, we use the bounded convergence theorem to change the order of derivative and expectation. Since $d(x,\, y) = \arccos(x \cdot y)$ for any $x,\, y \in S^D \subset \mathbb{R}^{D+1}$, we have

$$\begin{aligned}
Z_\epsilon(X) &:= \frac{\cos\big(d(X,\, f+\epsilon g)\big) - \cos\big(d(X,\, f)\big)}{\epsilon} \\
&= \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_{f+\epsilon g}(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_f(X))\big)\big)}{\epsilon} \\
&\leq \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_{f+\epsilon g}(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_{f+\epsilon g}(X))\big)\big)}{\epsilon} \\
&= \frac{X \cdot (f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - X \cdot f(\lambda_{f+\epsilon g}(X))}{\epsilon} \\
&= X \cdot \frac{(f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - f(\lambda_{f+\epsilon g}(X))}{\epsilon}, \qquad (3.22)
\end{aligned}$$

where $\cdot$ denotes the standard inner product in $\mathbb{R}^{D+1}$. Hence, we obtain the upper bound of $Z_\epsilon(X)$,

$$\begin{aligned}
Z_\epsilon(X) &\leq \|X\| \frac{\|(f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - f(\lambda_{f+\epsilon g}(X))\|}{\epsilon} \\
&\leq \frac{d\big((f+\epsilon g)(\lambda_{f+\epsilon g}(X)),\, f(\lambda_{f+\epsilon g}(X))\big)}{\epsilon} \\
&\leq \|g(\lambda_{f+\epsilon g}(X))\| \leq \|g\| \\
&\leq \pi,
\end{aligned}$$

where $\|\cdot\|$ denotes the standard norm in $\mathbb{R}^{D+1}$. Similarly, it follows that

$$\begin{aligned}
Z_\epsilon(X) &\geq \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_f(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_f(X))\big)\big)}{\epsilon} \\
&= \frac{X \cdot (f+\epsilon g)(\lambda_f(X)) - X \cdot f(\lambda_f(X))}{\epsilon} \\
&= X \cdot \frac{(f+\epsilon g)(\lambda_f(X)) - f(\lambda_f(X))}{\epsilon} \\
&\geq -\|X\| \frac{\|(f+\epsilon g)(\lambda_f(X)) - f(\lambda_f(X))\|}{\epsilon} \\
&\geq -\frac{d\big((f+\epsilon g)(\lambda_f(X)),\, f(\lambda_f(X))\big)}{\epsilon} \\
&\geq -\|g(\lambda_f(X))\| \geq -\|g\| \\
&\geq -\pi.
\end{aligned}$$

It means that $Z_\epsilon(X)$ is uniformly bounded for $0 \neq |\epsilon| \leq 1$. Next, to find the limit of $Z_\epsilon(X)$, we have

$$\begin{aligned}
Z_\epsilon(X) &= \frac{\cos\big(d\big(X,\, (f+\epsilon g)(\lambda_{f+\epsilon g}(X))\big)\big) - \cos\big(d\big(X,\, f(\lambda_f(X))\big)\big)}{\epsilon} \\
&= \frac{X \cdot (f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - X \cdot f(\lambda_f(X))}{\epsilon} \\
&= X \cdot \frac{(f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - f(\lambda_f(X))}{\epsilon}.
\end{aligned}$$

According to Proposition 5,

$$\begin{aligned}
\lim_{\epsilon \to 0} Z_\epsilon(X) &= X \cdot \lim_{\epsilon \to 0} \frac{(f+\epsilon g)(\lambda_{f+\epsilon g}(X)) - f(\lambda_f(X))}{\epsilon} \\
&=: X \cdot \phi(\lambda_f(X)).
\end{aligned}$$

For each $X \in A^c \cap B(\zeta)$, define a curve $C : I \to S^D$ by $\epsilon \mapsto C(\epsilon) = (f + \epsilon g)(\lambda_{f+\epsilon g}(X)) \in S^D \subset \mathbb{R}^{D+1}$, where I is an open interval containing zero and $C(0) = f(\lambda_f(X))$. For convenience, let $(f + \epsilon g)(\lambda) = f(\epsilon,\, \lambda)$, $\lambda_f(X) = \lambda(0)$ and $\lambda_{f+\epsilon g}(X) = \lambda(\epsilon)$. According to Proposition 1, $\lambda(\epsilon)$ is a smooth function on an interval I containing zero. As $f(\cdot,\, \cdot)$ is smooth on $[-1,\, 1] \times [0,\, 1]$ by Proposition 1 and $\lambda(\epsilon)$ is smooth on $\epsilon \in I$, $C(\epsilon) = f\big(\epsilon, \lambda(\epsilon)\big)$ is also smooth on $\epsilon \in I$. Thus, $\phi(\lambda)$ is well defined. Hence, by the definition of tangent space via tangent curves,

$$\phi(\lambda) = \lim_{\epsilon \to 0} \frac{C(\epsilon) - C(0)}{\epsilon} = C'(0) \in T_{f(\lambda)} S^d,$$

where $T_{f(\lambda)} S^D$ is the tangent space of S^D at $f(\lambda)$. Note that, by the symmetry of spheres, any tangent vector in $T_{f(\lambda)} S^D$ is orthogonal to the vector $f(\lambda)$, i.e., $\phi(\lambda) \cdot f(\lambda) = 0$. Finally, if f is an extrinsic principal curve, then

$$\mathbb{E}\big[X \mid \lambda_f(X) = \lambda\big] = c f(\lambda)$$

for $\exists c \in \mathbb{R}$. Hence, it follows, by the bounded convergence theorem, that

$$\begin{aligned}
\frac{\partial \mathbb{E}_X\big[\cos\big(d(X,\, f+\epsilon g)\big)\big]}{\partial \epsilon}\Big|_{\epsilon=0} &= \lim_{\epsilon\to 0}\frac{\mathbb{E}_X\big[\cos\big(d(X,\, f+\epsilon g)\big)\big]-\mathbb{E}_X\big[\cos\big(d(X,\, f)\big)\big]}{\epsilon}\\
&= \mathbb{E}_X\big[\lim_{\epsilon\to 0}\frac{\cos(d(X,\, f+\epsilon g))-\cos(d(X,\, f))}{\epsilon}\big]\\
&= \mathbb{E}_\lambda\big[\mathbb{E}\big[\lim_{\epsilon\to 0} Z_\epsilon(X)\ \big|\ \lambda_f(X)=\lambda\big]\big]\\
&= \mathbb{E}_\lambda\big[\mathbb{E}\big[\phi(\lambda)\cdot X\ \big|\ \lambda_f(X)=\lambda\big]\big]\\
&= \mathbb{E}_\lambda\big[\phi(\lambda)\cdot\mathbb{E}\big[X\ \big|\ \lambda_f(X)=\lambda\big]\big]\\
&= \mathbb{E}_\lambda\big[\phi(\lambda)\cdot cf(\lambda)\big]\\
&= 0.
\end{aligned}$$

To prove the converse, we assume that f satisfies

$$\begin{aligned}
0 &= \frac{\partial \mathbb{E}_X\big[\cos\big(d(X,\, f+\epsilon g)\big)\big]}{\partial \epsilon}\Big|_{\epsilon=0}\\
&= \lim_{\epsilon\to 0}\frac{\mathbb{E}_X\big[\cos\big(d(X,\, f+\epsilon g)\big)\big]-\mathbb{E}_X\big[\cos\big(d(X,\, f)\big)\big]}{\epsilon}\\
&= \mathbb{E}_X\big[\lim_{\epsilon\to 0}\frac{\cos\big(d(X,\, f+\epsilon g)\big)-\cos\big(d(X,\, f)\big)}{\epsilon}\big]\\
&= \mathbb{E}_\lambda\big[\mathbb{E}\big[\lim_{\epsilon\to 0} Z_\epsilon(X)\ \big|\ \lambda_f(X)=\lambda\big]\big]\\
&= \mathbb{E}_\lambda\big[\mathbb{E}\big[\phi(\lambda)\cdot X\ \big|\ \lambda_f(X)=\lambda\big]\big]\\
&= \mathbb{E}_\lambda\big[\phi(\lambda)\cdot\mathbb{E}\big[X\ \big|\ \lambda_f(X)=\lambda\big]\big],
\end{aligned}$$

for any smooth curve $h : [0,\, 1] \to S^D$. Since h is arbitrary, ϕ can become any vector in $T_{f(\lambda)}S^D$. In addition, h is only concerned with ϕ. We thus obtain, for a.e. λ, the following condition:

$$\phi\cdot\mathbb{E}\big[X\,|\,\lambda_f(X)=\lambda\big]=0 \text{ for any } \phi\in T_{f(\lambda)}S^D.$$

It means that $\mathbb{E}\big[X\,|\,\lambda_f(X)=\lambda\big]$ is orthogonal to $T_{f(\lambda)}S^D$. Therefore, it follows that

$$\mathbb{E}\big[X\,|\,\lambda_f(X)=\lambda\big]=cf(\lambda)$$

for $\exists c \geq 0$, which completes the proof. □

Proof of Theorem 2

Proof. In the case of $f = h$, the result is obvious. We thus assume that f and $f + g$ $(= h)$ are not identical. Further, suppose that $X \in A^c \cap B(\zeta)$ for a small $\zeta > 0$ and $\lambda_f(X) \in (0,\, 1)$ for a.e. X. As the proof of Theorem 1, we use the bounded convergence theorem to change the order of derivative and expectation. For this purpose, we define

$$\begin{aligned} Z_\epsilon(X) &= \frac{d^2(X,\, f + \epsilon g) - d^2(X,\, f)}{\epsilon} \\ &= \frac{d^2\big(X,\, f_\epsilon(\lambda_{f_\epsilon})\big) - d^2\big(X,\, f(\lambda_f)\big)}{\epsilon}, \end{aligned}$$

where $f_\epsilon := f + \epsilon g$ for $|\epsilon| \leq 1$. Let $\theta(\lambda,\, X)$ be the angle between segments of geodesics from $f(\lambda)$ to X and from $f(\lambda)$ to $(f + g)(\lambda)$. Then, from Lemma 3, it follows that

$$\begin{aligned} F(\epsilon) :=& \cos\big(d\big(X,\, f_\epsilon(\lambda_{f_\epsilon})\big)\big) \\ =& \cos\big(d\big(X,\, f(\lambda_{f_\epsilon})\big)\big) \cdot \cos\big(\epsilon\, \|g(\lambda_{f_\epsilon})\|\,\big) \\ &+ \sin\big(d\big(X,\, f(\lambda_{f_\epsilon})\big)\big) \cdot \sin\big(\epsilon\, \|g(\lambda_{f_\epsilon})\|\,\big) \cdot \cos\big(\theta(\lambda_{f_\epsilon}, X)\big), \end{aligned}$$

where $\|g(\lambda)\| = d\big(f(\lambda),\, (f + g)(\lambda)\big) < \pi$.

Firstly, we verify that $Z_\epsilon(X)$ is uniformly bounded for a small $|\epsilon| > 0$. By Lemma 5, there are constants $C > 0$ and $\eta > 0$ such that if $0 < |\epsilon_0| < \eta$, then $\lambda(\epsilon)$ is differentiable at $\epsilon = \epsilon_0$ and $\big|\frac{\partial \lambda(\epsilon)}{\partial \epsilon}\big|_{\epsilon = \epsilon_0}\big| < C$, where $\lambda(\epsilon) = \lambda_{f_\epsilon}(X)$. For convenience, let $\lambda_{f_\epsilon}(X) = \lambda_\epsilon$ and $\lambda_f(X) = \lambda_0$. If $0 < |\epsilon_0| < \eta$, then by the triangle inequality on sphere and mean value theorem, we have

$$\begin{aligned}
|Z_{\epsilon_0}(X)| &= \left|\frac{d\big(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})\big) - d\big(X,\, f(\lambda_f)\big)}{\epsilon_0}\right| \cdot \Big(d\big(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})\big) + d\big(X,\, f(\lambda_f)\big)\Big) \\
&\leq 2\pi \cdot \frac{d\big(f(\lambda_0),\, f_{\epsilon_0}(\lambda_{\epsilon_0})\big)}{\epsilon_0} \\
&\leq 2\pi \cdot \left[\frac{d\big(f(\lambda_0),\, f(\lambda_{\epsilon_0})\big)}{\epsilon} + \frac{d\big(f(\lambda_{\epsilon_0}),\, f_{\epsilon_0}(\lambda_{\epsilon_0})\big)}{\epsilon_0}\right] \\
&< 2\pi \cdot \big(s \cdot \frac{|\lambda_0 - \lambda_{\epsilon_0}|}{\epsilon_0} + \|g(\lambda_{\epsilon_0})\|\,\big) \\
&\leq 2\pi \cdot (s \cdot C + \pi),
\end{aligned}$$

where $s = |f'(\lambda)|$ for all λ. Therefore, $Z_\epsilon(X)$ is uniformly bounded on $X \in A^c \cap B(\zeta)$ for $0 < |\epsilon| < \eta$.

Secondly, we aim to find the limit of $Z_\epsilon(X)$. For this purpose, we define a map $u : (-1,\, 1] \to (1, \infty)$ by $u(x) = \arccos(x) \cdot \frac{1}{\sqrt{1-x^2}}$ if $x \in (-1,\, 1)$, and $u(1) = 1$. By simple calculations, u is a monotone decreasing continuous function on $(-1,\, 1]$. Note that $F(\epsilon)$ is differentiable for $|\epsilon| < \eta$. By the mean value theorem to find the limit of $Z_\epsilon(X)$, we have

$$\begin{aligned}
Z_{\epsilon_0}(X) &= \frac{d^2\big(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})\big) - d^2\big(X,\, f(\lambda_f)\big)}{\epsilon_0} \\
&= \frac{\arccos^2\big(F(\epsilon_0)\big) - \arccos^2\big(F(0)\big)}{\epsilon_0} \\
&= -2\arccos\big(F(\epsilon_1)\big) \cdot \frac{1}{\sqrt{1 - F^2(\epsilon_1)}} \cdot \frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1} \qquad (3.23)
\end{aligned}$$

for $0 < |\epsilon_1| < |\epsilon_0| < \eta$. When $F(\epsilon_1) = 1$, the last equality is considered as a limit that is well-defined, because $\lim_{x\to 1} u(x) = 1$ and $u(x)$ is smoothly extended on an open interval containing 1 such that $u(x)$ is differentiable at $x = 1$. By applying chain rule to the derivative of F, we obtain

$$\begin{aligned}
\lim_{\epsilon_0\to 0} \frac{\partial F(\epsilon)}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0} = \lim_{\epsilon_0\to 0} \Big[&\sin\big(d\big(X,\, f(\lambda_{f_{\epsilon_0}})\big)\big) \\
&\cdot \cos\big(\theta(\lambda_{f_{\epsilon_0}},\, X)\big) \cdot \Big(\|g(\lambda_{f_{\epsilon_0}})\| + \epsilon_0 \cdot \frac{\partial\, \|g(\lambda_{f_\epsilon})\|}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0}\Big)\Big] \\
&- \lim_{\epsilon_0\to 0}\Big[\sin\big(d\big(X,\, f(\lambda_{f_{\epsilon_0}})\big)\big) \cdot \frac{\partial d\big(X,\, f(\lambda_{f_\epsilon})\big)}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0}\Big].
\end{aligned}$$

In addition,

$$\frac{\partial \left\|g(\lambda_{f_\epsilon})\right\|}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} = \frac{\partial \left\|g(\lambda)\right\|}{\partial \lambda}\Big|_{\lambda=\lambda_{f_{\epsilon_0}}} \frac{\partial \lambda(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0},$$

which exists and does not diverge as $\epsilon_0 \to 0$, since $\|g(\lambda)\| = d\big(f(\lambda),\, (f+g)(\lambda)\big)$ is continuously differentiable for λ and $\frac{\partial \lambda(\epsilon)}{\partial \epsilon}\big|_{\epsilon=0}$ is bounded by Lemma 5. Moreover,

$$\begin{aligned}\lim_{\epsilon_0 \to 0} \frac{\partial d\big(X,\, f(\lambda_{f_\epsilon})\big)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} &= \lim_{\epsilon_0 \to 0} \frac{\partial d\big(X,\, f(\lambda)\big)}{\partial \lambda}\Big|_{\lambda=\lambda_{f_{\epsilon_0}}} \cdot \frac{\partial \lambda(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} \\ &= \frac{\partial d\big(X,\, f(\lambda)\big)}{\partial \lambda}\Big|_{\lambda=\lambda_f} \frac{\partial \lambda(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=0} = 0,\end{aligned}$$

where $\lambda(\epsilon) = \lambda_{f_\epsilon}$. The last equality is done by the definition of λ_f. We therefore get

$$\lim_{\epsilon \to 0} \frac{\partial F(\epsilon)}{\partial \epsilon} = \|g(\lambda_f)\| \cdot \cos\big(\theta(\lambda_f,\, X)\big) \cdot \sin\Big(d\big(X,\, f(\lambda_f)\big)\Big). \tag{3.24}$$

Thirdly, it follows from (3.23) and (3.24) that

$$\begin{aligned}\lim_{\epsilon_0 \to 0} Z_{\epsilon_0}(X) = \lim_{\epsilon_1 \to 0} \Big[&- 2 \arccos F(\epsilon_1) \cdot \frac{1}{\sqrt{1 - F^2(\epsilon_1)}} \cdot \frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1} \Big] \\ = &- 2u\Big(\cos\Big(d\big(X,\, f(\lambda_f)\big)\Big)\Big) \cdot \|g(\lambda_f)\| \cdot \cos\big(\theta(\lambda_f,\, X)\big) \cdot \sin\Big(d\big(X,\, f(\lambda_f)\big)\Big) \\ & \hspace{20em} (3.25) \\ = &- 2d\big(X,\, f(\lambda_f)\big) \cdot \frac{1}{\sin\Big(d\big(X,\, f(\lambda_f)\big)\Big)} \\ & \cdot \|g(\lambda_f)\| \cdot \cos\big(\theta(\lambda_f,\, X)\big) \cdot \sin\Big(d\big(X,\, f(\lambda_f)\big)\Big) \qquad (3.26) \\ = &- 2d\big(X,\, f(\lambda_f)\big) \cdot \|g(\lambda_f)\| \cdot \cos\big(\theta(\lambda_f,\, X)\big), \qquad (3.27)\end{aligned}$$

In the case of $d\big(X,\, f(\lambda_f)\big) = 0$, the same result is obtained since both (3.25) and (3.27) are zero. Thus, by Proposition 2, the equation (3.27) is established for a.e. $X \in B(\zeta)$. Next, we notice that, for a smooth curve f, it can be shown that $M_\lambda := \{x \in S^2 \,|\, \lambda_f(x) = \lambda\}$ is a subset of the great circle perpendicular to f at $f(\lambda)$ by Lemma 1. Let S_λ be the great circle perpendicular to f at $f(\lambda)$. That is, $M_\lambda \subset S_\lambda \cong S^1$. Moreover, a connected proper subset of S_λ is isometric to a line with the same length in $\mathbb{R}$, which makes the intrinsic mean on M_λ feasible. Note that if the length is less than $\pi/2$, the intrinsic mean is unique. Thus, f is an intrinsic-type principal curve of X, by the definition of $\theta(\lambda_f,\, X)$ and $\cos(\pi-\theta) = -\cos(\theta)$, if and only if

$$\mathbb{E}\big[d\big(X,\, f(\lambda_f)\big)\cdot\cos\big(\theta(\lambda_f,\, X)\big) \mid \lambda_f(X)=\lambda\big] = 0, \quad \text{for a.e. } \lambda.$$

By (3.27) and Lebesgue dominated convergence theorem,

$$\begin{aligned}
&\frac{\partial \mathbb{E}_X\big[d^2(X,\, f+\epsilon g)\big]}{\partial \epsilon}\Big|_{\epsilon=0} \\
&= \lim_{\epsilon\to 0}\Big[\frac{\mathbb{E}_X\big[d^2(X,\, f+\epsilon g)\big]-\mathbb{E}_X\big[d^2(X,\, f)\big]}{\epsilon}\Big] \\
&= \mathbb{E}_X\Big[\lim_{\epsilon\to 0}\frac{d^2(X,\, f+\epsilon g)-d^2(X,\, f)}{\epsilon}\Big] \\
&= \mathbb{E}_\lambda\big[\mathbb{E}\big[\lim_{\epsilon\to 0} Z_\epsilon(X) \mid \lambda_f(X)=\lambda\big]\big] \\
&= -2\mathbb{E}_\lambda\big[\mathbb{E}\big[d\big(X,\, f\big(\lambda_f(X)\big)\big)\cdot\big\|g\big(\lambda_f(X)\big)\big\|\cdot\cos\big(\theta(\lambda_f,\, X)\big) \mid \lambda_f(X)=\lambda\big]\big] \\
&= -2\mathbb{E}_\lambda\big[\,\|g(\lambda)\|\cdot\mathbb{E}\big[d\big(X,\, f\big(\lambda_f(X)\big)\big)\cdot\cos\big(\theta(\lambda_f,\, X)\big) \mid \lambda_f(X)=\lambda\big)\big]\big] \\
&= 0.
\end{aligned}$$

Conversely, we assume that

$$\mathbb{E}_\lambda\big[\,\|g(\lambda)\|\cdot\mathbb{E}\big[d\big(X,\, f\big(\lambda_f\big)\big)\cdot\cos\big(\theta(\lambda_f, X)\big) \mid \lambda_f(X)=\lambda\big]\big] = 0,$$

for all $f+g\ (=h)$ such that $\|g\| \neq \pi$ and $\|g'\| \leq 1$. It follows that

$$\mathbb{E}\big[d\big(X,\, f(\lambda_f)\big)\cdot\cos\big(\theta(\lambda_f,\, X)\big) \mid \lambda_f(X)=\lambda\big] = 0, \quad \text{for a.e. } \lambda,$$

which is equivalent to that f is an intrinsic principal curve of X. □

3.5 Concluding remarks

In this chapter, canonical principal curves are proposed for data on spheres. The extrinsic and intrinsic perspectives are considered and the stationarity of the principal curves is investigated, supporting that the proposed methods are a direct generalization of the principal curves of Hastie and Stuetzle (1989) to spheres.

For the data on spheres, both extrinsic and intrinsic approaches yield similar performance. However, it is questionable whether the extrinsic approach on non-homogeneous manifolds will still be efficient. For the spatially non-homogeneous manifolds like a torus, the intrinsic approach may yield better performance due to their inherency. Finally, the principal curve algorithm proposed in this study is a top-down approach. It approximates the structure of data with an initial curve and then gradually improves the estimation. However, for some data structures that are divided into several pieces or containing intersections, the initial estimate (curve) could significantly affect the final estimate. To cope with this limitation, it is worth studying a bottom-up approach. The approach on manifold will be given in Chapter 6.

Chapter 4

Robust spherical principal curves

This chapter is based on a paper which is under revision with *Pattern Recognition*. The chapter is organized as follows. The proposed robust method and its algorithm are provided in Section 4.1. The theoretical property of the proposed method is moreover provided in Section 4.2 and it rigorous proof is presented in Appendix A.2. Numerical experiments, including simulation study and real data analysis, are presented in Section 4.3. Conclusion and future work are finally given in Section 4.4.

4.1 The proposed robust principal curves

In practice, we denote a curve f by a sequence of T points, $f = \{C_1, C_2, \ldots, C_T\}$ joined by geodesic segments, as in Hastie (1984); Hastie and Stuetzle (1989); Hauberg (2016); Lee et al. (2021a). Suppose that we observe a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$. For each point C_t, the weighted barycenter (intrinsic mean) of the dataset can be obtained by the following optimization,

$$m_t(\mathcal{D}, f) = \underset{x \in M}{\arg\min} \sum_{i=1}^{n} w_{t,i} d^2(x, x_i), \quad t = 1, 2, \ldots, T.$$

Then C_t is replaced by $m_t(\mathcal{D},\, f)$ at each iteration in Algorithm 5.

Since the intrinsic mean is obtained by minimizing the sum of the squares of the geodesic distances, the spherical principal curves by such an intrinsic way can be sensitive to outliers. To overcome this problem, we consider the geometric median as an alternative measure for the central tendency of data, rather than intrinsic mean or extrinsic mean. Thus, we transform the expectation step of principal curves into *median step* as follows,

$$m_t(\mathcal{D},\, f) = \operatorname*{arg\,min}_{x \in M} \sum_{i=1}^{n} w_{t,\, i} d(x,\, x_i),\ t = 1,\, \ldots,\, T.$$

It is a sample version of the geometric median, as introduced in (2.6). C_t is then updated by $m_t(\mathcal{D},\, f)$ at each iteration in Algorithm 5. The resulting curves pass through the (geometric) *median* of given data. We now define an L_1-type spherical principal curve, by mimicking the definition of spherical principal curves (3.13), as

$$f(\lambda) = \mathrm{Median}\big[X \mid \lambda_f(X) = \lambda\big] \text{ for a.e. } \lambda, \tag{4.1}$$

where Median denotes the geometric median.

4.1.1 Exact projection step on S^D

Hauberg (2016) performed the projection step approximately, while Lee et al. (2021a) used an accurate projection, resulting in smooth and elaborated curves. In this study, we follow the exact projection step introduced in Lee et al. (2021a). The procedure of projection of x onto the curve f is as follows: (1) Find the projection point of x to each geodesic segment of f, and (2) select the closest projection point. So, it is sufficient to describe how to project a point into a one geodesic segment on S^D. Note that the geodesic distance on spheres is calculated by $d(a,\, b) = \arccos(a \cdot b)$ for $a,\, b \in S^D \subset \mathbb{R}^{D+1}$, where $\cdot$ denotes the dot product in $\mathbb{R}^{D+1}$. For given u, $v,\, w \in S^D \subset \mathbb{R}^{D+1}$, we aim to identify the closest point of w on the geodesic segment joining u and v, say $\overline{uv}$. The point is denoted as $\mathrm{proj}_{\overline{uv}}(w)$. In the case of $u = v$, obviously $\mathrm{proj}_{\overline{uv}}(w) = u = v$. If $u = -v \in \mathbb{R}^{D+1}$, the geodesic segment $\overline{uv}$

is not uniquely defined. Thus, the case $(u \cdot v)^2 \neq 1$ is only considered. Note that if $v \cdot w = w \cdot u = 0$, then a point x on $\overline{uv}$ is expressed by $x = \lambda u + \eta v$ for $\lambda,\, \eta \geq 0$ such that $\lambda^2 + \eta^2 = 1$. In this case, we obtain

$$d(w,\, x) = \arccos(w \cdot x) = \arccos(\lambda w \cdot u + \eta v \cdot w) = \arccos(0) = \pi/2,$$

which means that the geodesic distance between a point on $\overline{uv}$ and w is $\pi/2$. In this case, we can pick $\mathrm{proj}_{\overline{uv}}(w) = \lambda u + \eta v$ for appropriate $\lambda,\, \eta \geq 0$, say, $\lambda = \eta = \sqrt{2}/2$. Hence, we assume that u, v, and w do not hold the equation $v \cdot w = w \cdot u = 0$. Lee et al. (2021a) showed that

$$\mathrm{proj}_{\overline{uv}}(w) = \begin{cases} \mathrm{proj}(w), & \text{if } I \geq 0 \\ \arg\min_{y \in \{u,\, v\}} d(w,\, y), & \text{if } I < 0, \end{cases}$$

where $I = -(u - \mathrm{proj}(w)) \cdot (v - \mathrm{proj}(w))$ and

$$\mathrm{proj}(w) = \frac{(w \cdot u - (u \cdot v)(v \cdot w))u + (v \cdot w - (u \cdot v)(w \cdot u))v}{\|(w \cdot u - (u \cdot v)(v \cdot w))u + (v \cdot w - (u \cdot v)(w \cdot u))v\|}.$$

Then, the distance between w and $\overline{uv}$ can be calculated as

$$d(w,\, \overline{uv}) := d(w,\, \mathrm{proj}_{\overline{uv}}(w)) = \arccos(w \cdot \mathrm{proj}_{\overline{uv}}(w)). \tag{4.2}$$

For detailed description and justification, refer to Lee et al. (2021a,b).

4.1.2 Median step on S^D

The technical details of the median step follow the expectation step of principal curves (Hastie and Stuetzle, 1989; Lee et al., 2021a; Hauberg, 2016); that is, each point of the current curve is updated with the weighted geometric median at each iteration in Algorithm 5. Kernel smoothing can be used to produce stable curves. Suppose that we have n observations $\mathcal{D} = \{x_i\}_{i=1}^n$ and the corresponding projection indices $\{\lambda_f(x_i)\}_{i=1}^n$. For each iteration, the t^{th} point of the principal curve, say C_t, is replaced with the weighted geometric median of neighborhood data. In this chapter, the quartic kernel function $k(\lambda) = (1-\lambda^2)^2 I(|\lambda| \leq 1)$ is applied to obtain the weight of each data point with respect to C_t. The larger the distance between C_t and the projection point of x_i onto f along with the curve f, the smaller weight is given to x_i. The weight of x_i is defined as $w_{t,i} = k(|\lambda_f(C_t) - \lambda_f(x_i)|/q)$. The weighted geometric median of data points,

$$m_t(\mathcal{D},\, f) = \operatorname*{arg\,min}_{x \in S^D} \sum_{i=1}^{n} w_{t,\,i} d(x,\, x_i),\ t = 1,\, 2,\, \ldots,\, T$$

can be obtained by Algorithm 1.

4.1.3 L_1-type principal curves

In this section, we propose an L_1-type robust principal curve and its practical algorithm based on the principle of *self-consistency* that is a fundamental concept in statistics covering EM algorithm (Dempster et al., 1977), K-means clustering, and self-organizing map (Kohonen, 1990), as noted in Flury and Tarpey (1996). The key idea of this principle is to estimate a fixed point of (4.1). To this end, the strategy is to iterate the projection and median steps described in Sections 4.1.1 and 4.1.2 for a candidate curve to satisfy (4.1). Specifically, for an initialized curve f^0, $f^1 = \text{Median}[X \,|\, \lambda_{f^0}(X) = \lambda]$ is obtained through the first iteration of the projection and median steps. Through the second iteration, we obtain $f^2 = \text{Median}[X \,|\, \lambda_{f^1}(X) = \lambda]$. Recursively, for the i–th curve f^i, we obtain $f^{i+1} := \text{Median}[X \,|\, \lambda_{f^i}(X) = \lambda]$. We repeat this procedure for $i \in \mathbb{N}$ until convergence. The converged curve f satisfies $f = \text{Median}[X \,|\, \lambda_f(X) = \lambda]$ for any λ that is the definition of L_1-type principal curve. The proposed algorithm for L_1-type principal curve is as follows:

Let $f = \{C_1,\, C_2,\, \ldots,\, C_T\}$ be a curve, joined by geodesic segments in sequence, and $x \in \mathcal{D}$ be a data point. For each projection step of the above algorithm, we apply (4.2) to each geodesic segment $\overline{C_i C_{i+1}}$ in order to find the exact projection of x onto f. Formally, by the definition of $\lambda_f(x)$, we have $d(x,\, f(\lambda_f(x))) = \min_{1 \le i \le T-1} d(x,\, \overline{C_i C_{i+1}})$. Thus, for each $x \in \mathcal{D}$, the projection of x onto f is obtained by

$$f(\lambda_f(x)) = \text{proj}_{\overline{C_j C_{j+1}}}(x),$$

```
Algorithm 5: $L_1$-type spherical principal curves
- Initialize curve $f = \{C_1, C_2, \ldots, C_T\}$. ;
- Parameterize $f$ with unit interval [0, 1] by some constant speed. ;
- Calculate $\lambda_f(x_i)$ in (3.6) for $1 \le i \le n$. ;
- Calculate errors $\delta(\mathcal{D}, f) = \sum_{i=1}^{n} d(x_i, f(\lambda_f(x_i)))$. ;
while $(\Delta\delta(\mathcal{D}, f) \ge \textit{threshold})$ do
    - (Median step) $C_t \leftarrow m_t(\mathcal{D}, f) = \arg\min_{x \in S^D} \sum_{i=1}^{n} w_{t,i} d(x, x_i)$ for
      $1 \le t \le T$. ;
    - Reparameterize $f$ with unit interval [0, 1] by some constant speed. ;
    - (Projection step) Calculate $\lambda_f(x_i)$ for $1 \le i \le n$. ;
    - Calculate $\delta_1(\mathcal{D}, f) = \sum_{i=1}^{n} d(x_i, f(\lambda_f(x_i)))$. ;
end
```

where $j = \arg\min_{1 \le i \le T-1} d(x, \overline{C_i C_{i+1}})$. In addition, we calculate the $\delta_1(\mathcal{D}, f)$ using (4.2) for each iteration of the above algorithm; that is, for each $x \in \mathcal{D}$, we have

$$d(x, f(\lambda_f(x))) = d(x, \mathrm{proj}_{\overline{C_j C_{j+1}}}(x)) = \arccos(x \cdot \mathrm{proj}_{\overline{C_j C_{j+1}}}(x)).$$

The reconstruction error $\delta_1(\mathcal{D}, f)$ is obtained by the sum over $x \in \mathcal{D}$, which is related to the stopping condition of Algorithm 5. We iterate projection (P) and median (M) steps until the relative change of $\delta_1(\mathcal{D}, f) = \sum_{i=1}^{n} d(x_i, f)$ is below a certain threshold (e.g., 0.01 in our experiments). The relative change of δ_1 is defined as $|\delta_1(\mathcal{D}, f^{i+1}) - \delta_1(\mathcal{D}, f^i)|/\delta_1(\mathcal{D}, f^i)$. Note that $\delta_1(\mathcal{D}, f)$ is a sample version of L_1-type energy functional $\mathbb{E}_X\big[d(X, f)\big]$. This stopping condition comes from the consequence of Theorem 3 in Section 4.2.

We remark that as in most conventional principal curves (Hastie and Stuetzle, 1989; Hauberg, 2016; Kégl et al., 2000; Einbeck et al., 2005; Ozertem and Erdogmus, 2011), the convergence property of the proposed method cannot be guaranteed. The L_1-type principal curve is a fixed point of Algorithm 5, and if Algorithm 5 converges, the resulting curve satisfies the sample version of (4.1). In our experiments given in Section 4.3, the proposed algorithm converges at least empirically. Furthermore, the L_1-type principal curves obtained from Algorithm 5 are resistant to outliers, as shown in Figure 4.1. However, the L_1-type principal curve sometimes produces an unwelcomed non-smooth curve, as, at each median step, the central points are found by the L_1-absolute loss that is singular at zero. It is a well-known property of “median,” reported by several works of literature (Tukey, 1977; Arias-Castro and Donoho, 2009). For example, in the fields of time series analysis, signal, and image processing, the moving average (linear) filter smoothes out the sharp edges of the interpolating curve of data, while the moving median filter preserves the edges, as described in Justusson (1981); Petrus (1999). In the current study, the edge-preserving property of the median would be problematic if it is aimed to find a curve that smoothly represents the dataset. It motivates the Huber-type principal curves. Finally, we initialize a principal curve by a circle that minimizes the sum of squares of the distances from data on S^D. A detailed description of the circle and its algorithm is given in Lee et al. (2020) and Appendix B of Lee et al. (2021b). Several methods are available for initial circles. For more details, refer to Jung et al. (2012) or Section 4.1.2 in Hauberg (2016).

4.1.4 Huber-type principal curves

To define Huber-type principal curves, we consider the Huber loss function (Huber, 2004) defined as

$$\rho(t) = \begin{cases} t^2, & \text{if } |t| \le c \\ c(2|t| - c), & \text{if } |t| > c, \end{cases}$$

where $c > 0$ is a cutoff value. This function can be considered as a mixture of L_1 and L_2 functions since it becomes the L_1 or L_2 functions as c goes to zero or infinity, respectively.

In analogy, a Huber-type criterion on a manifold, $h : M \to \mathbb{R}$, can be defined as $h(x) = \frac{1}{2} \sum_{i=1}^{n} w_i \rho(d(x, x_i))$. The *Huber-type centroid* (Ilea et al., 2016) is defined as a minimizer of h. Namely, the Huber-type centroid is any

$$\arg\min_{m \in M} \sum_{i=1}^{n} w_i \rho(d(m,\, x_i)), \tag{4.3}$$

where w_i denote nonnegative weights of x_i with $\sum_{i=1}^{n} w_i > 0$. Note that the Huber-type centroid goes to the geometric median as $c \to 0$, and goes to the barycenter as $c \to \infty$. To find the Huber-type centroid, we obtain the derivative of h as

$$\nabla h(x) = \sum_{i=1}^{n} \log_x(x_i) I(d(x,\, x_i) \leq c) + c \cdot \frac{\log_x(x_i)}{\| \log_x(x_i) \|} I(d(x,\, x_i) > c) \in T_x M.$$

Using a gradient descent method, we have an algorithm for finding a Huber-type centroid as follows:

Algorithm 6: Huber-type centroid on manifold

```
For a given dataset $\{x_i\}_{i=1}^{n} \in M$ and their nonnegative weights $\{w_i\}_{i=1}^{n}$
  with $\sum_{i=1}^{n} w_i = 1$, set an initial value as $m_1 = x_1$. ;
while ($\Delta m \geq$ threshold) do
    - $\Delta m = \sum_{i=1}^{n} \left[ w_i \log_{m_k}(x_i) I(d(x,\, x_i) \leq c) + \frac{c w_i \log_x x_i}{\|\mathrm{Log}_x(x_i)\|} I(d(x,\, x_i) > c) \right]$ ;
    - $m_{k+1} = \exp_{m_k}(\Delta m)$ ;
end
```

In extensive experiments in Section 4.3, the algorithm converges for $c > 0$. That is, the algorithm at least empirically converges. However, it needs to prove the existence and uniqueness of Huber-type centroid for localized data and the convergence properties of the Algorithm 6, as in Fletcher and Joshi (2007). The algorithm for the Huber-type principal curves is the same as that for the L_1-type principal curves (Algorithm 5), except for the median step and the stopping condition. That is, the geometric median is replaced with the Huber-type measure

$$m_t(\mathcal{D},\, f) = \arg\min_{x} \sum_{i} w_{t,\,i} \rho(d(x,\, x_i)),\ t = 1,\, 2,\, \ldots,\, T.$$

Similarly, the algorithm for estimating the Huber-type principal curves is terminated when the relative change of a reconstruction error, defined as $|\delta_2(\mathcal{D},\, f^{i+1}) -$

$\delta_2(\mathcal{D},\, f^i)|/\delta_2(\mathcal{D},\, f^i)$, is less than a threshold, where $\delta_2(\mathcal{D},\, f) := \sum_{i=1}^{n} \rho(d(x_i,\, f))$ is a sample version of the Huber-type energy functional $\mathbb{E}_X\big[\rho\big(d(X,\, f)\big)\big]$, which is a connection to consequence of Theorem 4 in Section 4.2. Note that, by using the Huber loss, the edge-preserving property of "median", explained in Section 4.1.3, is relieved as illustrated later in Figure 4.1c and Figure 4.1d in Section 4.3. Intuitively, for some constant t with $t \approx 0$ and $0 < t < c$, we have $\rho(t) = t^2 \lll t = L_1(t)$ where $L_1(t) = |t|$. In this respect, under a new observation, the solution of (4.3) tends to less shrink to the location of the observation, compared to that of (2.6). Consequently, the non-singularity of Huber loss at zero usually causes a more softer and stable curves.

4.1.5 Roles and effects of parameters T, q, and c on fitted curves

There are several parameters related to the proposed algorithms. T is the number of points constituting the resulting curve. Through our extensive experiments, T has little effect on the performance of the proposed methods unless T is too small. q plays the same role as the bandwidth does in kernel regression. The roles and effects of T and q on the fitted curves are very similar to the original spherical principal curves Lee et al. (2021a). For the influence of T and q on the fitted principal curves, refer to Section 3.3.2 or Appendix D in Lee et al. (2021b).

The performance of the Huber-type principal curves changes as the predetermined parameter c varies. The L_1-type principal curves induced by $c = 0$ are less affected by outliers, but they are coarser than the ordinary spherical principal curves under the same data structure. As the value of c increases, the Huber-type principal curves become smoother but gradually loosen their robustness. Thus, the cutoff value c controls the trade-off between the smoothness and robustness of the fitted curve. Note that Figure 4.1e and Figure 4.3e in Section 4.3 show the effect of c on the fitted curves. From this perspective, the Huber-type principal curves can be considered as a generalization of the two methods. For the choice of c, we use $c = 0.1$ in our experiments. Alternatively, it can be chosen in an objective way, such as cross-validation. Finally, in the experiments in Section 4.3, the above algorithm converges on spheres for $c > 0$. In other words, the proposed algorithm converges at least empirically. However, it is necessary to study the existence and uniqueness of the Huber-type measure for localized data and the convergence properties of Algorithm 6, as in Fletcher et al. (2009); Afsari (2011). It is left for future research.

4.2 Stationarity of robust spherical principal curves

In this section, we aim to provide the theoretical properties of the proposed methods. It can be shown that $f+\epsilon g : [0,\, 1] \to S^D$ is differentiable for λ. (See Proposition 1 in Section 3.4). Note that the definition of ϵ-perturbation, $f+\epsilon g$, coincides with that of Euclidean case because geodesics on Euclidean space are straight lines. $\|g'\|$ is also defined by mimicking the Euclidean case as shown in Section 3.4. The properties of stationarity of L_1-type and Huber-type principal curves are then as follows.

Theorem 3. *Under $(A1) - (A3)$ with $D = 2$, f is a L_1-type principal curve of X if and only if*

$$\frac{\partial \mathbb{E}_X\big[d(X,\, f+\epsilon g)\big]}{\partial \epsilon}\Big|_{\epsilon=0} = 0, \tag{4.4}$$

for any $f+g$ with $\|g\| < \pi$ and $\|g'\| \le 1$.

Proof. See Appendix A.2. □

Theorem 4. *Under $(A1) - (A3)$ with $D = 2$, f is a Huber-type principal curve of X if and only if*

$$\frac{\partial \mathbb{E}_X\big[\rho\big(d(X,\, f+\epsilon g)\big)\big]}{\partial \epsilon}\Big|_{\epsilon=0} = 0, \tag{4.5}$$

for any $f+g$ with $\|g\| < \pi$ and $\|g'\| \le 1$.

Proof. See Appendix A.2. □

It can be shown that the area of $S^D \setminus B(\zeta) \to 0$ as $\zeta \to 0$. (For a proof, see Lemma 4). Thus, the assumption (A2) in Theorems 3 and 4 is almost negligible for arbitrarily small $\zeta > 0$. Theorems 3 and 4 respectively state that f is a robust principal curve if and only if f is a stationary (critical) point of the energy functionals, $\mathbb{E}_X[d(X,\, f)]$ and $\mathbb{E}_X[\rho\big(d(X,\, f)\big)]$, under the ϵ-perturbation $f + \epsilon g$ for $\epsilon \in [-1,\, 1]$. The stationarity property of the proposed curves on S^D for $D \geq 3$ remains as future work.

The consequences of Theorems 3 and 4 are related to the stopping condition of algorithms for estimating L_1- and Huber-type principal curves. The reconstruction errors $\delta_1(\mathcal{D},\, f) = \sum_{i=1}^{n} d(x_i,\, f)$ and $\delta_2(\mathcal{D},\, f) = \sum_{i=1}^{n} \rho(d(x_i,\, f))$, defined on Sections 4.1.3 and 4.1.4, are sample versions of the energy functionals, $\mathbb{E}_X[d(X,\, f)]$ and $\mathbb{E}_X[\rho(d(X,\, f))]$, respectively. The relative changes of δ_1 and δ_2 approximately correspond to (4.4) and (4.5), respectively. Therefore, the proposed algorithms are terminated when the relative changes of the errors are below a certain threshold to obtain the robust principal curves.

We note that the stationarity of the principal curves in Euclidean space supports that the original principal curve (Hastie, 1984; Hastie and Stuetzle, 1989) is a nonlinear generalization of the principal component. The stationarity of (4.4) and (4.5) along the lines of Hastie (1984); Hastie and Stuetzle (1989); Lee et al. (2021a) justifies that the L_1 and Huber-type principal curves are proper extensions of the principal curves. Moreover, it is noteworthy that the stationarity property of the principal curves on Euclidean space is a path-breaking theorem of the curves, which produces numerous variations of principal curves methods. For example, these include bias-corrected principal curves (Banfield and Raftery, 1992; Tibshirani, 1992; Kégl et al., 2000), probabilistic approaches (Tibshirani, 1992; Stanford and Raftery, 2000; Chang and Ghosh, 2001), principal curves and surfaces generalizing total variance (Delicado, 2001), density estimation approaches (Einbeck et al., 2005; Ozertem and Erdogmus, 2011), curves constructed by bottom-up ways (Kégl et al., 2000; Einbeck et al., 2005; Liu et al., 2017), and curves capable of identifying complex data (Einbeck et al., 2005; Ozertem and Erdogmus, 2011; Wang and Lee, 2008; Zhang et al., 2013).

Before closing this section, we remark that the robust principal curves are defined in a “almost everywhere” sense, instead of “everywhere”. For instance, we define the notion of L_1-type principal curves if f satisfies

$$f(\lambda) = \mathrm{Median}[X \mid \lambda_f(X) = \lambda] \text{ for “}a.e.\text{” } \lambda.$$

The condition of “almost everywhere” is related to our theoretical results. This condition is necessary to make Theorems 3 and 4 “if and only if”. Specifically, if we define the principal curve as $f(\lambda) = \mathrm{Median}[X \mid \lambda_f(X) = \lambda]$ for “any” λ, then the “only if” part is still proved in Theorems 1 and 2, that is, if f is a L_1-type principal curve, then it can be shown that the stationarity,

$$\left.\frac{\partial \mathbb{E}[d(X,\, f + \epsilon g)]}{\partial \epsilon}\right|_{\epsilon=0} = 0. \tag{4.6}$$

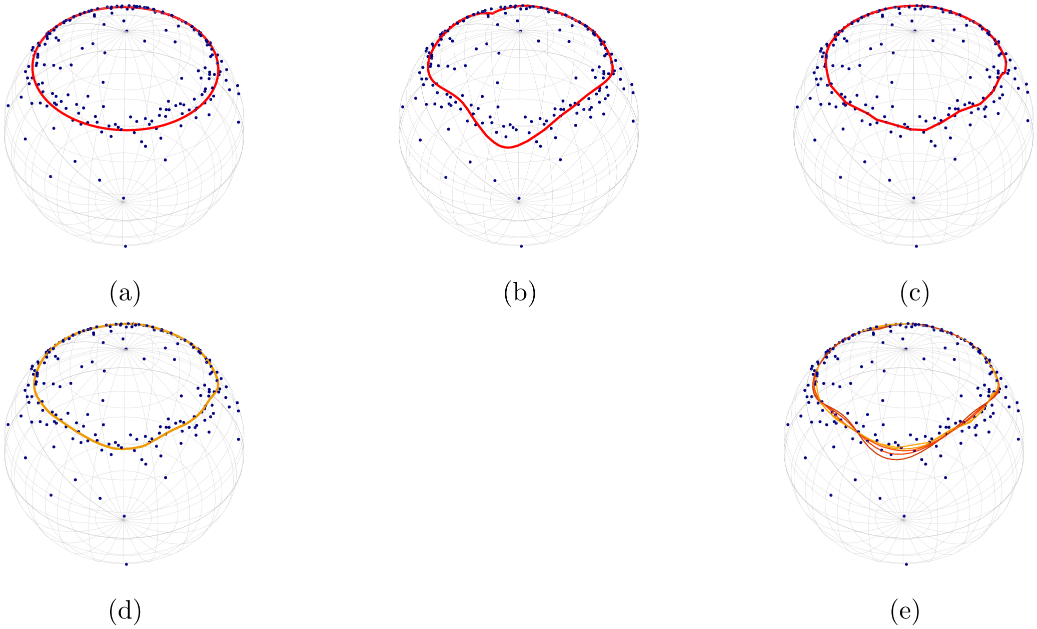
However, the “if” part cannot be proven. Formally, if f satisfies the stationarity of (4.6), then f should be $f(\lambda) = \mathrm{Median}[X \mid \lambda_f(X) = \lambda]$ for “$a.e.$” λ. Therefore, the condition “almost everywhere” cannot be excluded. For the same reason, Hastie Hastie (1984) and Hastie and Stuetzle Hastie and Stuetzle (1989) defined the principal curves in “almost everywhere” way.

4.3 Numerical experiments

4.3.1 Simulation study on S^2

In this section, we conduct a simulation study to evaluate the performances of the proposed robust principal curves. To compare the ordinary spherical principal curves and the proposed methods, we consider two types of data on S^2, circular and waveform data.

We generate a circular dataset on S^2, which is formed of $(r = 1, \theta, \phi)$ with $\phi = \pi/4$, where $0 \le \theta < 2\pi$ and $0 \le \phi < \pi$ denote azimuth and polar angles in the spherical coordinates, respectively. Observations of $n = 200$ data are generated by sampling θ uniformly in $[0, 2\pi)$ and adding random noises from Cauchy(0, 0.05) on ϕ. Figure 4.1 shows a realization (blue dots) of simulated dataset with the true curve (solid red) and the fitting results by the ordinary spherical principal curves (SPC) and the proposed methods (L_1- and Huber-type) with $q = 0.1$ and $T = 100$. As one can see, the proposed curves are less influenced by outliers; thus, they represent the underlying structure efficiently, compared to the conventional spherical principal curve in panel (b) damaged by outliers. The Huber-type principal curve is smoother than the L_1-type curve, as described in Section 4.1.3. Furthermore, from the results in panel (e), we observe that the Huber-type principal curve yields several results over different $c = 0.03,\ 0.1,\ 0.2,\ 0.3,$ and 0.5, which cover L_1 and L_2 fits.



(a) (b) (c) (d) (e)

Figure 4.1: Simulated circular data with outliers (blue) and the resulting curves: (a) true circular curve, (b) extrinsic spherical principal curve, (c) L_1-type principal curve, (d) Huber-type principal curve with $c = 0.1$, and (e) Huber-type principal curves with $c = 0.03,\ 0.1,\ 0.2,\ 0.3$ and 0.5 (from yellow to brown). The proposed methods are implemented with $q = 0.1$ and $T = 100$.

Next, we consider waveform data, as in Lee et al. (2021a); Liu et al. (2017); Panaretos et al. (2014), which are formed of $\phi = \alpha \sin(\theta f) + \frac{\pi}{6}$, where α and f denote amplitude and frequency of the wave, respectively. In the case of $\alpha = 0$, the waveform becomes a circle with radius $1/2$ on the unit 2-sphere. We generate

$n = 200$ data points by sampling θ uniformly in $[0,\, 2\pi)$ and adding random noises from Gaussian mixture $0.7N(0,\, 0.02^2) + 0.3N(0,\, 0.3^2)$ on ϕ. Figure 4.2 shows a realization (blue dots) of simulated dataset with $f = 4$ and $\alpha = 0.2$, the underlying curve marked by solid red, and the fitting results by the ordinary spherical principal curves proposed by (Lee et al., 2020, 2021a) (SPC), principal geodesic analysis (Fletcher et al., 2004) (PGA), principal nested spheres (Jung et al., 2012) (PNS), and the proposed methods (L_1- and Huber-type). For implementation of the proposed methods, we use $q = 0.05$ and $T = 100$. As shown in Figure 4.2, the proposed L_1-type and Huber-type principal curves are capable of identifying the true structure of data well with retaining their resistance to outliers, compared to the others. Note that SPC in extrinsic and intrinsic ways provides similar results, and SPC outperforms the principal curves proposed by Hauberg (2016) in terms of reconstruction error. So, we only consider SPC in the extrinsic way. For more information, see Lee et al. (2021a). To evaluate the performance of each method, the reconstruction error $\sum_{i=1}^{n} d^2\big(x_i,\, \hat{f}(\lambda_{\hat{f}}(\tilde{x_i}))\big)$ is defined as an evaluation measure, where $\{x_i\}_{i=1}^{n}$, $\{\tilde{x}_i\}_{i=1}^{n}$, and $\hat{f}$ denote true points of population curve, noisy data points, and the fitted curve, respectively. We generate $n = 100$ waveform data points with $f = 2, 4$ and $\alpha = 0.2$. Three noise types are considered as $N(0,\, 2\sigma^2)$, contaminated Gaussian mixtures $0.9N(0,\, 0.05^2) + 0.1N(0,\, 16\sigma^2)$, and t-distribution with degrees of freedom three, i.e., $\sigma \cdot t(3)$, where $\sigma = 0.07$ and 0.1. The proposed methods and spherical principal curves (SPC) are implemented with $q = 0.03,\, 0.05,\, 0.07$, $T = 50$, and $c = 0.1$. For each combination of noise type, f, σ and q, we generate waveform data points with size $n = 100$ and then compute the reconstruction error by each method. Over 50 simulation data sets, Table 4.1 lists the average values of the reconstruction errors and their standard deviations. In Table 4.1, we make several observations: (a) SPC and the proposed methods provide similar results for Gaussian noise errors. (b) When random noises occur in contaminated Gaussian mixtures or t-distribution, the proposed L_1-type and Huber-type principal curves outperform others. (c) In all cases, SPC and the proposed methods are superior to PGA and PNS.

Table 4.1: Averages of reconstruction errors and their standard deviations in the parentheses by each method with $T = 50$

Noise type	True form	Method	Noise level					
			$\sigma = 0.07$			$\sigma = 0.1$		
			$q = 0.03$	$q = 0.05$	$q = 0.07$	$q = 0.03$	$q = 0.05$	$q = 0.07$
$N(0,\, 2\sigma^2)$	$f = 2$	SPC (extrinsic)	0.715 (0.135)	**0.484 (0.104)**	**0.390 (0.101)**	1.640 (0.279)	**1.326 (0.278)**	**1.073 (0.269)**
		Proposed (L_1)	**0.690 (0.146)**	0.537 (0.106)	0.445 (0.098)	1.540 (0.277)	1.335 (0.278)	1.136 (0.286)
		Proposed (Huber)	0.693 (0.145)	0.497 (0.101)	0.418 (0.108)	1.605 (0.288)	1.337 (0.284)	1.093 (0.289)
		PGA	9.748 (0.106)			10.347 (0.215)		
		PNS	2.019 (0.048)			2.234 (0.320)		
	$f = 4$	SPC (extrinsic)	**0.689 (0.122)**	**0.627 (0.105)**	0.718 (0.162)	1.655 (0.272)	**1.257 (0.258)**	1.400 (0.285)
		Proposed (L_1)	0.709 (0.132)	0.634 (0.102)	**0.686 (0.147)**	**1.591 (0.284)**	1.308 (0.292)	**1.351 (0.285)**
		Proposed (Huber)	0.691 (0.120)	0.639 (0.103)	0.729 (0.163)	1.615 (0.281)	1.282 (0.269)	1.387 (0.276)
		PGA	14.867 (0.228)			15.530 (0.236)		
		PNS	2.017 (0.079)			2.296 (0.399)		
$0.9N(0,\, 0.05^2) + 0.1N(0,\, 16\sigma^2)$	$f = 2$	SPC (extrinsic)	0.995 (0.541)	0.713 (0.373)	0.598 (0.572)	1.663 (0.794)	2.293 (1.659)	1.460 (0.817)
		Proposed (L_1)	**0.672 (0.414)**	**0.585 (0.367)**	0.462 (0.336)	**1.180 (0.616)**	**1.795 (0.824)**	**0.967 (0.668)**
		Proposed (Huber)	0.836 (0.471)	0.596 (0.360)	**0.460 (0.328)**	1.422 (0.657)	1.829 (0.916)	0.971 (0.677)
		PGA	9.823 (0.338)			10.326 (0.618)		
		PNS	2.374 (0.558)			2.919 (0.887)		
	$f = 4$	SPC (extrinsic)	0.930 (0.399)	0.790 (0.350)	0.902 (0.432)	1.536 (0.645)	1.798 (0.893)	1.527 (0.840)
		Proposed (L_1)	**0.702 (0.325)**	**0.594 (0.265)**	**0.768 (0.403)**	**1.018 (0.578)**	**1.068 (0.469)**	**1.036 (0.556)**
		Proposed (Huber)	0.752 (0.346)	0.641 (0.287)	0.851 (0.403)	1.184 (0.620)	1.167 (0.514)	1.078 (0.530)
		PGA	15.022 (0.439)			15.324 (0.741)		
		PNS	2.298 (0.455)			2.709 (0.953)		
$\sigma \cdot t(3)$	$f = 2$	SPC (extrinsic)	1.195 (0.661)	0.988 (0.580)	0.624 (0.452)	2.081 (0.705)	1.912 (0.856)	1.869 (1.138)
		Proposed (L_1)	**0.935 (0.453)**	0.727 (0.441)	0.579 (0.363)	**1.717 (0.634)**	**1.549 (0.660)**	1.454 (0.607)
		Proposed (Huber)	1.060 (0.536)	**0.666 (0.245)**	**0.563 (0.333)**	1.896 (0.717)	1.622 (0.763)	**1.427 (0.641)**
		PGA	9.977 (0.372)			11.037 (1.486)		
		PNS	2.305 (0.440)			2.895 (0.964)		
	$f = 4$	SPC (extrinsic)	1.125 (0.769)	1.134 (0.881)	1.060 (0.594)	2.814 (1.593)	2.163 (1.028)	2.251 (1.787)
		Proposed (L_1)	**0.898 (0.400)**	**0.809 (0.375)**	**0.925 (0.540)**	**2.411 (1.436)**	**1.690 (0.695)**	**1.650 (0.703)**
		Proposed (Huber)	1.061 (0.774)	0.827 (0.389)	0.993 (0.547)	2.636 (1.592)	1.772 (0.744)	1.673 (0.711)
		PGA	15.019 (0.459)			16.453 (1.571)		
		PNS	2.229 (0.381)			3.099 (0.826)		

4.3.2 Simulation study on S^4

This section conducts a simulation study on S^4. For a point $x = (x_1,\, x_2,\, x_3,\, x_4) \in S^4 \subset \mathbb{R}^5$, it can be parameterized by spherical coordinates, as $x_1 = \cos(\varphi_1)$, $x_2 = \sin(\varphi_1)\cos(\varphi_2)$, $x_3 = \sin(\varphi_1)\sin(\varphi_2)\cos(\varphi_3)$, $x_4 = \sin(\varphi_1)\sin(\varphi_2)\sin(\varphi_3)\cos(\varphi_4)$, and $x_5 = \sin(\varphi_1)\sin(\varphi_2)\sin(\varphi_3)\sin(\varphi_4)$ where φ_1, φ_2, φ_3, and φ_4 are angular coordinates with $\varphi_4 \in [0,\, 2\pi)$ and the others ranging over $[0,\, \pi)$. For our analysis, we consider a simulated waveform dataset on S^4 which is formed of $\phi_1 = \phi_2 = \phi_3 = \alpha \sin(\phi_4 f) + \frac{\pi}{4}$. When $\alpha = 0$, the waveform becomes a one-dimensional circle with radius $\sqrt{2}/4$ on the unit 4-sphere. Two noise types are considered as $N(0,\, \sigma^2)$ and t-distribution with degrees of freedom three, i.e., $\sigma \cdot t(3)$, where $\sigma = 0.03$. We generate a dataset with size $n = 200,\ \alpha = 0.1$ and $f = 3$, by sampling ϕ_4 uniformly in $[0,\, 2\pi)$ and adding random noises from $N(0,\, \sigma^2)$ or $\sigma \cdot t(3)$ on ϕ_1, where $\sigma = 0.03$. Each method with $q = 0.005,\ 0.007,\ 0.02,$ and $T = 100$ is applied to the dataset. Over 50 simulation data sets, Table 4.2 lists the averages of reconstruction errors defined on Section 4.3.1 and their standard deviations. As listed, we observe several features: (a) SPC and the proposed methods provide similar results for Gaussian noise errors. (b) When random noises occur in the t-distribution that is a heavy-tailed distribution, the proposed L_1- and Huber-type principal curves outperform SPC, which implies that the proposed methods are more resistant to outliers. (c) The L_1-type principal curves appear to work slightly better than the Huber-type principal curves for t-distributed noise errors.

Figure 4.3 shows the original motion capture data, the artificially contaminated data, and the fitting results by SPC and the proposed methods. As shown, SPC in Figure 4.3b is significantly distorted after the corruption, while the proposed L_1-type principal curve is less affected by the outliers, but somewhat rougher, as shown in Figure 4.3c. The proposed Huber-type principal curve with $c = 0.1$ in Figure 3(d) appears to be smooth with preserving robustness. In addition, Figure 4.3e shows several fitting results by the Huber-type principal curve according to different $c = 0.03,\ 0.1,\ 0.2,\ 0.3$ and 0.5, ranging over L_1 and SPC fits.

Table 4.2: Averages of reconstruction errors and their standard deviations in the parentheses by each method with $T = 100$ in waveform simulated data on S^4

Noise type	Method	$q = 0.005$	$q = 0.007$	$q = 0.02$
$N(0, \sigma^2)$	SPC (extrinsic)	0.372 (0.297)	0.317 (0.218)	0.401 (0.160)
	Proposed (L_1)	0.382 (0.461)	0.281 (0.259)	**0.439 (0.304)**
	Proposed (Huber)	**0.371 (0.561)**	**0.251 (0.228)**	0.487 (0.386)
$\sigma \cdot t(3)$	SPC (extrinsic)	0.563 (1.072)	0.661 (1.041)	0.766 (1.133)
	Proposed (L_1)	**0.452 (0.657)**	**0.437 (0.571)**	**0.512 (0.751)**
	Proposed (Huber)	0.534 (0.997)	0.526 (1.023)	0.585 (0.542)

4.3.3 Real data analysis: motion capture data

In application, we consider motion capture data of a person walking in a periodic pattern used in Lee et al. (2021a); Mallasto and Feragen (2018); Ionescu et al. (2011, 2014); Hauberg (2016). The dataset represents the direction of the person's left *femur* and thus lie on S^2. This dataset contains 338 observations $\{x_i\}_{i=1}^{338}$ with a circular pattern. To quantify the performance of each method, we artificially contaminate the data by generating 30 outliers from the original data points, as shown in Figure 4.3. The contaminated data are denoted as $\{\tilde{x}_i\}_{i=1}^{338}$. Unlike simulation studies, the true structure of the real data is not available. Therefore, a spherical principal curve with $T = 300$ and $q = 0.1$ is applied to the original (uncontaminated) data as a baseline curve, and the fitted curve in Figure 4.3a can be considered as a pseudo-true structure of the data, denoted by f_0.

To further evaluate the robustness of each method, we consider the following reconstruction error for a fitted curve $\hat{f}$ as

$$\sum_{i=1}^{338} d^2\big(f_0(\lambda_{f_0}(\tilde{x}_i)),\, \hat{f}(\lambda_{\hat{f}}(\tilde{x}_i))\big),$$

which measures how far $\hat{f}$ deviates from the pseudo-true structure f_0 in the presence of outliers. We apply SPC and the proposed methods with $q = 0.1$, 0.12, 0.15, and, 0.17

Table 4.3: Reconstruction errors by each method with $T = 300$ in the contaminated motion capture data

	$q = 0.1$	$q = 0.12$	$q = 0.15$	$q = 0.17$
SPC (extrinsic)	7.990	4.569	3.509	3.409
Proposed (L_1)	0.641	0.583	0.623	0.639
Proposed (Huber)	0.753	0.539	0.567	0.635

to the contaminated data and compute the reconstruction errors. As listed in Table 4.3, the proposed methods are superior to SPC and both proposed methods provide comparable results.

4.4 Summary and future work

In this chapter we have proposed robust principal curves for nonparametric dimensionality reduction on spheres. For this purpose, we have considered L_1-type and Huber-type principal curves. The Huber-type principal curves can be considered as a generalization of the L_1-type principal curves and conventional L_2-type principal curves. Numerical experiments, including simulation studies and real data analysis, demonstrate that the proposed methods are resistant to outliers; thus, they are more capable of extracting true structures of data than other methods. For a theoretical aspect, we have investigated the stationarity of the robust principal curves, which is a theoretical justification that the proposed methods are a proper extension of the conventional principal curves Hastie and Stuetzle (1989); Lee et al. (2021a).

As future work, the proposed L_1- and Huber-type principal curves can be extended into model spaces such as hyperbolic space $\mathbb{H}^D$ and $\mathbb{R}^D$ and other manifolds, including space of symmetric positive definite matrices (SPD) and product space of spheres such as torus or polyspheres (product of spheres), by modifying the projection step of the proposed algorithms. In the empirical studies in Section 4.3, the Huber-type centroid appears to have a mixed property of the intrinsic mean (barycenter) and the geometric median. Therefore, it is necessary to deeply investigate the Huber-type centroid itself, such as the existence, uniqueness, and convergence of related algorithm, as in Yang (2010); Afsari (2011). Finally, the proposed principal curves go through the median of a given data. To explore the hidden structures of the data, it is worth considering the principal curves that pass through quantile, M-quantile, or expectile of a given data. We believe that this extension provides more fruitful information beyond the centrality of the data.

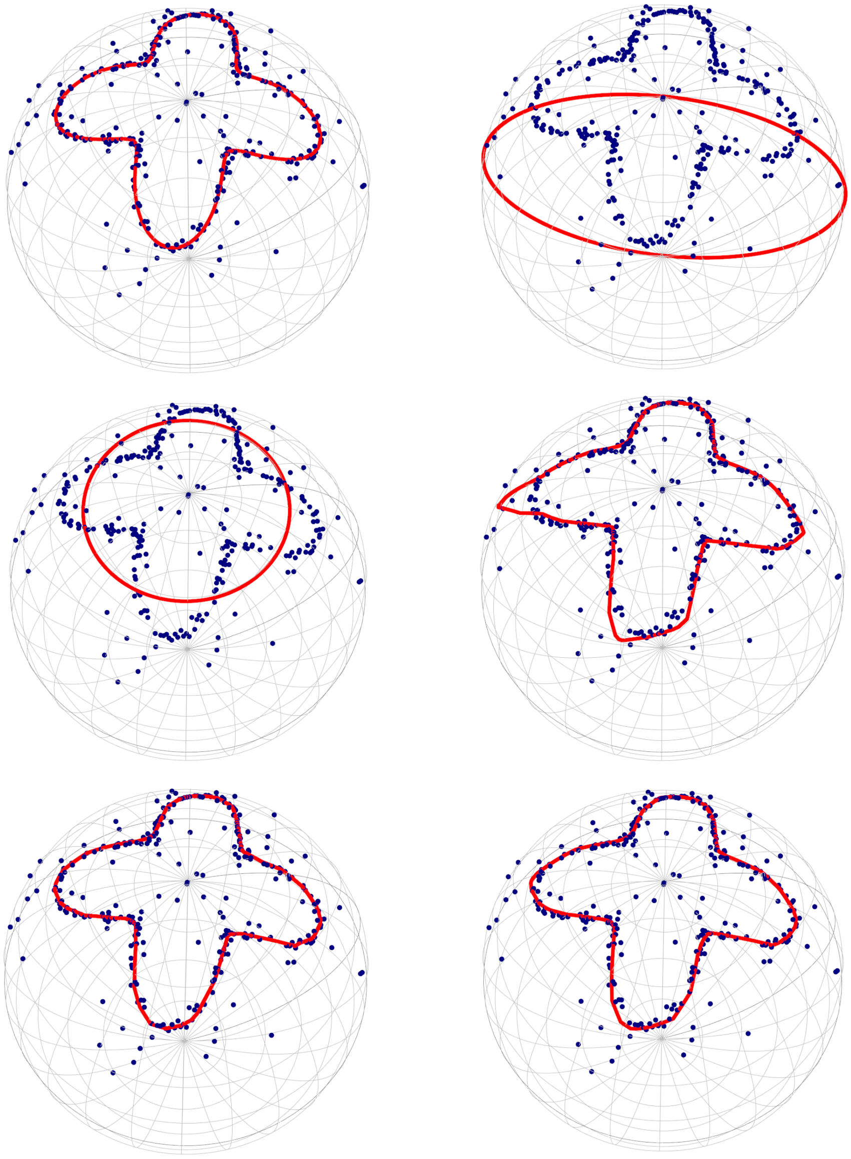
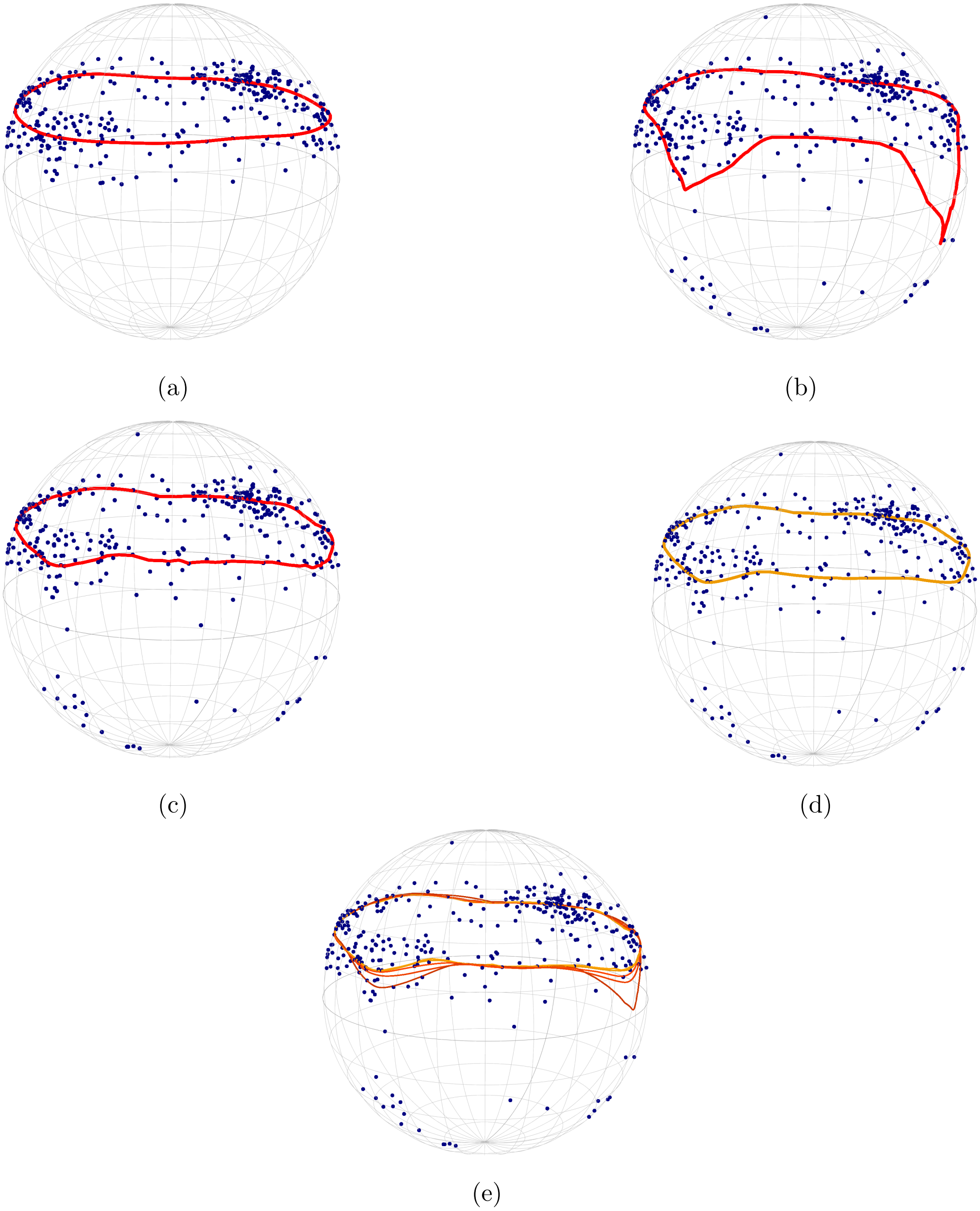


Figure 4.2: From left to right and top to bottom, contaminated waveform data (blue) and population curve (red), principal geodesic analysis, principal nested sphere, spherical principal curve obtained by an extrinsic way, L_1-type principal curve and Huber-type principal curve. The last three methods are implemented with $q = 0.05$ and $T = 100$. Huber-type is additionally implemented with $c = 0.1$.


(a)

(b)

(c)

(d)

(e)

Figure 4.3: (a) Motion capture real data and a pseudo-true curve obtained by spherical principal curve. (b)-(d) Contaminated motion capture data and fitted results by spherical principal curve, the L_1-type principal curve, and the Huber-type principal curve with $c = 0.1$. (e) The fitted results by the Huber-type principal curves with $c = 0.03,\ 0.1,\ 0.2,\ 0.3$ and 0.5 (from yellow to brown). The proposed method are implemented with $q = 0.1$ and $T = 300$.

Chapter 5

spherepc: An R package for dimension reduction on a sphere

This chapter is based on Lee et al. (2022a) which has been published in *R Journal*, **14**(1), 167-181. The purpose of this chapter is to introduce an R package **spherepc** that considers several dimension reduction techniques on a sphere, which encompass recently developed approaches such as SPC and LPG as well as some existing methods, and discuss how to implement these methods through **spherepc**.

In recent times, Lee et al. (2021a) proposed a new method, termed spherical principal curves (SPC), that constructs principal curves, ensuring a stationary property on spheres. SPC is useful for representing circular or waveform data with smaller reconstruction errors than conventional methods, including principal geodesic analysis (Fletcher et al., 2004), exact principal circle (Lee et al., 2021a), and principal curves proposed by Hauberg (2016). However, SPC has the disadvantage of being sensitive to initialization. As a result, there are some data structures that SPC does not apply to, for example, data with spirals, zigzag, or branches like tree-shape. To resolve such a problem, a localized version of SPC, called local principal geodesics (LPG), is being developed. A function for LPG is also provided in the package **spherepc**. Research on the LPG is underway progress.

This chapter is organized as follows. Section 5.1 introduces the existing methods for dimension reduction on the sphere and relevant functions covered in the package **spherepc**, which is available on CRAN. Furthermore, their usages are discussed with examples in detail. Spherical principal curves proposed by Lee et al. (2021a) and principal curves of Hauberg (2016) are briefly described. In addition, implementations of the SPC() and SPC.Hauberg() functions in the **spherepc** are presented. Section 5.3 discusses local principal geodesics (LPG) with implementing it to various simulated data, demonstrating its promising usability. In Section 5.4, all the mentioned methods are performed for analysis of real seismological data. Finally, conclusions are given in last section.

5.1 Existing methods

5.1.1 Principal geodesic analysis

Principal geodesic analysis (PGA) proposed by Fletcher et al. (2004) can be regarded as a generalization of principal component analysis (PCA) to Riemannian manifolds. Fletcher et al. (2004) particularly performed dimension reduction of data on the Cartesian product space of the manifolds. In detail, the data are projected onto the tangent spaces at the intrinsic means of each component of the manifolds; thus, the given data are approximated as points on Euclidean vector space, and subsequently, PCA is applied to the points. As a result, the dimension reduction can be performed through the inverse of the tangent projections. For more details, see Fletcher et al. (2004).

The principal geodesic analysis can be implemented by the PGA() function available in the **spherepc**. The detailed usage of the PGA() function is described as follows.

```
PGA(data, col1 = "blue", col2 = "red")
```

Before using the PGA() function, it requires loading the packages **rgl** (Adler and Murdoch, 2020), **sphereplot** (Robotham, 2013), and **geosphere** (Hijmans et al., 2017). The following codes yield an implementation of the PGA() function.

```
#### for all simulated datasets, longitude and latitude
#### are expressed in degrees
#### example 1: half-great circle data
> circle <- GenerateCircle(c(150, 60), radius = pi/2, T = 1000)
> sigma <- 2                             # noise level
> half.circle <- circle[circle[, 1] < 0, , drop = FALSE]
> half.circle <- half.circle + sigma * rnorm(nrow(half.circle))
> PGA(half.circle)


#### example 2: S-shaped data
# the dataset consists of two parts: lon ~ Uniform[0, 20],
# lat = sqrt(20 * lon - lon^2) + N(0, sigma^2),
# lon ~ Uniform[-20, 0], lat = -sqrt(-20 * lon - lon^2) + N(0, sigma^2)
> n <- 500
> sigma <- 1                             # noise level
> lon <- 60 * runif(n)
> lat <- (60 * lon - lon^2)^(1/2) + sigma * rnorm(n)
> simul.S1 <- cbind(lon, lat)
> lon2 <- -60 * runif(n)
> lat2 <- -(-60 * lon2 - lon2^2)^(1/2) + sigma * rnorm(n)
> simul.S2 <- cbind(lon2, lat2)
> simul.S <- rbind(simul.S1, simul.S2)
> PGA(simul.S)
```

Because a principal geodesic always is a great circle, the PGA() function is suitable for identifying the global trend of data. The implementations to half-circle and S-shaped data are displayed in Figure 5.1, where the principal geodesic properly extracts the global trends in the half-great circle and S-shaped data, while it cannot identify the circular variations in the S-shaped case. In addition, the arguments and outputs of the PGA() function are described in Tables 5.1 and 5.2.

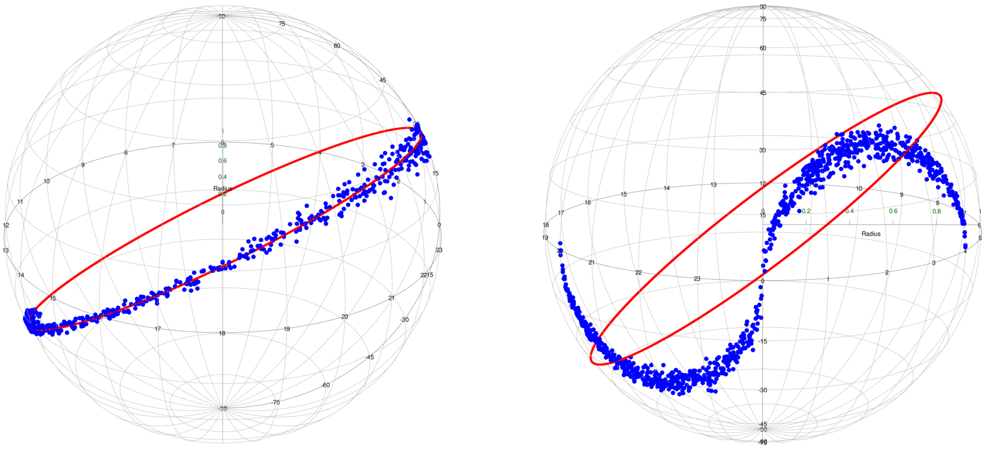


Figure 5.1: From left to right, half-great circle and S-shaped data (blue) and the results (red) of principal geodesic analysis (PGA). The principal geodesic detects the global trends of the noisy half-great circle and the S-shaped data but cannot identify the circular variation of the S-shaped data.

Argument	Description
data	matrix or data frame consisting of spatial locations with two columns. Each row represents longitude and latitude (denoted by degrees).
col1	color of data. The default is blue.
col2	color of the principal geodesic line. The default is red.

Table 5.1: Arguments of the PGA()

Output	Description
plot	plotting of the result in 3D graphics.
line	spatial locations (longitude and latitude by degrees) of points in the principal geodesic line.

Table 5.2: Outputs of the PGA()

5.1.2 Principal circle

In a spherical surface, as shown in Figure 5.1, the principal geodesic analysis always results in a great circle, which cannot be sufficient to identify the non-geodesic structure of data. The circle on a sphere that minimizes a reconstruction error is called principal circle, where the reconstruction error is defined as the total sum of squares of geodesic distances between the circle and data points. However, the existing method for generating the principal circle is still based on the tangent space approximation and its inverse process, thereby leading to numerical errors. Lee et al. (2021a) have proposed an exact principal circle in an intrinsic way and its practical algorithm based on gradient descent. The details are described in Section 3 of Lee et al. (2020) and Appendix B of Lee et al. (2021b). The **spherepc** package provides the PrincipalCircle() function to implement the intrinsic principal circle. Its usage is followed by

```
PrincipalCircle(data, step.size = 1e-3, thres = 1e-5, maxit = 10000).
```

Argument	Description
data	matrix or data frame consisting of spatial locations (longitude and latitude denoted by degrees) with two columns.
step.size	step size of gradient descent algorithm. For convergence of the algorithm, step.size is recommended to be below 0.01. The default is 1e-3.
thres	threshold of the stopping condition. The default is 1e-5.
maxit	maximum number of iterations. The default is 10000.

Table 5.3: Arguments of the PrincipalCircle()

The arguments of the PrincipalCircle() are described in Table 5.3, and its output is a three-dimensional vector, where the first and second components are longitude and latitude (represented by degrees) respectively. The last one is the radius of the principal circle. To display the circle, the GenerateCircle() function should be implemented. Its usage is followed by

```
GenerateCircle(center, radius, T = 1000).
```

The output of the GenerateCircle() function is a matrix consisting of spatial locations (longitude and latitude by degrees) with two columns, which can be plotted by the sphereplot::rgl.sphgrid() and sphereplot::rgl.sphpoints() functions from the **sphereplot** package (Robotham, 2013). Note that the **sphereplot** package depends on the **rgl** package (Adler and Murdoch, 2020). The detailed arguments of the GenerateCircle() function are described in Table 5.4.

Argument	Description
center	center of circle with spatial locations (longitude and latitude denoted by degrees).
radius	radius of circle. It should be range from 0 to π.
T	the number of points that make up a circle. The points in a circle are equally spaced. The default is 1000.

Table 5.4: Arguments of the GenerateCircle()

The following codes implement principal circles by using the PrincipalCircle() and GenerateCircle() functions.

```
#### for all the following examples, longitude and latitude
#### are denoted by degrees
#### example 1: half-great circle data
> circle <- GenerateCircle(c(150, 60), radius = pi/2, T = 1000)
> half.great.circle <- circle[circle[, 1] < 0, , drop = FALSE]
> sigma <- 2                              # noise level
> half.great.circle <- half.great.circle
>                     + sigma * rnorm(nrow(half.great.circle))
```

```
## find a principal circle
> PC <- PrincipalCircle(half.great.circle)
> result <- GenerateCircle(PC[1:2], PC[3], T = 1000)
## plot the half-great circle data and principal circle
> sphereplot::rgl.sphgrid(col.lat = "black", col.long = "black")
> sphereplot::rgl.sphpoints(half.great.circle, radius = 1, col = "blue"
>                                                            , size = 9)
> sphereplot::rgl.sphpoints(result, radius = 1, col = "red", size = 6)


#### example 2: circular data
> n <- 700                             # the number of samples
> sigma <- 5                           # noise level
> x <- seq(-180, 180, length.out = n)
> y <- 45 + sigma * rnorm(n)
> simul.circle <- cbind(x, y)
## find a principal circle
> PC <- PrincipalCircle(simul.circle)
> result <- GenerateCircle(PC[1:2], PC[3], T = 1000)
## plot the circular data and principal circle
> sphereplot::rgl.sphgrid(col.lat = "black", col.long = "black")
> sphereplot::rgl.sphpoints(simul.circle, radius = 1, col = "blue"
>                                                       , size = 9)
> sphereplot::rgl.sphpoints(result, radius = 1, col = "red", size = 6)
```

The results of principal circle are shown in Figure 5.2. As one can see, the principal circle identifies the circular patterns of the noisy half-great circle and circular dataset well.

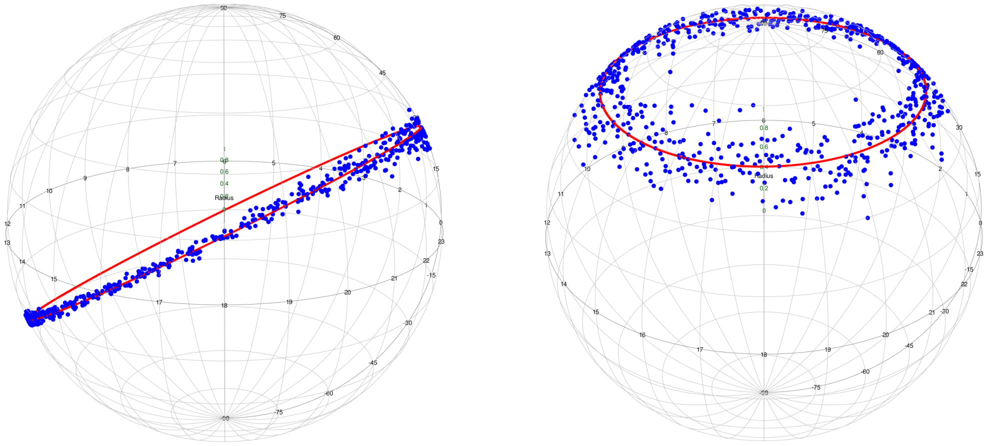


Figure 5.2: Half-great circle data and circular data (blue) and the results (red) of the principal circle from left to right. The principal circle can identify the relatively small circular structure (right) and the great circle structure (left).

5.2 Spherical principal curves

Principal curves proposed by Hastie and Stuetzle (1989) can be considered as a nonlinear generalization of the principal component analysis, in the sense that the principal curves pass through the middle of given data and reserve a stationary property. The curve is a smooth function from a one-dimensional closed interval to a given space; then, a curve f is said to be a *principal curve* of X or self-consistent if the curve satisfies

$$f(\lambda) = \mathbb{E}_X\big[X \mid \lambda_f(X) = \lambda\big],$$

where $f(\lambda_f(x))$ is the closest (projection) point in the curve f from the point x.

Hauberg (2016) provided an algorithm for principal curves on Riemannian manifold. On the other hand, Hauberg (2016) used approximations for finding the closest point of each data point, which may lead to numerical errors. Recently, Lee et al. (2021a) presented theoretical results of principal curves on spheres and a practical algorithm for constructing principal curves without any approximations, called spherical principal curves (SPC), thereby causing the given data to be represented more precisely and smoothly compared to principal curves of Hauberg (2016). In the both ways of extrinsic and intrinsic approaches, the method of SPC updates curves on the spherical surfaces to represent the given data and fits curves that satisfy the stationary conditions. For more details, refer to Lee et al. (2020, 2021a).

The package **spherepc** provides the SPC() function for implementing spherical principal curves and the SPC.Hauberg() function for principal curves of Hauberg (2016). The usage of the SPC() function is as follows.

```
SPC(data, q = 0.05, T = nrow(data), step.size = 1e-3, maxit = 30,
      type = "Intrinsic", thres = 1e-2, deletePoints = FALSE,
      plot.proj = FALSE, kernel = "quartic", col1 = "blue",
      col2 = "green", col3 = "red").
```

The usage of the SPC.Hauberg() function is the same to that of the SPC() function. Before implementing the SPC() and SPC.Hauberg() functions, it requires loading the **rgl** (Adler and Murdoch, 2020), **sphereplot** (Robotham, 2013), and **geosphere** (Hijmans et al., 2017) packages. To implement the SPC() and SPC.Hauberg() functions, we consider the waveform data used in Liu et al. (2017); Lee et al. (2020, 2021a). The generating equation of waveform is

$$\phi = \alpha \cdot \sin(f\theta \cdot \pi/180) + 10,$$

where ϕ, θ, α, and f denote the longitude, latitude in degrees, amplitude and frequency of the waveform, respectively. θ is uniformly sampled from the interval $[-180, 180]$ and a Gaussian random noise from $N(0, \sigma^2)$ is added on each ϕ where $\sigma = 2$, 10. The generating waveform data and implementations of the SPC() and SPC.Hauberg() functions are as follows.

```
#### longitude and latitude are expressed in degrees
#### example: waveform data
> n <- 200
> alpha <- 1/3; freq <- 4        # amplitude and frequency of wave
> sigma1 <- 2; sigma2 <- 10      # noise levels
> lon <- seq(-180, 180, length.out = n) # uniformly sampled longitude
> lat <- alpha * 180/pi * sin(freq * lon * pi/180) + 10 # latitude vector
## add Gaussian noises on the latitude vector
> lat1 <- lat + sigma1 * rnorm(length(lon))
> lat2 <- lat + sigma2 * rnorm(length(lon))
> wave1 <- cbind(lon, lat1); wave2 <- cbind(lon, lat2)
## implement Hauberg's principal curves to the waveform data
> SPC.Hauberg(wave1, q = 0.05)
## implement SPC to the (noisy) waveform data
> SPC(wave1, q = 0.05)
> SPC(wave2, q = 0.05)
```

The above codes generate the results in Figure 5.3. As one can see, the SPC() and SPC.Hauberg() functions identify the waveform pattern of the simulated data. Especially, the SPC() generates a smoother curve. The detailed arguments and outputs of the SPC() are described in Tables 5.5 and 5.6 respectively, which are the same for the SPC.Hauberg().

5.2.1 Options for spherical principal curves

There are some options for the SPC() and SPC.Hauberg() functions. In particular, we implement using the arguments plot.proj and deletePoints, described in Table 5.5. If plot.proj = TRUE is used, then the projection line for each data point is plotted. If the argument deletePoints = TRUE is performed, the SPC() function deletes the points in curves that do not have adjacent data for each expectation step required to fit the principal curves, returning an open curve, i.e., a curve with endpoints. As a result, the principal curves are more parsimonious since a redundant part of the resulting curves is removed. The SPC.Hauberg() function also contains the same options. For implementing these two arguments, the following codes are performed through real earthquake data.

```
> data(Earthquake)
## collect spatial locations
## (longitude and latitude denoted by degrees) of data
> earthquake <- cbind(Earthquake$longitude, Earthquake$latitude)

#### example 1: plot the projection lines (option of plot.proj)
> SPC(earthquake, q = 0.1, plot.proj = TRUE)

#### example 2: open principal curves (option of deletePoints)
> SPC(earthquake, q = 0.04, deletePoints = TRUE)
```

The results are illustrated in Figure 5.4. The left panel shows a closed principal curve (red) with projection lines (black) of each data point onto the curve, and the right panel displays an open principal curve due to the option deletePoints = TRUE. It is a parsimonious result because the redundant part on the upper right side of sphere is removed.

5.3 Local principal geodesics

Suppose that observations have a non-geodesic structure. Then the PGA may not be beneficial to represent such data because PGA always results in a geodesic line. To overcome this problem, we consider performing PGA locally and repeatedly to detect the non-geodesic and complex structures of data, which can be interpreted as a localized version of the PGA. The newly proposed method is called local principal geodesics (LPG). The main idea behind the LPG is that non-geodesic structures can be regarded as a part of geodesic when viewed locally. Although there is no reference to the LPG because research on LPG is underway, there is a localized principal curve method on Euclidean space (Einbeck et al., 2005), which is similar to LPG and may share some motivation with the LPG. For more details, refer to Einbeck et al. (2005).

The package **spherepc** offers the LPG() function, which can recognize the various structures of data such as spirals, zigzag, and tree data. The usage of the function is

```
LPG(data, scale = 0.04, tau = scale/3, nu = 0, maxpt = 500,
    seed = NULL, kernel = "indicator", thres = 1e-4, col1 = "blue",
    col2 = "green", col3 = "red").
```

Like the previous functions, before the LPG() function is implemented, it requires to load the **rgl** (Adler and Murdoch, 2020), **sphereplot** (Robotham, 2013), and **geosphere** (Hijmans et al., 2017) packages. The detailed arguments and outputs of this function are described in Tables 5.7 and 5.8. To apply the LPG() function to the following spiral, zigzag, and tree simulated data illustrated in Figures 5.5, 5.6, and 5.7 respectively, we implement the following codes.

```
## longitude and latitude are expressed in degrees
#### example 1: spiral data
> set.seed(40)
> n <- 900                      # the number of samples
> sigma1 <- 1; sigma2 <- 2.5;   # noise levels
> radius <- 73; slope <- pi/16  # radius and slope of the spiral
## polar coordinate of (longitude, latitude)
> r <- runif(n)^(2/3) * radius; theta <- -slope * r + 3
## transform to (longitude, latitude)
> correction <- (0.5 * r/radius + 0.3)  # correction of noise level
> lon1 <- r * cos(theta) + correction * sigma1 * rnorm(n)
> lat1 <- r * sin(theta) + correction * sigma1 * rnorm(n)
> lon2 <- r * cos(theta) + correction * sigma2 * rnorm(n)
> lat2 <- r * sin(theta) + correction * sigma2 * rnorm(n)
> spiral1 <- cbind(lon1, lat1); spiral2 <- cbind(lon2, lat2)
## plot the spiral data
> rgl.sphgrid(col.lat = ’black’, col.long = ’black’)
> rgl.sphpoints(spiral1, radius = 1, col = 'blue', size = 12)
## implement the LPG to (noisy) spiral data
> LPG(spiral1, scale = 0.06, nu = 0.1, seed = 100)
> LPG(spiral2, scale = 0.12, nu = 0.1, seed = 100)

#### example 2: zigzag data
> set.seed(10)
> n <- 50                  # the number of samples is 6 * n = 300
> sigma1 <- 2; sigma2 <- 5  # noise levels
> x1 <- x2 <- x3 <- x4 <- x5 <- x6 <- runif(n) * 20 - 20
> y1 <- x1 + 20 + sigma1 * rnorm(n); y2 <- -x2 + 20 + sigma1 * rnorm(n)
> y3 <- x3 + 60 + sigma1 * rnorm(n); y4 <- -x4 - 20 + sigma1 * rnorm(n)
> y5 <- x5 - 20 + sigma1 * rnorm(n); y6 <- -x6 - 60 + sigma1 * rnorm(n)
> x <- c(x1, x2, x3, x4, x5, x6); y <- c(y1, y2, y3, y4, y5, y6)
> simul.zigzag1 <- cbind(x, y)
## plot the zigzag data
> sphereplot::rgl.sphgrid(col.lat = 'black', col.long = 'black')
> sphereplot::rgl.sphpoints(simul.zigzag1, radius = 1, col = 'blue'
>                                                         , size = 12)
## implement the LPG to the zigzag data
> LPG(simul.zigzag1, scale = 0.1, nu = 0.1, maxpt = 45, seed = 50)

## noisy zigzag data
> set.seed(10)
> z1 <- z2 <- z3 <- z4 <- z5 <- z6 <- runif(n) * 20 - 20
> w1 <- z1 + 20 + sigma2 * rnorm(n); w2 <- -z2 + 20 + sigma2 * rnorm(n)
> w3 <- z3 + 60 + sigma2 * rnorm(n); w4 <- -z4 - 20 + sigma2 * rnorm(n)
> w5 <- z5 - 20 + sigma2 * rnorm(n); w6 <- -z6 - 60 + sigma2 * rnorm(n)
> z <- c(z1, z2, z3, z4, z5, z6); w <- c(w1, w2, w3, w4, w5, w6)
> simul.zigzag2 <- cbind(z, w)
```

```
## implement the LPG to the noisy zigzag data
> LPG(simul.zigzag2, scale = 0.2, nu = 0.1, maxpt = 18, seed = 20)
```

Note that the LPG() function may return several curves. We now implement the function in a complex simulation dataset composed of several curves. As shown in the left panel of Figure 5.7, the tree object has twenty-six geodesic (linear) structures composed of one stem, five branches, and twenty subbranches. It is not informative to show the generating formula for the tree dataset. Instead, we provide its generating code with explanatory notes as follows.

```
#### example 3: tree dataset
## the tree dataset consists of stem, branches and subbranches
## generate stem
> set.seed(10)
> n1 <- 200; n2 <- 100   # the number of samples in a stem, a branch,
> n3 <- 15               # and a subbrach
> sigma1 <- 0.1; sigma2 <- 0.05; sigma3 <- 0.01  # noise levels
> noise1 <- sigma1 * rnorm(n1); noise2 <- sigma2 * rnorm(n2)
> noise3 <- sigma3 * rnorm(n3)
> l1 <- 70; l2 <- 20;    # length of stem, branches,
> l3 <- 1                # and subbranches
> rep1 <- l1 * runif(n1) # repeated part of stem
> stem <- cbind(0 + noise1, rep1 - 10)
## generate branch
> rep2 <- l2 * runif(n2) # repeated part of branch
> branch1 <- cbind(-rep2, rep2 + 10 + noise2)
> branch2 <- cbind(rep2, rep2 + noise2)
> branch3 <- cbind(rep2, rep2 + 20 + noise2)
> branch4 <- cbind(rep2, rep2 + 40 + noise2)
> branch5 <- cbind(-rep2, rep2 + 30 + noise2)
> branch <- rbind(branch1, branch2, branch3, branch4, branch5)
```

```
## generate subbranches
> rep3 <- l3 * runif(n3) # repeated part in subbranches
> branches1 <- cbind(rep3 - 10, rep3 + 20 + noise3)
> branches2 <- cbind(-rep3 + 10, rep3 + 10 + noise3)
> branches3 <- cbind(rep3 - 14, rep3 + 24 + noise3)
> branches4 <- cbind(-rep3 + 14, rep3 + 14 + noise3)
> branches5 <- cbind(-rep3 - 12, -rep3 + 22 + noise3)
> branches6 <- cbind(rep3 + 12, -rep3 + 12 + noise3)
> branches7 <- cbind(-rep3 - 16, -rep3 + 26 + noise3)
> branches8 <- cbind(rep3 + 16, -rep3 + 16 + noise3)
> branches9 <- cbind(rep3 + 10, -rep3 + 50 + noise3)
> branches10 <- cbind(-rep3 - 10, -rep3 + 40 + noise3)
> branches11 <- cbind(-rep3 + 12, rep3 + 52 + noise3)
> branches12 <- cbind(rep3 - 12, rep3 + 42 + noise3)
> branches13 <- cbind(rep3 + 14, -rep3 + 54 + noise3)
> branches14 <- cbind(-rep3 - 14, -rep3 + 44 + noise3)
> branches15 <- cbind(-rep3 + 16, rep3 + 56 + noise3)
> branches16 <- cbind(rep3 - 16, rep3 + 46 + noise3)
> branches17 <- cbind(-rep3 + 10, rep3 + 30 + noise3)
> branches18 <- cbind(-rep3 + 14, rep3 + 34 + noise3)
> branches19 <- cbind(rep3 + 16, -rep3 + 36 + noise3)
> branches20 <- cbind(rep3 + 12, -rep3 + 32 + noise3)
> sub.branches <- rbind(branches1, branches2, branches3, branches4,
> branches5, branches6, branches7, branches8, branches9, branches10,
> branches11, branches12, branches13, branches14, branches15, branches16,
> branches17, branches18, branches19, branches20)
## tree dataset consists of stem, branch, and subbranches
> tree <- rbind(stem, branch, sub.branches)
## plot the tree dataset
> sphereplot::rgl.sphgrid(col.lat = ’black’, col.long = ’black’)
> sphereplot::rgl.sphpoints(tree, radius = 1, col = 'blue', size = 12)
## implement the LPG function to the tree dataset
> LPG(tree, scale = 0.03, nu = 0.2, seed = 10)
```

As displayed in Figures 5.5, 5.6, and 5.7, the LPG() function identifies the non-geodesic or complex patterns of the simulated datasets well as long as the parameters of scale and ν are properly chosen. The arguments and outputs of the function are respectively described in Tables 5.7 and 5.8.

5.4 Application

In application, we use earthquake data from the U.S. Geological Survey that has collected significant earthquakes (8+ Mb magnitude) around the Pacific Ocean since 1900. As shown in Figure 5.8, the data contain 77 observations distributed in the borders between the Eurasian, Pacific, North American, and Nazca tectonic plates. The data have three features: the observations are distributed globally, scattered, and form non-geodesic structures. Because the tectonic plates are constantly moving towards different directions, identifying the hidden patterns of borders is useful in geostatistics and seismology, as noted in Biau and Fischer (2011); Mardia (2014). It can be possible to identify the borders of plates by applying dimension reduction methods to the earthquake data.

To apply the aforementioned dimension reduction methods to the earthquake data, the code is performed.

```
> data(Earthquake)
## collect spatial locations of data
> earthquake <- cbind(Earthquake$longitude, Earthquake$latitude)

#### example 1: principal geodesic analysis (PGA)
> PGA(earthquake)
```

```
#### example 2: principal circle
## get center and radius of principal circle
> circle <- PrincipalCircle(earthquake)
## generate the principal circle
> PC <- GenerateCircle(circle[1:2], circle[3], T = 1000)
## plot the principal circle
> sphereplot::rgl.sphgrid(col.long = "black", col.lat = "black")
> sphereplot::rgl.sphpoints(earthquake, radius = 1, col = "blue"
>                                                      , size = 12)
> sphereplot::rgl.sphpoints(PC, radius = 1, col = "red", size = 9)
```

Examples 1 and 2 implement the principal geodesic and the principal circle respectively. As illustrated in Figure 5.9, the principal geodesic (left) fails to identify the variations of the earthquake data. The principal circle (right) captures the global trend of the data; whereas the circle could not extract the local variations of the data.

```
#### example 3: spherical principal curves
####            and principal curves of Hauberg
> SPC.Hauberg(earthquake, q = 0.1) # principal curves of Hauberg
> SPC(earthquake, q = 0.1)         # spherical principal curves
```

Example 3 fits the spherical principal curve and Hauberg's principal curve with $q = 0.1$. As shown in Figure 5.10, both methods identify the curved feature of the earthquake data. The spherical principal curve particularly tends to be more continuous than Hauberg's principal curve.

```
#### example 4: spherical principal curves with q = 0.15, 0.1
####                                         , 0.03, and 0.02
> SPC(earthquake, q = 0.15)
> SPC(earthquake, q = 0.1)
> SPC(earthquake, q = 0.03)
> SPC(earthquake, q = 0.02)
```

Example 4 applies the spherical principal curve to the earthquake data with varying $q = 0.15$, 0.1, 0.03, 0.02. The parameter q plays a role of the bandwidth in the SPC() function. As shown in Figure 5.11, the smaller q is, the rougher the curve is. On the contrary, the larger q is, the smoother the curve is.

```
#### example 5: local principal geodesics (LPG)
> LPG(earthquake, scale = 0.5, nu = 0.2, maxpt = 20, seed = 50)
> LPG(earthquake, scale = 0.4, nu = 0.3, maxpt = 22, seed = 50)
```

Lastly, example 5 implements the LPG() function with different scale and nu. As shown in Figure 5.12, the function represents the curved pattern of the data, illustrating the slightly different features.

5.5 Conclusions

In this chapter, existing and newly developed dimension reduction methods on 2-sphere that are covered in **spherepc** R package have been introduced with various simulated examples. It includes not only principal geodesic analysis (PGA), principal circle, and principal curves of Hauberg (2016) as existing methods but also spherical principal curves (SPC) and local principal geodesics (LPG) as new approaches. The **spherepc** package has demonstrated its usefulness by applying the functions to several simulation examples and real earthquake data. We believe that **spherepc** is helpful for applications in various fields, ranging from statistics to engineering such as geostatistics, image analysis, pattern recognition, and machine learning.

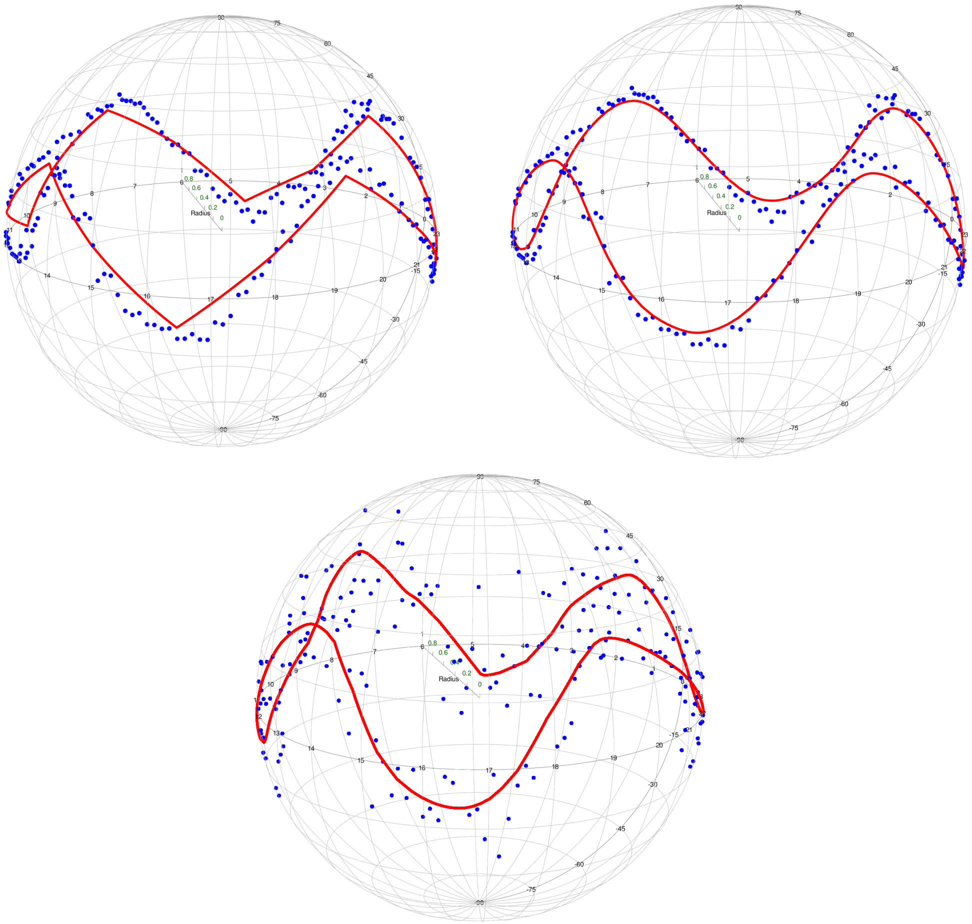


Figure 5.3: Top: The waveform data (blue) and the results (red) of Hauberg's principal curves (left) and spherical principal curves. Bottom: The noisy waveform data (blue) and the result (red) of spherical principal curves. All cases are implemented with $q = 0.05$. The two methods find the true waveform of the data well. In particular, the spherical principal curve tends to be smoother.

Argument	Description
data	matrix or data frame consisting of spatial locations with two columns. Each row represents longitude and latitude (denoted by degrees).
q	numeric value of the smoothing parameter. The parameter plays the same role, as the bandwidth does in kernel regression, in the SPC function. The value should be a numeric value between 0.01 and 0.5. The default is 0.1.
T	the number of points making up the resulting curve. The default is 1000.
step.size	step size of the PrincipalCircle function. The default is 0.001. The resulting principal circle is used as an initialization of the SPC function.
maxit	maximum number of iterations. The default is 30.
type	type of mean on the sphere. The default is "Intrinsic" and the other choice is "Extrinsic".
thres	threshold of the stopping condition. The default is 0.01.
deletePoints	logical value. The argument is an option of whether to delete points or not. If deletePoints is FALSE, this function leaves the points in curves that do not have adjacent data for each expectation step. As a result, the function usually returns a closed curve, i.e. a curve without endpoints. If deletePoints is TRUE, this function deletes the points in curves that do not have adjacent data for each expectation step. As a result, The SPC function usually returns an open curve, i.e. a curve with endpoints. The default is FALSE.
plot.proj	logical value. If the argument is TRUE, the projection line for each data point is plotted. The default is FALSE.
kernel	kind of kernel function. The default is the quartic kernel, and the alternative is indicator or Gaussian.
col1	color of data. The default is blue.
col2	color of points in principal curves. The default is green.
col3	color of resulting principal curves. The default is red.

Table 5.5: Arguments of the SPC()

Output	Description
plot	plotting of the result in 3D graphics.
prin.curves	spatial locations (denoted by degrees) of points in the resulting principal curves.
line	connecting lines between points in prin.curves.
converged	whether or not the algorithm converged.
iteration	the number of iterations of the algorithm.
recon.error	sum of squared distances between the data and their projections.
num.dist.pt	the number of distinct projections.

Table 5.6: Outputs of the SPC()

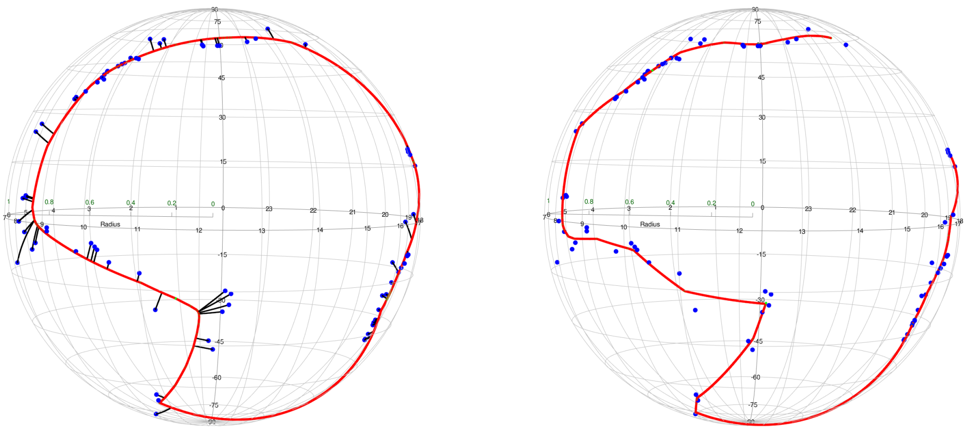


Figure 5.4: Left: Projection result (black) of SPC with $q = 0.1$. The spherical principal curve (red) continuously represents the earthquake data (blue). Right: The open curve of SPC with $q = 0.04$ and deletePoints=TRUE. The less q is, the more the curve tends to overfit the data.

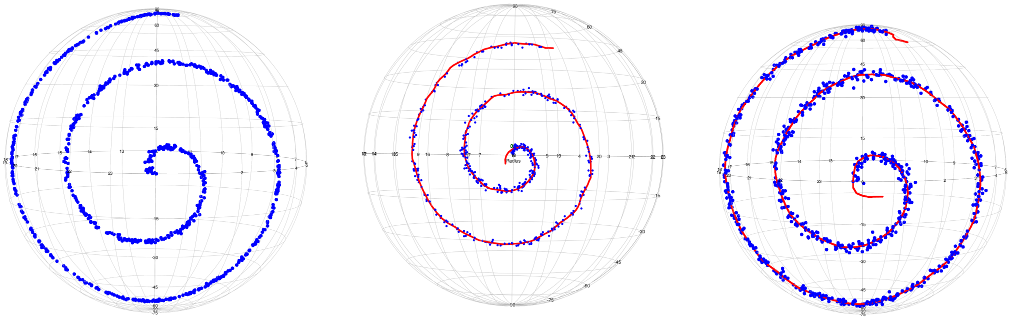


Figure 5.5: Left: Spiral data (blue) and the result (red) of LPG with scale = 0.06 and ν = 0.1. Right: Noisy spiral data (blue) and the result (red) of LPG with scale = 0.12 and ν = 0.1. Local principal geodesics represent the spiral patterns of the (noisy) spiral data well. The larger the noise is, the larger scale is required.

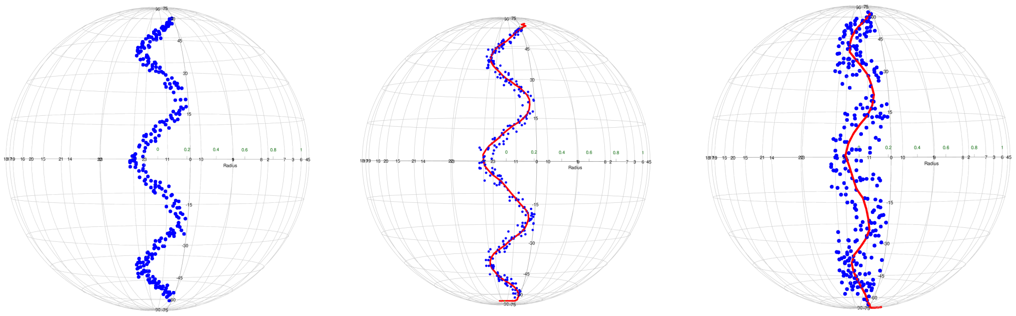


Figure 5.6: Left: zigzag data (blue); Middle: zigzag data (blue) and the result (red) of with scale = 0.1 and ν = 0.1; Right: Noisy zigzag data (blue) and the result (red) of LPG with scale = 0.2, and ν = 0.1; Local principal geodesics extract the zigzag structures of the (noisy) zigzag data properly. The larger the noise is, the larger scale is needed.

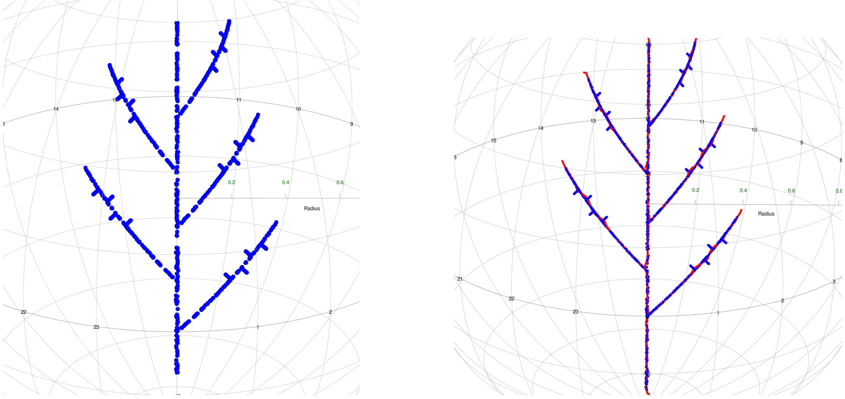


Figure 5.7: Tree data (blue) and the result (red) of LPG with scale = 0.03 and ν = 0.2. The LPG function captures the complex structures of the data well, provided that scale and ν are properly chosen.

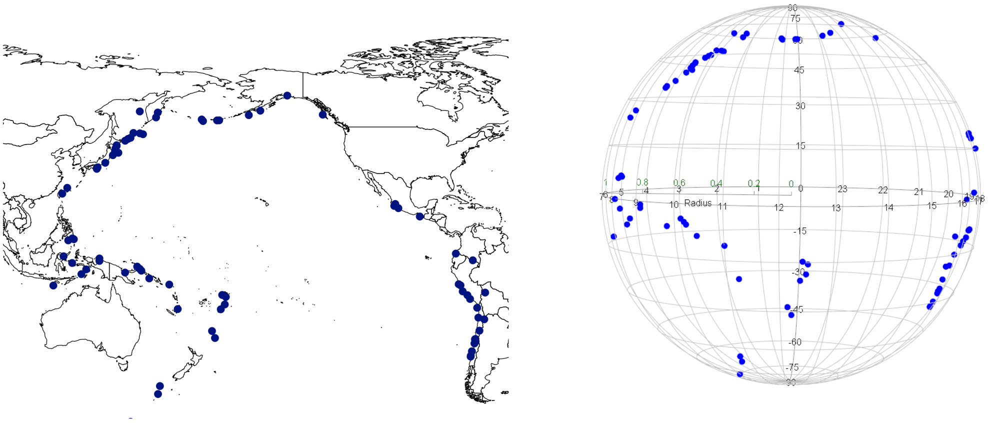


Figure 5.8: Left: The distribution of significant (8+ Mb magnitude) earthquakes (colored in blue); Right: The earthquake is represented in three-dimensional visualization.

Argument	Description
data	matrix or data frame consisting of spatial locations with two columns. Each row represents longitude and latitude (denoted by degrees).
scale	scale parameter for this function. The argument is the degree to which the LPG function expresses data locally; thus, as the scale grows, the result of the LPG becomes similar to that of the PGA function. The default is 0.4.
tau	forwarding or backwarding distance of each step. It is empirically recommended to choose a third of scale, which is the default of this argument.
nu	parameter to alleviate bias of resulting curves. nu represents the viscosity of the given data and it should be selected in [0, 1). The default is zero. When nu is close to 1, the curve usually swirls similarly to the motion of a large viscous fluid. The argument maxpt can control the swirling.
maxpt	maximum number of points in each curve. The default is 500.
seed	random seed number.
kernel	kind of kernel function. The default is the indicator kernel, and the alternative is quartic or Gaussian.
thres	threshold of the stopping condition for the IntrinsicMean function in the process of the LPG function. The default is 1e-4.
col1	color of data. The default is blue.
col2	color of points in the resulting principal curves. The default is green.
col3	color of the resulting curves. The default is red.

Table 5.7: Arguments of the LPG()

Output	Description
plot	plotting of the result in 3D graphics.
num.curves	the number of resulting curves.
prin.curves	spatial locations (represented by degrees) of points in the resulting curves.
line	connecting lines between points in prin.curves.

Table 5.8: Outputs of the LPG()

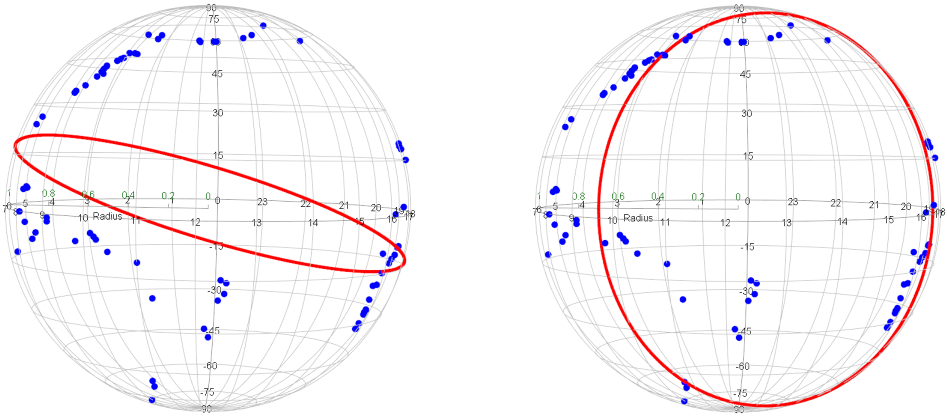


Figure 5.9: Earthquake data (blue) and the results (red) of the principal geodesic analysis and principal circle, from left to right. The principal geodesic fails to find the non-geodesic feature of the data, and the principal circle captures the circular pattern but cannot identify the local variations of the data.

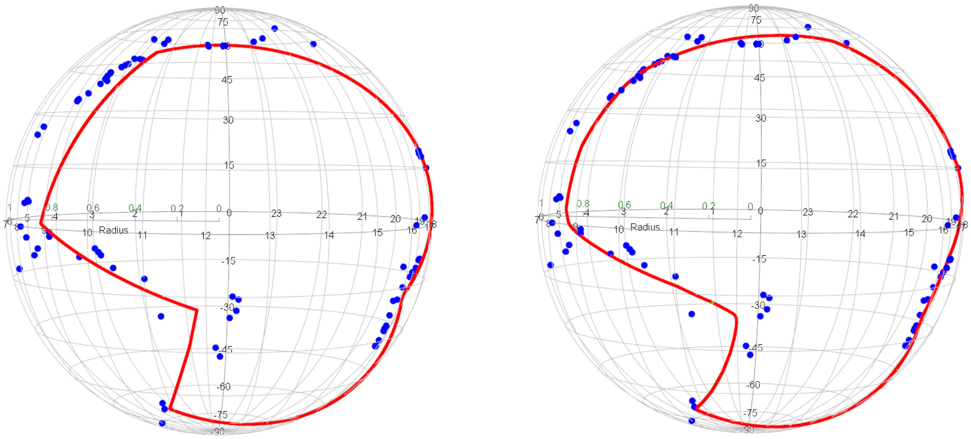


Figure 5.10: Earthquake data (blue) and implementation results (red) with $q = 0.1$ of the SPC.Hauberg and SPC functions respectively, from left to right. Both methods can represent the non-geodesic structure of the earthquake data. The spherical principal curve particularly tend to be smoother.

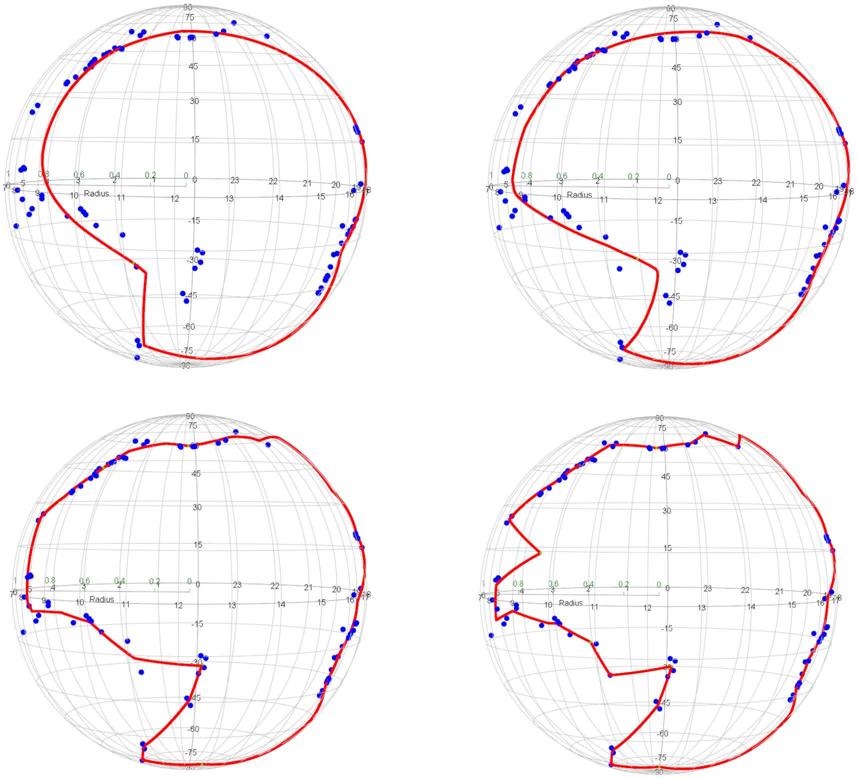


Figure 5.11: From left to right and top to bottom, Earthquake data (blue) and the results (red) of the SPC with $q = 0.15,\ 0.1,\ 0.03$ and 0.02. The larger the parameter q is, the smoother the curve is, while it tends to underfit the data. Conversely, the smaller the parameter q is, the rougher the curve is.

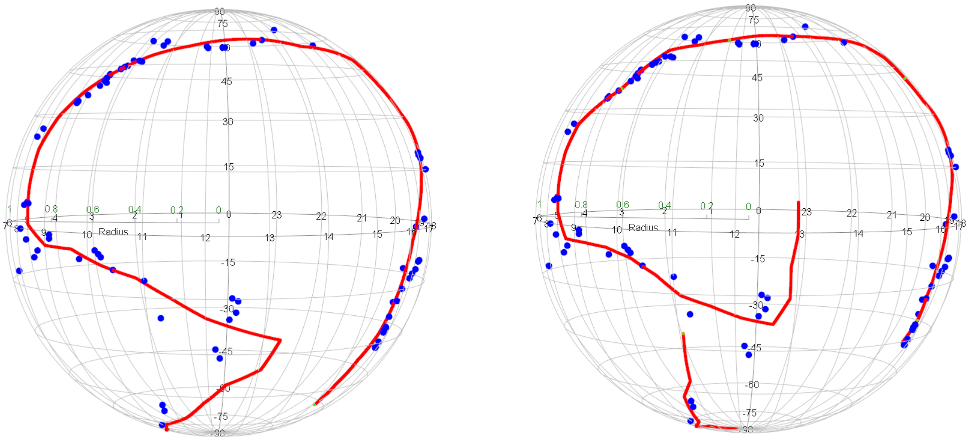


Figure 5.12: From left to right, earthquake data (blue) and the results of the LPG function with scale $= 0.5$, $\nu = 0.2$ and scale $= 0.4$, $\nu = 0.3$ are illustrated. Both the local principal geodesics implemented by different parameters recognize the non-geodesic and scattered pattern of the data, illustrating the different features.

Chapter 6

Local principal curves on Riemannian manifolds

Studies on dimension reduction on manifold domains have drawn attentions over the recent decades. For examples, Dai and Müller (2018) proposed a extrinsic method of functional principal component analysis on manifold, which takes into account on structure of space living in data, with applications to longitudinal compositional data and flight trajectories data. As a followup study, Lin and Yao (2019) have provided a method of functional dimension reduction on manifold in an intrinsic way. Subsequently, Dai et al. (2021) proposed sparse functional principal component modeling for analyzing the brain data which can be treated as functions taking values in manifolds. For examples, principal geodesic analysis (PGA) proposed by Fletcher et al. (2004) is an extension of principal component analysis from Euclidean space to Riemannian manifolds which are, roughly speaking, curved smooth surface equipped with metric tensor and tangent plane for each point of that surface. However, if a set of data is not distributed on a local region, the variational feature of the data is often not identified by geodesics. In regression problems, Fletcher (2013) have proposed geodesic regression on symmetric spaces with applications to corpus callosum shape data.

As related work, Jung et al. (2011) have suggested a dimension reduction method on direct product of manifold choosing geodesic or least square circle in some criterion, as termed principal arc analysis. It is beneficial to the case of manifold whose total intrinsic dimension is high, such as direct product of manifold. Furthermore, Jung et al. (2012) have proposed principal nested sphere which is a dimension reduction for arbitrary dimension of hypersphere. Nevertheless, it seems that two methods may not be effective if underlying distribution of data is not periodic or has crossing structures like T-shape or X-shape data. Banfield and Raftery (1992) modified the ordinary algorithm suggested by Hastie and Stuetzle (1989) to reduce the estimation bias when the curvature of the underlying curve highly varies. Tibshirani (1992) suggested a probabilistic definition of principal curves based on a Gaussian mixture model and applied an EM algorithm for estimation to alleviate bias.

Panaretos et al. (2014) extended a canonical interpretation of principal component analysis from Euclidean space to curved space, such as sphere or cone embedded in $\mathbb{R}^3$. From a mechanics manner, they presented a smooth curve attempting to follow the direction of maximal variation, subject to bounded length. However, because this method starts at center of the data set, it may not work if the center of the data set is far away from the data cloud like circle or C-shape structure. Moreover, it may not be expected that the method captures complex data structures, such as crossing or separated ones. There are several related follow-up studies. For example, Liu et al. (2017) applied a level set-based approach to estimate flexible and robust curves. Yao et al. (2019) relaxed the constraint of boundary conditions imposed on principal flows, and Yao and Zhang (2020) used a principal flow method to deal with a classification problem on manifolds. However, these methods used variational approaches like the Euler-Lagrange equation involved with differential equations on manifolds, making it rather difficult to reproduce the methodologies.

Hauberg (2016) proposed principal curves on Riemannian manifolds which uses nonlinear approach; principal curves, firstly suggested by Hastie and Stuetzle (1989) instead of linear one, PGA. It is more flexible than PGA from two reasons: First, there is no reliance on starting point; mainly, intrinsic mean of the data points. Secondly, it is also able to capture non-geodesic variation of data set. Since the method of Hauberg (2016) depends on an initial estimate of the principal curve, it is effective in representing a simple curve. Unfortunately, it would be failed if underlying structures of data are separated or self-intersecting. Meanwhile, Hauberg (2016) suggested an algorithm of principal curves on Riemannian manifolds. Owing to the large class of the Riemannian manifolds, the principal curves are estimated using approximations. Recently, on spheres S^D for $D \geq 2$, Lee et al. (2021a) presented a newly method, named spherical principal curve (SPC) that constructs a principal curve without any approximations on spheres S^D, resulting in the given data to be represented more precisely and smoothly compared to the method of Hauberg (2016), thereby making more elaborated curves. Moreover, SPC is a direct generalization of original principal curves (Hastie and Stuetzle, 1989) to spheres since SPC ensures the theoretical properties of stationarity on the spheres. However, the above-mentioned methods both are the top-down approach of feature extraction in the sense that this algorithm sets an initial curve and find a proper principal curves adaptive to data set in an iterative manner. The initial curve have to capture the structure of data set to some extent, if not, the consequence is wiggly and eventually fail, as shown in Figure 6.1. Therefore, selecting initial curves properly plays an essential role to estimate a principal curve.

In Euclidean domain there are many variations of principal curves to cope with choosing model complexity, model bias and existence of principal curves. For examples, Kégl et al. (2000) presented a new definition of principal curves which always exist under some conditions. They also proved convergence of optimal principal curves and proposed a practical algorithm, termed as polygonal line algorithm where it is further studied by Biau and Fischer (2011). They selected smoothness parameters such as the number of segments, the length, and the turn of the curve via empirical risk minimization principle.

There are several methodology for finding complex structures which are either crossing on itself or are divided into several pieces (Einbeck et al., 2005; Ozertem and Erdogmus, 2011; Kirov and Slepčev, 2017) in Euclidean space. For examples,

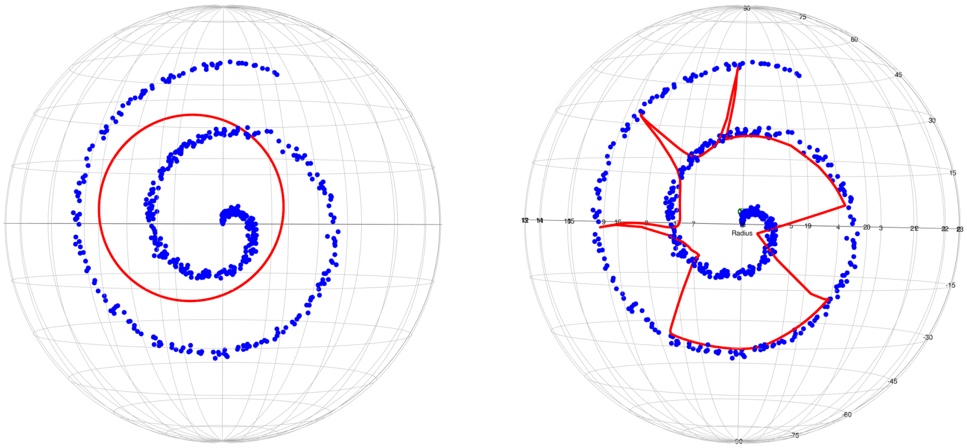


Figure 6.1: Noisy spiral data (blue) and the consequences (red) of principal circle and principal curves (Hauberg, 2016) initialized by the principal circle for $q = 0.07$, from left to right.

Einbeck et al. (2005) proposed an approach, termed local principal curves, which is able to identify crossing or branching curves based on the density estimation like mean-shift algorithm. This method and local principal geodesics have common in that they are bottom up approaches and are able to capture a trend of complicated structures of dataset. On the other hand, the proposed approach differs from that of Einbeck et al. (2005) in that it interpreted the initial curve of main trend of data set as a particle flow of fluid inspired by Panaretos et al. (2014). Specifically, leading eigenvector of sample covariance, analogous to PCA, is the direction of maximal variability which is interpreted in main flow of a particle. Moreover, cohesive force between particles in fluid have to be considered, thus the resultant force of maximal variability and cohesive is the finial direction of the particle for each step of the local principal geodesics. Although the local principal curves identify the pattern of complex structures, it somewhat lacks of theoretical background. On the other hand, our method has a theoretical justification like canonicality which is described and proved in Section 6.2. Moreover, the method can be applied to Riemannian manifolds including Euclidean vector space.

The aim of this chapter is to propose a novel framework for *local principal curves on Riemannian manifolds* (LPCRM) which may be several one-dimensional descriptions of a dataset which is observed from one or several curvilinear structures with noises. The remainder of the chapter constructed as follows. Section 6.1 briefly describes the necessary notions for our method. In Section 6.2, a detailed procedure of methodology are described and canonical property is proved. Section 6.2 also apply the new procedure to various datasets such as spiral, T-shaped, X-shaped data. LPCRM is presented and concrete theories including existence, consistency, and convergence rate, are established by means of empirical risk minimization principle in Section 6.3. Section 6.4 performs a seismological real data analysis. Section 6.5 lastly conclude this chapter with remarks about further work.

6.1 Preliminaries

For each $p \in M$, *exponential map* is a differentiable map from a neighborhood of p in T_pM to M. For a vector v in the neighborhood, the geodesic at p with direction v, $\gamma : [0,\, 1] \to M$, uniquely exists so that γ satisfies that $\gamma(0) = p$, $\gamma^{'}(0) = v$, and $\|\gamma^{'}(t)\| = \|v\|$ for any $t \in [0,\, 1]$. The exponential map at p is defined as

$$\exp_p(v) := \gamma(1) \in M. \tag{6.1}$$

If $(M,\, d)$ is connected and complete as a metric space, then the geodesic continues as much as we want from the Hopf-Rinow theorem (e.g., Theorem 6.13. of Lee (2006)). In other words, the exponential map at p, $\exp_p : T_pM \to M$, is defined on the entire T_pM. For the simplest case, $M = \mathbb{R}^D$, since $T_p\mathbb{R}^D \simeq \mathbb{R}^D$ for any $p \in \mathbb{R}^D$, the exponential and logarithm maps are both identity. In particular, when $M = S^D := \left\{(x_1,\, x_2,\, \ldots x_{D+1}) \in \mathbb{R}^{D+1} \mid \sum_{i=1}^{D+1} x_i^2 = 1\right\}$, naturally embedded into the ambient space $\mathbb{R}^{D+1}$, the exponential map at $p = (0,\, 0,\, \ldots,\, 0,\, 1) \in \mathbb{R}^{D+1}$ can be written as

$$\exp_p(v) = (v_1 \frac{\sin \|v\|}{\|v\|},\, v_2 \frac{\sin \|v\|}{\|v\|},\, \ldots,\, v_D \frac{\sin \|v\|}{\|v\|},\, \cos \|v\|),$$

for any $v \in T_pS^D \simeq \mathbb{R}^D$ with $\|v\| \leq \pi$ in which $\|\cdot\|$ denotes the standard norm in $\mathbb{R}^D$. *logmap* is the inverse map of exponential map. The logmap at p, $\log_p : S^D \to T_pS^D$, is written by

$$\log_p(w) = (w_1 \frac{\theta}{\sin\theta}, \, w_2 \frac{\theta}{\sin\theta}, \, \ldots, \, w_D \frac{\theta}{\sin\theta})$$

for any $w = (w_1, w_2, \ldots, w_{D+1}) \in S^D \setminus (0, 0, \ldots, 0, -1) \subset \mathbb{R}^{D+1}$, where $\theta = \arccos(w_{D+1})$. See Buss and Fillmore (2001) for details. Principal geodesic analysis (PGA) (Fletcher et al., 2004) can be regarded as a generalization of principal component analysis (PCA) to Riemannian manifolds. Fletcher et al. (2004) especially performed dimension reduction of data on the Cartesian product space of the manifolds. In details, the data are projected onto the tangent spaces at the intrinsic means of each component of the manifolds; thus, the given data are approximated as points on Euclidean vector space, and subsequently, PCA is applied to the points. As a result, the dimension reduction can be performed through the inverse of the tangent projections.

The main idea of PGA is that approximating a data set on a Riemmanian manifold to its tangent space at the center of the data set via logarithm map and then applying PCA to the approximated data. However, PGA does often not work when global trend of data set is not captured by a single geodesic. On the other hand, a single geodesic is still a good description for a localized data; thus, PGA could work well in local region, which is the main motivation for the local principal geodesics (LPG). To find the locally maximal direction of variability, locally defined sample covariance should be defined. For a given Riemannian manifold M and a data set $\mathcal{D} = \{x_i\}_{i=1}^n \subset M$, Panaretos et al. (2014) have naturally defined a (sample) local tangent covariance by approximating a set of data on manifold to its tangent space via logarithm map as follows.

Definition 4 (Definition 2.2 in Panaretos et al. (2014)). *A h-scale local tangent covariance at $\overline{x} \in M$ is defined as*

$$\Sigma_h(\overline{x}) = \frac{1}{\sum_i I_h(x_i, \overline{x})} \sum_{i=1}^n (\log_{\overline{x}} x_i) \cdot (\log_{\overline{x}} x_i)^T I_h(x_i, \overline{x}),$$

where $\log_{\overline{x}} x_i$ *is a* D*-dimensional column vector,* $I(x) = 1$ *if* $|x| \leq 1$*, otherwise 0, and* $I_h(x, \overline{x}) = I(h^{-1} \|\log_{\overline{x}} x - \overline{x}\|)$ *for* $h > 0$*.*

A h-scale local tangent covariance at $\overline{x}$ is a sample covariance of data points of which distance from $\overline{x}$ is less than h. The h-scale local tangent covariance at $\overline{x}$ is the sample covariance w nearby data points . We define a h-radius closed ball of x as $B_h(x) := \{y \in M \,|\, d(x, y) \leq h\}$. Without loss of generality we may assume that $B_h(x) \cap \mathcal{D} = \{x_1, x_2, \ldots, x_m\}$ are a set of points whose distances from $x \in M$ are equal to or less than h. We further suppose that the intrinsic mean of them are $\overline{x} \in M$ and $B_h(\overline{x}) \cap \mathcal{D} = \{x_1, x_2, \ldots, x_m\}$ for $m \geq 1$ and $h > 0$. Then a scale h tangent covariance at x is given by

$$\begin{aligned} \Sigma(x) = \frac{1}{m}\sum_{i=1}^{m}(\log_x x_i) \cdot (\log_x x_i)^T \;&\approx\; \frac{1}{m}\sum_{i=1}^{m}(\log_{\overline{x}} x_i - \log_{\overline{x}} x) \cdot (\log_{\overline{x}} x_i - \log_{\overline{x}} x)^T \\ &\approx\; \frac{1}{m}\sum_{i=1}^{m}(\log_{\overline{x}} x_i) \cdot (\log_{\overline{x}} x_i)^T + (\log_{\overline{x}} x) \cdot (\log_{\overline{x}} x)^T \\ &=\; \Sigma(\overline{x}) + (\log_{\overline{x}} x) \cdot (\log_{\overline{x}} x)^T. \end{aligned}$$

Thus, $\Sigma(\overline{x})$ is approximately calculated by $\Sigma(x) - (\log_{\overline{x}} x) \cdot (\log_{\overline{x}} x)^T$. Note that if $M = \mathbb{R}^D$ then the above approximations are replaced by equals, since the logmap is just subtraction; that is, $\log_x y = y - x$ for any $x, y \in \mathbb{R}^D$, and the intrinsic mean $\overline{x}$ becomes the center of gravity of $\{x_i\}_{i=1}^m$, $\frac{1}{m}\sum_{i=1}^m x_i$.

6.2 Local principal geodesics

In this section, local principal geodesics (LPG) is introduced. LPG is a one-dimensional description of data set lying on Riemannian manifolds, and can be an initial estimate of local principal curves. The basic idea for LPG is that the method of principal geodesic is applied and proceeded in local region alternatively. For a given Riemannian manifold M with dimension D, scale parameter $h > 0$, step size $\tau > 0$ and a data set $\mathcal{D} = \{x_i\}_{i=1}^n \subset M$. Suppose that the data set $\mathcal{D}$ is collected from one or several curvilinear structures with noises.

(1) Set $c_0 = x_1 \in \mathcal{D}$ randomly and $c_1 = \text{argmin}_{x \in M} \sum_{x_i \in B_h(f_0)} d^2(x_i, x)$ (local centering), where $B_x(h) = \{y \in M \,|\, d(x, y) < h\}$ for all $x \in M$ and c_1 is the intrinsic mean of nearby h-neighborhood points from a randomly chosen initial point $c_0 \in \mathcal{D}$.

For extracting various structures of data automatically, an initial point c_0 can be randomly chosen from the data set. However, if f_0 is selected outside of data cloud, then the curves may go outward, resulting in meaningless components. Therefore, to obtain stable consequence of the curves, select a starting point as c_1, the local center of nearby points of c_0. Although it is more likely to be inside the cloud of data than c_0, the local center c_1 may not be actual center of nearby points of itself if c_0 is chosen outside the cloud. To cope with this problem, one can select a starting point after exploring the data, or perform local centering several times as follows, $c_{j+1} = \text{argmin}_{x \in M} \sum_{x_i \in B_{c_j}(h)} d^2(x_i, x)$ for $j \geq 0$, until its change below some threshold and let f_1 be the limit of this sequence.

(2) (Forward direction step) Find a h-scale local tangent covariance $D \times D$ matrix at c_1, Σ_h and a unit eigenvector $v_1 \in \mathbb{R}^D$ corresponding to the largest eigenvalue of Σ_h. Forward in that direction by τ; that is, set $c_2 = \exp_{c_1}(\tau v_1)$.
Repeat previous procedure using center c_{i+1} instead of c_i for $i \geq 1$. Recursively, select a direction v_{i+1} so that $v_{i+1} \cdot v_i \geq 0$ and define $f_{i+1} = \exp_{c_i}(\tau v_i)$ until $B_h(f_m) = \phi$ for some $m_1 \geq 1$. Note that, for smoothness of curve, forward direction is chosen so that the angle from the previous direction is less than $\pi/2$ for each step of (2).

(3) (Backward direction step) Define $c_{-1} = c_1$ and direction $w_1 = -v_1$. Inductively, find a h-scale local tangent covariance $D \times D$ matrix at c_{-i}, Σ_h, and a unit eigenvector $w_i \in \mathbb{R}^D$ corresponding to the largest eigenvalue of Σ_h. for each step, backward in that direction by τ, that is, $c_{-i-1} = \exp_{c_{-i}}(\tau w_i)$ recursively. Similarly in (2), for each step w_{i+1} is selected, by satisfying $w_{i+1} \cdot w_i \geq 0$ until $B_h(c_{-m_2}) = \phi$

for some $m_2 \geq 1$. In the same way to (2), backward direction is also chosen so that the angle from the previous direction is less than $\pi/2$ for each step of (3). From (2) and (3), the set of points $\mathcal{C}_1 := (c_{m_1}, c_{m_1-1}, \ldots, c_1 = c_{-1}, c_{-2}, c_{-3}, \ldots, c_{-m_2})$ is a principal geodesic curve.

(4) Let $\mathcal{D}_1 := \{x_i \in \mathcal{D} \,|\, d(x_i, \mathcal{C}_1) < h\}$. For the data set $\mathcal{D} \setminus \mathcal{D}_1 \subset M$ repeat (1), (2) and (3) to find local principal geodesic segments $\mathcal{C}_2$ and $\mathcal{D}_2 := \{x_i \in \mathcal{D} \,|\, d(x_i, \mathcal{C}_2) < h\}$. Similarly, for data set $\mathcal{D} \setminus (\mathcal{D}_1 \cup \mathcal{D}_2)$ find local principal geodesics $\mathcal{C}_3$ and $\mathcal{D}_3 := \{x_i \in \mathcal{D} \,|\, d(x_i, \mathcal{C}_3) < h\}$. In the same way, for the data set $\mathcal{D} \setminus (\bigcup_{j=1}^{i} \mathcal{D}_j)$ find $\mathcal{C}_{i+1}$ and $\mathcal{D}_{i+1}$ iteratively, until $\mathcal{D} = \bigcup_{i=1}^{s} \mathcal{D}_s$ for some $s \geq 1$. Then, local principal geodesic curves $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s$ and corresponding neighbor data set $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_s$ are the result of local principal geodesics (LPG).

LPG has an effect of clustering in that the data set D divided into $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_s\}$ which are chosen as maximally connected components via LPG, and its corresponding of one-dimensional descriptions $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s\}$. Although, strictly speaking, it is not a clustering because $\mathcal{D}_i's$ are not disjoint, it is the reason why LPG works for crossing curvilinear structures.

6.2.1 Bias relaxation

Here we improve a bias problem of LPG. Since the principal component analysis is sensitive to centering, it is performed at the mean of the data points by centering data matrix. However, when seeking a direction of maximal variability at each stage of the local principal geodesics, the leading eigenvector of the covariance matrix is a distorted result if the starting point is usually not placed on the neighborhood data points. For this reason, once the local principal geodesic curve is out of the vicinity of the data cloud, it will proceed in a more distorted direction, as the distance between the starting point and data points surrounding it has increased. Eventually, if the local principal geodesic curve deviated from the data cloud, then it does not return to the distribution of the data, instead go outwardly, causing a negative impact on the feature extraction of the data set. To cope with this bias, the following improved local principal geodesics, the cohesive-maximum variation force local principal geodesics, is considered. Inspired by Panaretos et al. (2014), specifically, data points are considered fluid and local principal geodesic curve is the trajectory in which one fluid particle travels. Then, the force at which the overall flow of the fluid acts on the particle is a maximal variability force. Furthermore, we are able to consider a cohesive force, the gravitational pull between the particles. If there is no inter-fluid cohesion, it will not return when the fluid overflows from the waterways, but will flow outward only. This is similar to the phenomenon of the local principal geodesics without returning from the data cloud. Taking into account the cohesion between the fluid particles, therefore, the direction of the two resultant force is the one in which the particles finally move for each step. Thus, in forward and backward direction step the followings below should be added.

(2)' (forward direction step)

For i-th step, define a cohesion vector at f_i as the normalized vector of $\sum_{j=1}^{n}(\log_{c_i} x_j) I_h(c_i,\, x_j)$, and new direction $v_i^{'}$ as $\frac{\nu \cdot cohesion + v_i}{\|\nu \cdot cohesion + v_i\|}$ where ν is a predetermined viscosity $\forall i = 1,\, 2,\, \ldots,\, s$.

(3)' (backward direction step)

In the same way, define a cohesion vector at f_i as the normalized vector of $\sum_{j=1}^{n}(\log_{c_-i} x_j) I_h(c_{-i},\, x_j)$ and new direction $w_i^{'}$ as $\frac{\nu \cdot cohesion + w_i}{\|\nu \cdot cohesion + w_i\|}$ $\forall i = 1,\, 2,\, \ldots,\, s$.

The procedure of LPG with bias relaxation is illustrated in Figure 6.2. In Euclidean space, the first principal component line is an one-dimensional reduction of a data set maximizing the variability of it. Panaretos et al. (2014) have verified that the first principal flows in Euclidean space is the first principal component and k-th order principal flow in Euclidean space is the k-th principal component on local region around the center of data; that is, it is an *canonical* extension of principal component analysis. In the same way, it can be shown that the proposed method is also a *canonical* extension of the first principal component from Euclidean space to

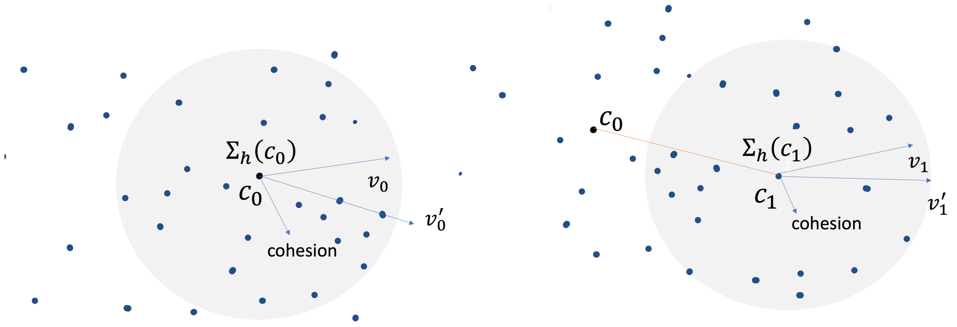


Figure 6.2: Left: LPG starts at a point c_0 and $\Sigma_h(c_0)$ is calculated in the h-neighborhood (gray shade). It forward to the direction v_0' that is the resultant direction between v_0 and cohesion. Right: LPG moves forward by $\tau v_0'$ from c_0 and then the next direction is calculated from $\Sigma_h(c_1)$ and cohesion. The procedure proceeds in the same way; that is, for $i = 1, 2, 3, \ldots, c_{i+1}$ is found from c_i.

Riemannian manifolds.

Theorem 5. *The local principal geodesics becomes to be the first principal geodesic when scale parameter $h = \infty$, step size $\tau > 0$, and viscosity $\nu = 0$. In particular, LPG becomes to be first principal component line when $M = \mathbb{R}^D$.*

Proof. See Appendix A.3. □

As Theorem 5 states, LPG becomes to be the first principal geodesic when $h = \tau = \infty$, which means that LPG is the non-geodesic generalization of first principal geodesic analysis. Particularly, when $M = \mathbb{R}^D$, Theorem 5 implies that LPG is a canonical extension of PCA from Euclidean space to Riemannian manifold and is a justification that LPG proceeds in the largest eigenvector of tangent covariance matrix for each forward and backward direction step.

6.2.2 Overlap of curves and merging

In the process of local principal geodesics, if the data sets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_s\}$ are disjoint then curves are separated and there is no overlap. If not, overlap of curves may take place. If much energy is focused on an intersection of curves, then the local principal geodesic curves have a common region to some degree. Nonetheless, the share may not be a big problem. In such a case, various choice could be considered, for examples, getting rid of the redundant or express the overlap as the mean of it.

6.2.3 Consideration of parameters

To implement locally defined principal curves, parameters which have to be chosen are scale h, step size τ and viscosity ν. Choosing such parameters is very important. Inadequate parameters have a negative effect on the result of local principal geodesics. For example, step size τ should be set smaller than h, if not, local principal geodesics go outwardly from the data cloud, which result in failure of identifying the features. Generally, the larger the noise level of the data, the greater h is required. At the extremes, the different values of scale h show the various features of data. Specifically, if $h \approx 0$ (e.g., $h < \min_{i,j} d(x_i, x_j)$), the number of the connected component of result, s, becomes to n while the large h (e.g., $h, \tau > \max_{i,j} d(x_i, x_j)$) causes $s = 1$. In the case that we know the number of clusters s_0 a priori, we can choose the scale h by gradually decreasing it from large to zero until $s = s_0$. We suppose that $s = s_0$ when $h \in [h_s, h_{s+1}]$. Under the interval $[h_s, h_{s+1}]$, subsequently h is selected so that the goodness of fit measure, defined as $\sum_{i=1}^{n} d^2(x_i, \hat{f})$, is minimized at $h \in [h_s, h_{s+1}]$, where $\hat{f}$ is the fitted local principal geodesics. In our experiments, we have chosen h so that $s \leq 5$. Moreover, the curve may be circulated when step size τ is too small compared to the scale h, or viscosity ν is very large. Throughout experiments, one third of h is recommended for τ empirically. That is, we set $\tau = h/3$ and $\nu = 0.1$.

6.2.4 Connection with existing methods

Principal flows

From the perspective for *principal flows*, a local principal geodesic is an integral curve with respect to given covariance vector field W; that is, $f'(p)$ compatible to $W(p)$ for each p in the local principal geodesic curve. Thus, both LPG and principal flows maximally attempt to follow the vector field W. Although principal flows are constrained in length limit for regularization, LPG has no upper bound for its length so as to capture various patterns of data. When it comes to the regularization issue, it can be relieved via principal curves algorithm, which is described in Section 6.3.

Local principal curves on Euclidean space

The method proposed by (Einbeck et al., 2005), termed as local principal curves, is a method that can be used to detect the several curvilinear structures of data on Euclidean space, along the line of Delicado (2001). The method and LPG both use iterative algorithm to identify the complicated features of data. The local principal curve (Einbeck et al., 2005) is based on the density estimation of underlying distribution, while LPG depends on covariance of given data. The method of Einbeck et al. (2005) can be used only on $M = \mathbb{R}^D$ while LPG can be applied to generic Riemannian manifolds. Moreover, as Einbeck et al. (2005) stated, the method somewhat little lacks of theoretical basis, while LPG is supported by a related justification (Theorem 5 in this chapter).

6.2.5 Results of local principal geodesics

In this section, a variety of simulation data sets having curvilinear structures, such as zigzag, spiral used in Kégl et al. (2000), T-shaped, X-shaped, doubly circular used in Liu et al. (2017), and spherical helix data, are considered. The simulation sets are involved in the example codes of R package **spherepc** (Lee et al., 2022a). For each data set, we the proposed method is applied. Figure 6.3 shows the LPG results of various synthetic datasets.

6.3 Local principal curves

Even if the improved local principal geodesics already captures features of the data set, due to the tangent approximation errors, it is slightly not the curves that go through the middle of the data set. For regularization of the LPG, method of principal curves is needed. To this end, we adopt the procedure proposed by Kégl (1999); Kégl et al. (2000). The procedure and its algorithm (polygonal line algorithm) are conceptually suitable for our method.

We assume M to be generic Rimannian manifold. Specifically, M is complete and connected as a metric space, which makes the Riemannian distance, $d(\cdot,\, \cdot) : M \times M \to [0,\, \infty)$, feasible. For any continuous curve $f : [0,\, 1] \to M$ and a point $x \in M$, $d(x,\, f) := \inf_{\lambda \in [0,\, 1]} d(X,\, f(\lambda))$. For an M-valued random variable X, let $d(X,\, f)$ be the random distance from X to f. We denote $d(x,\, f) = \inf_{\lambda \in [0,\, 1]} d(x,\, f(\lambda))$. Then *risk* of f is defined as

$$R(f) := \mathbb{E} d^2(X,\, f) = \mathbb{E}\big[\inf_{\lambda \in [0,\, 1]} d^2(X,\, f(\lambda))\big] = \mathbb{E}\big[d^2(X,\, f(\lambda_f(X)))\big], \qquad (6.2)$$

where $\lambda_f(x) := \min\big\{\lambda \in [0,\, 1] \,|\, d(x,\, f(\lambda)) = \inf_{\mu \in [0,\, 1]} d(x,\, f(\mu))\big\}$. Note that $d(X,\, f)$ and $R(f)$ are invariant under parametrizations of f. That is, for any surjective monotone map $m : [0,\, 1] \to [0,\, 1]$,

$$d(x,\, f) = d(x,\, f \circ m), \quad \forall x \in M \;\text{ and }\; R(f) = R(f \circ m), \qquad (6.3)$$

where $\circ$ denotes the composition of function. In other words, the distance a point and a function f does not depends on any parametrization of f but on the graph of f, $G_f := \{f(\lambda) \,|\, \lambda \in [0,\, 1]\} \subset M$. We now denote the collection of all continuous functions from $[0,\, 1]$ to M by $C([0,\, 1],\, M)$ and define that

$$\begin{aligned}
\mathcal{G} &= \{f \in C([0,\, 1],\; M) \,|\, d(f(\lambda_1),\, f(\lambda_2)) \le \ell|\lambda_1 - \lambda_2| \text{ for any } \lambda_1,\, \lambda_2 \in [0,\, 1]\} \\
\mathcal{G}_0 &= \{f \in C([0,\, 1],\; M) \,|\, L(f) \le \ell\}
\end{aligned}$$

for some constant $\ell \ge 0$, where $L(f)$ denotes the length of f and $\mathcal{G}_0$ is thus the collection of functions in $C([0,\, 1],\, M)$ whose lengths are not greater than ℓ. If $f \in \mathcal{G}$, then

$$L(f) = \sup_{\mathcal{P}} \sum_{i=0}^{m-1} d(f(\lambda_i),\, f(\lambda_{i+1})) \leq \ell \sum_{i=0}^{m-1} (\lambda_i - \lambda_{i+1}) = \ell, \tag{6.4}$$

where $\mathcal{P} = \{0 = \lambda_0 < \lambda_1 < \ldots < \lambda_{m-1} < \lambda_m = 1\}$ with $m \geq 1$ is a partition of $[0,\, 1]$ and the above supremum is taken over all partitions of $[0,\, 1]$. That is, $f \in \mathcal{G}_0$ and then $\mathcal{G}_0 \subset \mathcal{G}_0$. Conversely, for a $f \in C([0,\, 1],\, M)$ with $L(f) \leq \ell$, from (6.3) we may assume that, without loss of generality, f is parametrized with the unit interval $[0,\, 1]$ by a constant speed $s = L(f) \leq \ell$. Then $d(f(\lambda_1),\, f(\lambda_2)) = s|\lambda_1 - \lambda_2| \leq \ell|\lambda_1 - \lambda_2|$, thereby implying that the reparametrization of f belongs to $\mathcal{G}$. More rigorously, the quotient space $\mathcal{G}_0/\sim$ is same to $\mathcal{G}$, in which $\sim$ denotes the equivalent relation such that $f \sim g$ if and only if there is a monotone map $m : [0,\, 1] \to [0,\, 1]$ such that $g = f \circ m$. It means that, by the reparametrization, $\mathcal{G}$ is nearly same to $\mathcal{G}_0$ as far as the risk is concerned from (6.3). Therefore it is suffices to consider the $\mathcal{G}$ if we inspect $\mathcal{G}_0$. For more details, see Kégl (1999) or Appendix A in Kégl et al. (2000).

Definition 5. *For an M-valued random variable X and some constant $\ell \geq 0$, $f^* \in C([0,\, 1],\, M)$ is said to be a principal curve of X with length ℓ, if f^* satisfies*

$$R(f^*) = R^* := \inf_{f \in C([0,\, 1],\, M)} \{R(f) \,|\, L(f) \leq \ell\} \;\; \textit{and } L(f^*) \leq \ell \tag{6.5}$$

Before investigating the existence of a principal curve, the following lemma needs to be proved.

Lemma 6. *$\mathcal{G}$ is a closed set in $C([0,\, 1],\, M)$ with respect to the uniform distance*

$$\textit{dist}(f,\, g) := \sup_{\lambda \in [0,\, 1]} d(f(\lambda),\, g(\lambda)) \quad \textit{for any } f,\, g \in C([0,\, 1],\, M). \tag{6.6}$$

Proof of Lemma 6. Suppose that $\{f_n\}_{n \in \mathbb{N}} \subset \mathcal{G} \subset C([0,\, 1],\, M)$ converges to f, that is, $\text{dist}(f_n,\, f) \to 0$ as $n \to \infty$. It follows that $f \in C([0,\, 1],\, M)$ owing to the fact that the limit (with the uniform distance) of continuous functions is also continuous. Since the uniform convergence implies the point convergence, for any $\lambda_1,\, \lambda_2 \in [0,\, 1]$

$$\ell|\lambda_2 - \lambda_1| \geq \lim_{n \to \infty} d(f_n(\lambda_1),\, f_n(\lambda_2)) = d(f(\lambda_1),\, f(\lambda_2)).$$

Therefore $f \in \mathcal{G}$, which completes the proof. □

We now say that an M-valued random variable X has second moment if $\mathbb{E}d^2(X, p) < \infty$ for some point $p \in M$. The point $p \in M$ serves as an origin point in M. (When $M = \mathbb{R}^D$, p is the origin). It is easy to show that $\mathbb{E}d^2(X, p) < \infty$ is equivalent to $\mathbb{E}d^2(X, p_0) < \infty$ for any point $p_0 \in M$ due to triangle inequality and Jensen's inequality. The following theorem shows that, for any given X with second moment and nonnegative constant ℓ, there exists a principal curve with length ℓ.

Theorem 6 (Existence of principal curves)**.** *Assume that* $\mathbb{E}d^2(X, p) < \infty$ *for some point* $p \in M$*. For any* $\ell \geq 0$*, there exists a principal curve of* X *with length* ℓ*, say* $f^* \in C([0, 1], M)$*. Namely,* f^* *satisfies* (6.5)*.*

Proof of Theorem 6. Denote $R^* = \inf_{f \in C([0,1], M)} \{R(f) \,|\, L(f) \leq \ell\}$ for simplicity. By the definition, it follows that $R^* \leq \mathbb{E}d^2(X, p)$ by considering the constant function $f \equiv p$. In case $R^* = \mathbb{E}d^2(X, p)$, the trivial curve $f \equiv p$ is the principal curve with length ℓ we wish to find and then the proof ends. For this reason, we consider the case that $R^* < \mathbb{E}d^2(X, p)$. Let P be a probability distribution of X on M. Denote $\Delta := [R^* + \mathbb{E}d^2(X, p)]/2 < \mathbb{E}d^2(X, p)$ and a closed ball $\bar{B}_p(r) = \{x \in M \,|\, d(x, p) \leq r\}$. Since $\mathbb{E}d^2(X, p) = \lim_{r \to \infty} \int_{\bar{B}_p(r)} d^2(x, p) P(dx)$, some constant $r_0 > 0$ can be chosen so that

$$\int_{\bar{B}_p(r_0)} d^2(x, p) P(dx) \geq \Delta.$$

Now set $r_1 = \max(r_0, \ell)$. We first aim to prove that a candidate for principal curve with length ℓ is totally contained in $\bar{B}_p(3r_1)$, which in turn means that it is enough to only consider curves contained in the closed ball. Suppose that f is not fully contained in the $\bar{B}_p(3r_1)$, meaning that $G_f := \{f(\lambda) \,|\, \lambda \in [0, 1]\} \not\subset \bar{B}_p(3r_1)$ where G_f denotes the graph of f. It is easy to show that for any $x \in \bar{B}_p(r_1)$ $d(x, p) \leq r_1 \leq d(x, f)$ from $L(f) \leq \ell \leq r_1$ and $\bar{B}_p(2r_1) \cap G_f = \phi$. Thus,

$$\begin{aligned} R^* < \Delta &\leq \int_{\bar{B}_p(r_1)} d^2(x, f) dP(x) \\ &\leq \int_M d^2(x, f) dP(x) = R(f), \end{aligned}$$

which implies that

$$\begin{aligned} R^* &= \inf_{f\in C([0,\,1],\,M)} \{R(f) \,|\, L(f) \le \ell\} \\ &= \inf_{f\in C([0,\,1],\,M)} \{R(f) \,|\, L(f) \le \ell,\ G_f \subset \bar{B}_p(3r_1) =: N\} \\ &= \inf_{f\in C([0,\,1],\,N)} \{R(f) \,|\, L(f) \le \ell\}, \end{aligned}$$

where $N = \bar{B}_p(3r_1)$ is the closed ball of center p with radius $3r_1$ $(=: r)$ Secondly, it is now enough to consider N, instead of M. Note that N is compact by Hopf-Rinow theorem because N is also connected and complete. Let $\mathcal{G} \subset C([0,\,1],\,N)$ be the collection of all continuous functions from [0, 1] to N satisfying

$$d(f(\lambda_1),\, f(\lambda_2)) \le \ell|\lambda_1 - \lambda_2| \ \text{ for any } \lambda_1,\, \lambda_2 \in [0,\,1]. \tag{6.7}$$

As previously mentioned, it is suffices to consider the $\mathcal{G}$ if we inspect all the continuous functions $f : [0,\,1] \to N$ with $L(f) \le \ell$ by reparametrization and (6.3) as far as risk is concerned. Now, a generalization of Arzelà-Ascoli theorem (Chapter 7 in Kelly (1991)) is needed to end this proof. The generalization of Arzelà-Ascoli theorem states that $\mathcal{G} \subset C([0,\,1],\,N)$ is compact if and only if (i) $\mathcal{G}$ is closed, (ii) equi-continuous, and (iii) for each $\lambda \in [0,\,1]$ $\mathcal{G}_\lambda := \{f(\lambda) \,|\, f \in \mathcal{F}\} \subset N$ is relatively compact, meaning that the closure of $\mathcal{G}_\lambda$ in N is compact. The condition (i) follows by Lemma 6 and the condition (ii) follows by (6.7). Owing to the facts that N is compact (by Hopf-Rinow theorem) and that a closed subset of a compact set is also compact, the closure of $\mathcal{G}_\lambda \subset N$ is compact for each $\lambda \in [0,\,1]$. Consequently, $\mathcal{G}_\lambda$ is relatively compact and thus the condition (iii) follows.

According to the generalization of Arzelà-Ascoli theorem, $\mathcal{G}$ is compact in $C([0,\,1],\,N)$ with respect to the uniform distance $\mathrm{dist}(f,\,g) := \sup_{\lambda\in[0,\,1]} d(f(\lambda),\,g(\lambda))$ for $f,\,g \in \mathcal{G}$. Denote $R^* = \inf_{f\in C([0,\,1],\,N)} \{R(f) \,|\, l(f) \le \ell\}$ for simplicity. There exists a sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{G}$ satisfying $R(f_n) < R^* + 1/n$. Apparently, $G_{f_n} \subset N = B_p(r)$ from the definition of $\mathcal{G}$. Note that usual notion of compactness is equivalent to that of sequential compactness on metric space by a well-known fact of point-set topology. Because $\mathcal{G}$ is sequentially compact, there is a subsequence $\{f_{n_k}\}_{k\in\mathbb{N}} \subset \{f_n\}_{n\in\mathbb{N}}$

and a limit $f^* \in \mathcal{G}$ so that $\mathrm{dist}(f_{n_k},\, f^*) \to 0$ as $k \to \infty$. At that time, the limit f^* is a candidate for a principal curve of X with length ℓ. The condition of length, $l(f^*) \le \ell$, is obtained by (6.4). The remaining part is to prove that f^* achieves the infimum in (6.5). We now denote $\{f_{n_k}\}_{k\in\mathbb{N}}$ by $\{f_n\}_{n\in\mathbb{N}}$ for simplicity. By definition, $\lim_{n\to\infty} \mathrm{dist}(f_n,\, f^*) = 0$ and $\lim_{n\to\infty} R(f_n) = R^*$. For a point $x \in M$, suppose that $d^2(x,\, f^*) \le d^2(x,\, f_n)$. Note that $\{f_n\}_{n\in\mathbb{N}}$ and f^* are contained in the closed ball $N = \bar{B}_p(r)$. By triangle inequality,

$$
\begin{aligned}
|d^2(x,\, f^*) - d^2(x,\, f_n)| &= d^2(x,\, f_n) - d^2(x,\, f^*) \\
&\le d^2(x,\, f_n(\lambda_{f^*}(x))) - d^2(x,\, f^*(\lambda_{f^*}(x))) \\
&\le \big[d(x,\, f_n(\lambda_{f^*}(x))) + d(x,\, f^*(\lambda_{f^*}(x)))\big] \cdot d(f_n(\lambda_{f^*}(x)),\, f^*(\lambda_{f^*}(x))) \\
&\le \big[d(x,\, p) + d(p, f_n(\lambda_{f*}(x))) + d(x,\, p) + d(p,\, f^*(\lambda_{f^*(x)}))\big] \\
&\qquad \cdot d(f_n(\lambda_{f^*}(x)),\, f^*(\lambda_{f^*}(x))) \\
&\le 2(r + d(x,\, p)) \cdot \sup_{\lambda\in[0,\,1]} d(f_n(\lambda),\, f^*(\lambda)) \\
&= 2(r + d(x,\, p)) \cdot \mathrm{dist}(f_n,\, f^*) \qquad (6.8) \\
&\to 0 \quad \text{as } n \to \infty. \qquad (6.9)
\end{aligned}
$$

In case $d(x,\, f^*) < d(x,\, f_n)$, the inequality (6.8) is also derived in the same way. Namely, (6.8) holds for all $x \in M$, thereby implying that

$$
\begin{aligned}
\lim_{n\to\infty} |R(f_n) - R(f^*)| &= \lim_{n\to\infty} |\mathbb{E}[d^2(X,\, f_n) - d^2(X,\, f^*)]| \\
&\le \lim_{n\to\infty} \mathbb{E}|d^2(X,\, f^*) - d^2(X,\, f_n)| \\
&\overset{\text{“LDCT”}}{=} \mathbb{E} \lim_{n\to\infty} |d^2(X,\, f^*) - d^2(X,\, f_n)| \\
&= 0
\end{aligned}
$$

where the last equality holds by (6.9) and the second equality holds by $|d^2(X,\, f^*) - d^2(X,\, f_n)| \le 2r + 2d(X,\, p)$ from (6.8) for sufficiently large n, $2r + 2\mathbb{E}d(X,\, p) \le 2r + 2\sqrt{\mathbb{E}d^2(X,\, p)} < \infty$ (Jensen's inequality), and Lebesgue's dominated convergence theorem (LDCT). Accordingly, above limits exist and

$$
R^* = \lim_{n\to\infty} R(f_n) = R(f^*).
$$

Therefore, $f^* \in C([0,\,1],\,N) \subset C([0,\,1],\,M)$ is a principal curve with length ℓ. □

Theorem 6 shows that a principal curve with length ℓ always exists. The uniqueness is not assured as in Kégl (1999); Kégl et al. (2000). It means that there can be several principal curves for a given M-valued random variable X. However, their risks are all same. Suppose that the observations $\{X_i\}_{i=1}^n$ are i.i.d. samples from an M-valued random variable X, where n is the number of observations. Let $k \geq 1$ be the number of vertex point of a curve to be estimated, let $\mathcal{F}$ be a family of continuous curves in M whose lengths are not greater than $\ell \geq 0$, and denote $\mathcal{F}_k$ as a family of curves whose lengths are not greater than ℓ consisting of k-geodesic segments. With an increasing complexity of function classes, we clearly have $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_k \subset \ldots \subset \mathcal{F}$ since a single geodesic segment can considered as two geodesic segments by splitting the single geodesic segment. Formally, the principal curve with length ℓ is any f^* such that

$$f^* = \mathrm{argmin}_{f \in \mathcal{F}}\, \mathbb{E}d^2(X,\,f) = \mathrm{argmin}_{f \in \mathcal{F}}\, R(f)$$

We proved that the principal curves f^* always exists when X has second moment by Theorem 6. Alternatively, the minimizer of risk in the restricted class $\mathcal{F}_k$, say f_k^*, is any

$$f_k^* = \mathrm{argmin}_{f \in \mathcal{F}_k}\, \mathbb{E}d^2(X,\,f) = \mathrm{argmin}_{f \in \mathcal{F}_k}\, R(f), \tag{6.10}$$

as long as it exists. The existence of f_k^* is guaranteed by the following Theorem 7.

Theorem 7. *Under conditions $(B1)-(B2)$ stated later in next page, f_k^* exists for any $k \geq 1$.*

Proof. See Appendix A.3. □

Theorem 7 means that (6.10) is well-defined. In practice, empirical minimizer $f_{k,\,n}$ is naturally defined as

$$f_{k,\,n} = \mathrm{argmin}_{f \in \mathcal{F}_k}\, \frac{1}{n}\sum_{i=1}^{n} d^2(X_i,\,f). \tag{6.11}$$

Kégl (1999); Kégl et al. (2000) proved the consistency of principal curve and moreover proved its $n^{-1/3}$-convergence rate of *excess risk (accuracy)* by means of empirical risk minimization. In this work, this properties are extended into general Riemannian manifolds with same sign of curvatures. To this end, the following are assumed:

(B1) M is complete and connected as a metric space. The M-valued random variable X is supported on some closed, bounded, and convex subset $N \subset M$, i.e, $\mathbf{P}(X \in N) = 1$.

(B2) M satisfies either of the following:

(a) Any sectional curvatures of M are nonpositive.

(b) Any sectional curvatures of M, say K, are bounded below by some positive constant δ, i.e., $0 < \delta \leq K$. It also holds that $\mathrm{diam}(N) < \pi/\sqrt{\kappa}$ where κ is the supremum of curvatures on M.

In the case (b) of (B2), the curvature K is bounded below by $\delta > 0$. By Bonnet's theorem (e.g., Chapter 11, page 200–201, in Lee (2006)), M is compact and then the supremum of sectional curvature, say κ, is achieved. (For a proof, see Chapter 9.3, page 166, in Bishop and Crittenden (2011)). It means that $\kappa < \infty$ and that the condition, $\mathrm{diam}(N) < \pi/\sqrt{\kappa} \neq 0$, is hence meaningful. (In the case of D-sphere with radius $1/\sqrt{\kappa}$, the diameter of any half D-sphere equals to $\pi/\sqrt{k}$). In the assumption (B2), we restrict that M has same sign of curvatures, as in Fletcher et al. (2009); Dai and Müller (2018); Lin and Yao (2019). To the best of our knowledge, the class of manifolds that is more wider than (B2) has not been covered in statistical and computer science communities. The class of manifold with same sign of curvature assumed in (B2) is sufficiently *wide* in statistical views, while a more larger class of spaces, such as $\mathrm{CAT}(\kappa)$ (the collection of spaces with curvatures at most κ for some $\kappa \in \mathbb{R}$), mainly belongs to the area of differential geometry, a field of mathematics.

For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we denote $a_n \asymp b_n$ when $C_1|b_n| < |a_n| < C_2|b_n|$ for some constants C_1, $C_2 > 0$. The number of vertices k is chosen by

$k \asymp n^{1/3}$. To prove the asymptotic properties of the proposed procedure thoroughly, we lastly assume that

(B3) For any $k \geq 1$,

$$f_{k,\,n} = \operatorname*{arg\,min}_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} d^2(X_i,\, f)$$

exists.

Let $\mathcal{X}_n = \{X_1,\, X_2,\, \ldots,\, X_n\}$ and X be the i.i.d. copies of a probability distribution on M. By the definition of risk (6.2), we can denote $R(f_{k,\,n}) = \mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n]$ where $\mathbb{E}$ is the expectation taken over X. $\mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n}) - R(f^*)] = \mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n})] - R(f^*)$, is nonnegative by the definition of f^* where $\mathbb{E}_{\mathcal{X}_n}$ denotes the expectation taken over $\mathcal{X}_n$. By the following theorem, the procedure is *consistent* and its convergence rate is $n^{-1/3}$ order.

Theorem 8 (Consistency and cubic-convergence rate). *Under $(B1) - (B3)$, if $k \asymp n^{1/3}$, then the procedure is consistent and has $n^{-1/3}$-convergence rate in the sense that*

$$R(f_{k,\,n}) \xrightarrow{L_1} R(f^*) \quad \textit{and} \quad \mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n}) - R(f^*)] = O(n^{-1/3}) \qquad \textit{as } n \to \infty,$$

where $\xrightarrow{L_1}$ denotes the L_1-convergence.

Proof. See Appendix A.3. □

Note that L_1-convergence implies convergence in probability by Markov's inequality. The theorem thus implies that the procedure is *consistent*; that is,

$$R(f_{k,\,n}) \xrightarrow{P} R(f^*) \quad \text{as } \to \infty$$

where $\xrightarrow{P}$ denotes the convergence in probability. In the terminology of empirical risk minimization principle, moreover, $R(f_{k,\,n}) - R(f^*)$ is termed as *excess risk* (accuracy) of $f_{k,\,n}$. A nonasymptotic concentration inequality for the excess risk of $f_{k,\,n}$ can be established as follows.

Theorem 9. *Under* $(B1)-(B3)$*, for any* $0<\delta<1$*, with probability at least* $1-\delta$ *we get*

$$R(f_{k,n})-R(f^*)\le\sqrt{\frac{C(\ell,\,r)k-2r^4\log\delta}{n}}+\frac{2(r\ell+1)}{k}.$$

Proof. See Appendix A.3. □

Theorem 9 is statistically important because it gives an $100(1-\delta)\%$ confidence interval for $R(f^*)$. Even if the theoretical properties for LPC is developed, the practical algorithm for constructing LPC procedure is in progress.

6.4 Real data analysis

For real data analysis, seismological data near the East Sea (6.5+ Mb magnitude) near the U.S. since 1900 are collected from U.S. Geological Survey. The data have 165 observations and are distributed vicinity from the borders of plates, as shown in Figure 6.4.

Figure 6.5 represents the consequences of existing methods. As shown, the methods cannot identify the local structures of the seismological data. On the other hands, Figure 6.6 shows the results which are implemented by LPG with $h=0.3$, $\nu=0.05$ (left) and $h=0.11$ and $\nu=0.05$ (right). The consequence (right) of LPG with suitably adjusted parameters is able to extract a local structure of the seismological data.

6.5 Further work

As mentioned previously, the work for local principal curves on Riemannian manifolds (LPCRM) is in progress. So far, initialization method (LPG) to capture complex structured data is provided and theoretical consequences on generic Riemannian manifold with same sign of curvatures are established. The remaining part is to thoroughly construct the practical algorithm (polygonal line algorithm) based on Kégl (1999); Kégl et al. (2000). There is another line of further work. To robustify the whole procedure, along the same line with Chapter 4, an M-type local principal curve on Riemannian manifolds could be developed as a minimizer of L_1- or Huber type risk instead of L_2 risk. In this case, the existence, consistency, and convergence rate of the robust procedure could also be guaranteed in the similar argument. However it is beyond the scope of a single study. We hence let this topic be left as future study.

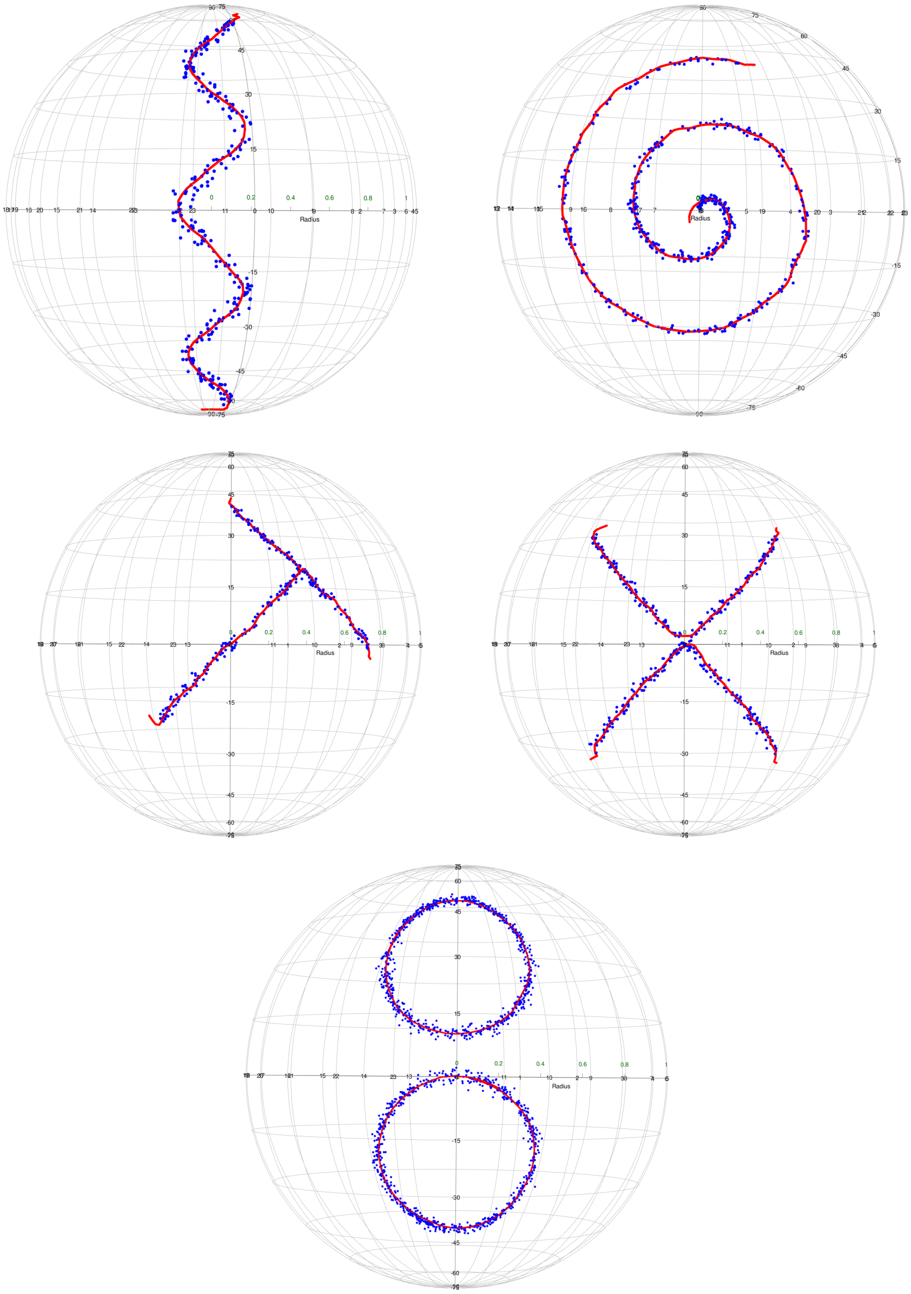


Figure 6.3: From top to bottom and left to right, simulated zigzag, spiral, T-shaped, X-shaped, doubly circular data are colored in blue. The LPG consequences of zigzag for $h = 0.07$, spiral for $h = 0.07$, T-shaped for $h = 0.05$, X-shaped for $h = 0.05$, double circle for $h = 0.1$, and helix for $h = 0.15$ are colored in red. In all cases, viscosity is set to be $\nu = 0.1$.

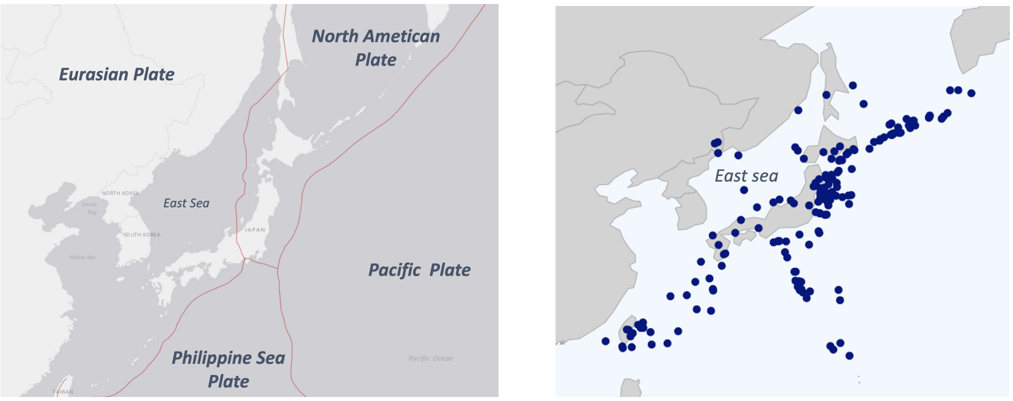


Figure 6.4: Left: The borders (red) of plates near the East Sea (source: U.S. geological Survey); Right: The distribution of earthquakes (blue) near the East Sea

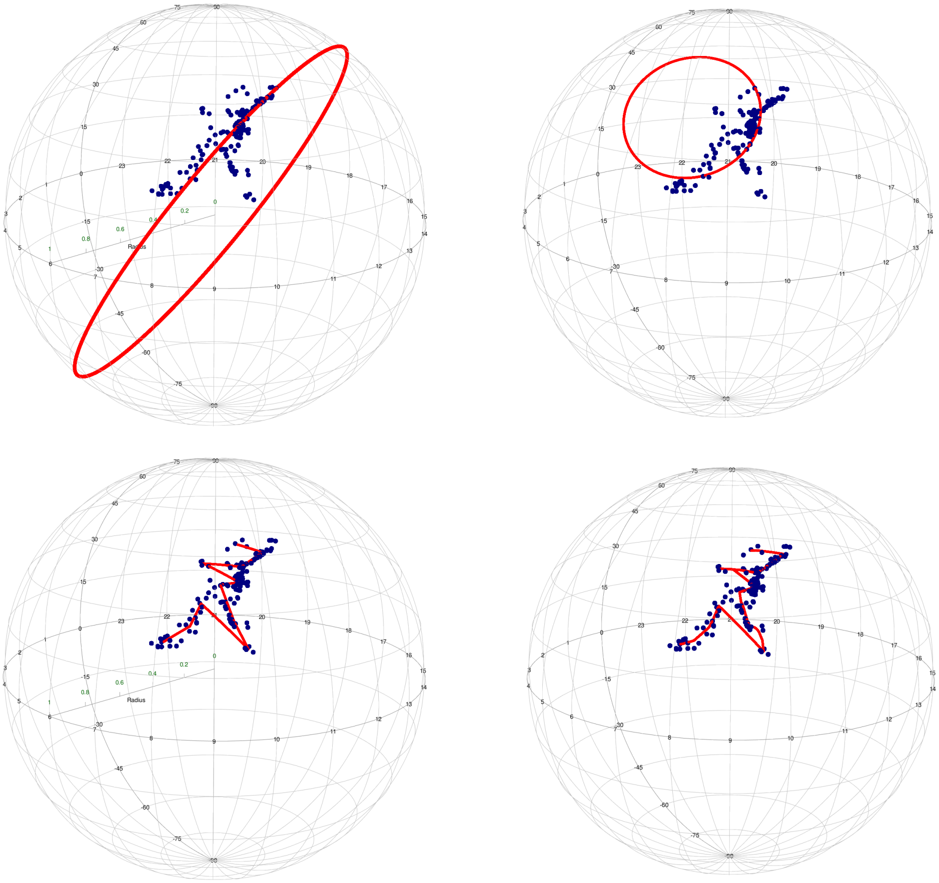


Figure 6.5: From top to bottom and left to right, the consequences of PGA, principal circle, principal curves of Hauberg, and SPC with $q = 0.05$, are colored in red.

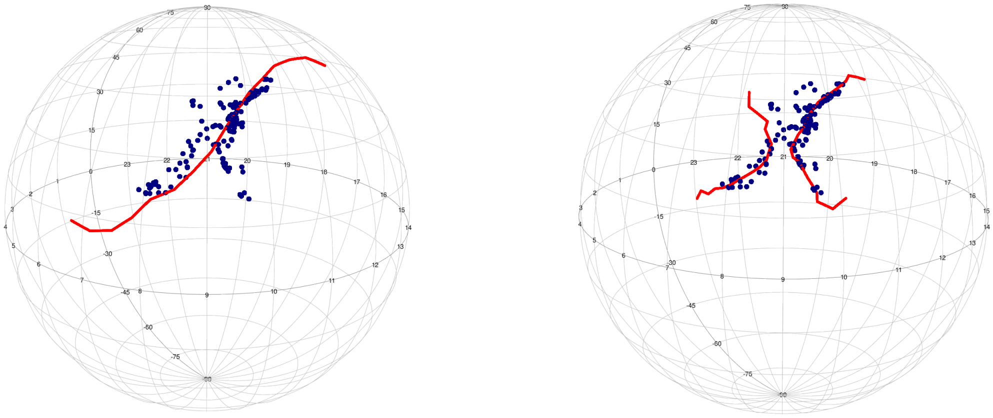


Figure 6.6: From left to right, the consequences of LPG with $h = 0.3$, $\nu = 0.05$ and with $h = 0.11$, $\nu = 0.05$ are colored in red.

Chapter 7

Conclusion

In the thesis, nonparametric dimension reduction methods on spheres and Riemannian manifolds are presented. Specifically, Chapter 3 provided spherical principal curves on spheres and establish the theoretical properties of the method. The procedure is robustified and corresponding theory is established in Chapter 4. The methods are implemented to real earthquake data and real motion capture data. Chapter 5 introduces an R package **spherepc** which provides existing methods and the proposed methods throughout this thesis. Finally, local principal curves on Riemannian manifolds (LPCRM) is presented in Chapter 6. The procedure has advantages in that (i) it is capable of identifying the complicated structures of a given data, (ii) the spaces to which the method can be applied is extended to generic Riemannian manifolds, and (iii) solid theories on the procedure are established.

There are several ways to further work. Examples of possible future studies are given as follows: (a) As shown in Chapter 3, we observe that for sphere-valued data both extrinsic and intrinsic approaches yield similar performance. It is compatible to the fact that the intrinsic and extrinsic means are **indistinguishable** if distribution of data has small support (Bhattacharya and Patrangenaru, 2003, 2005; Bhattacharya et al., 2012). In my experience, moreover, the extrinsic and intrinsic means are very close each other when data are *symmetrically* distributed and the space where the data is lying are *symmetric* (e.g. $\mathbb{R}^D, S^D$, Riemannian symmetric space, and homogeneous Lie group). In this case, it is worth to investigating the upper and lower bounds on differences between the two means. On the one hand, it is questionable whether the extrinsic approach on non-symmetric manifold will still be efficient. For the non-symmetric manifold like a torus, it seems that the intrinsic approach may yield better performance due to their inherency. (b) Investigate the theoretical properties for Huber-type centroid in Chapter 4 such as existence, uniqueness, and convergence of related algorithm, along the line of Fletcher et al. (2009); Yang (2010); Afsari (2011). (c) Beyond the central tendency, develop a quantile-based principal curve in the sense of Chaudhuri (1996); Chowdhury and Chaudhuri (2019). The quantile-based approach could be useful. (d) Apply the consequence of dimension reduction presented in the thesis to other statistical and machine learning tasks, such as clustering, regression or classification (Goh and Vidal, 2008; Mallasto and Feragen, 2018; Yao and Zhang, 2020).

Appendix A

Appendix

A.1 Appendix for Chapter 3

Proof of Proposition 3. It suffices to show that, for any constant $c \in [0, 1]$, $\{\lambda_f(x) \geq c\}$ is measurable. If $c = 0$, it is the entire space; thus, we may assume that $c \in (0, 1]$. Let A be the set of ambiguity points. A is a measure zero set by Proposition 2. It can be shown that the spherical measure inherits completeness from the Lebesgue measure, which means that any subsets of a null set are measurable. Therefore A, A^c, and their subsets are measurable in which A^c denotes the complement of A. Note that

$$\{\lambda_f(x) \geq c\} = \big[\,\{\lambda_f(x) \geq c\} \cap A\big] \cup \big[\,\{\lambda_f(x) \geq c\} \cap A^c\big],$$

where the former is measurable since it is a subset of A. In this respect, it is enough to show that

$$\{\lambda_f(x) \geq c\} \cap A^c$$

is measurable. $y \in \big\{x \in S^D \,|\, \lambda_f(x) \geq c\big\} \cap A^c \overset{(*)}{\iff} y$ is not an ambiguity point and for any $\mu \in [0, c) \cap \mathbb{Q}$ there exist a $\lambda \in [c, 1] \cap \mathbb{Q}$ such that $d(y, f(\lambda)) < d(y, f(\mu))$ where $\mathbb{Q}$ is the set of rational numbers. Once this is proven,

$$\{\lambda_f(x) \geq c\} \cap A^c = \left\{ \bigcap_{\mu \in [0,\, c) \cap \mathbb{Q}} \; \bigcup_{\lambda \in [c,\, 1] \cap \mathbb{Q}} \{d(x, f(\lambda)) < d(x, f(\mu))\} \right\} \cap A^c. \quad \text{(A.1)}$$

Each set in the right-hand side of (A.1) is measurable since, for any $\mu,\ \lambda$, the function $x \mapsto d(x,\, f(\mu)) - d(x,\, f(\lambda))$ is continuous. Accordingly, $\{\lambda_f(x) \geq c\} \cap A^c$ is measurable due to the fact that countable unions and intersections of measurable sets are also measurable, which completes the assertion.

Proof of (*). ($\Rightarrow$): Since the closest point in f from y is unique, it follows from the continuity of f and the definition of the projection index.

($\Leftarrow$): Suppose that $\lambda_f(y) < c$. From the non-ambiguity of y, observe that

$$d(y,\, f(\lambda_f(y))) < \min_{\lambda \in [c,\, 1]} d(y,\, f(\lambda)).$$

By the continuity of f again, there exists a $\mu_0 \in [0,\, c] \cap \mathbb{Q}$ such that for any $\lambda_0 \in [c,\, 1] \cap \mathbb{Q}$

$$d(y,\, f(\mu_0)) < \min_{\lambda \in [c,\, 1]} d(y,\, f(\lambda)) \leq d(y,\, f(\lambda_0)).$$

It completes the $(*)$, as desired. □

Proof of Proposition 4. The proof follows the line of the proof of Lemma 4.1 in Hastie and Stuetzle (1989). It is enough to show that, for any small $\eta > 0$ there exists a $\delta > 0$ such that $|\epsilon| < \delta$ implies $|\lambda_{f_\epsilon}(x) - \lambda_f(x)| < \eta$. Define a set $C = [0,\, 1] \cap (\lambda_f(x) - \eta,\, \lambda_f(x) + \eta)^c$ and $d_C = \inf_{\lambda \in C} d(x,\, f(\lambda)) > d(x,\, f(\lambda_f(x)))$ where d_C is achieved by some $\lambda \in C$ from the compactness of C, and the inequality holds since x is not an ambiguity point of f. Choose $\delta = \frac{1}{3}\big[d_C - d(x,\, f(\lambda_f(x)))\big] > 0$. Then if $|\epsilon| < \delta$ then

$$\begin{aligned}
&\inf_{\lambda \in C} d\big(x,\, f_\epsilon(\lambda)\big) - d\big(x,\, f_\epsilon(\lambda_f(x))\big) \\
&\qquad \geq \inf_{\lambda \in C} d\big(x,\, f(\lambda)\big) - d(f(\lambda),\, f_\epsilon(\lambda)) - d(x,\, f(\lambda_f(x))) - d(f(\lambda_f(x)),\, f_\epsilon(\lambda_f(x))) \\
&\qquad \geq d_C - \delta - d(x,\, f(\lambda_f(x))) - \delta \\
&\qquad = 3\delta - 2\delta > 0.
\end{aligned}$$

Hence,

$$\inf_{\lambda \in C} d(x,\, f_\epsilon(\lambda)) > d(x,\, f_\epsilon(\lambda_f(x))).$$

If $\lambda_{f_\epsilon}(x) \in C$, then $\inf_{\lambda \in C} d(x,\, f_\epsilon(\lambda)) = d(x,\, f_\epsilon(\lambda_{f_\epsilon}(x))) > d(x,\, f_\epsilon(\lambda_f(x)))$, which is a contradiction by the definition of λ_{f_ϵ}. It follows that $\lambda_{f_\epsilon}(x) \notin C$; thus, $|\lambda_f(x) - \lambda_{f_\epsilon}(x)| < \eta$. It completes the assertion. □

Proof of Lemma 4. Suppose that $I : [0,\, 1] \to S^D$ is smooth (actually $\mathcal{C}^2$). The domain and range of I are the second countable (with usual topology) and differentiable manifolds whose dimensions are 1 and D, respectively. Since f is $\mathcal{C}^2$ and the differential dI has rank 1 that is less than intrinsic dimension of S^D, by a generalization of Sard's Theorem (for details see Theorem 1 in Sard (1965)), the image $I([0,\, 1]) = \{I(x) \in S^D \,|\, x \in [0,\, 1]\}$ has (D-dimensional Hausdorff) measure zero. Next, for simplicity, assume $D = 2$. The general case is also proved in the same way. Each point $x \in S^2 \setminus B(0)$ satisfying $\lambda_f(x) \neq 0,\, 1$ is characterized by two equations $f'(\lambda) \cdot x = 0$ and $f''(\lambda) \cdot x = 0$ for some $\lambda \in [0,\, 1]$. Therefore, we define functions I_1, I_2 as follows: For all $\lambda \in [0,\, 1]$,

$$\begin{aligned} I_1(\lambda) &= f'(\lambda) \times f''(\lambda) / \left\| f'(\lambda) \times f''(\lambda) \right\|, \\ I_2(\lambda) &= -f'(\lambda) \times f''(\lambda) / \left\| f'(\lambda) \times f''(\lambda) \right\|. \end{aligned}$$

It is well known that the curvature of a smooth curve lying on the unit sphere is more than 1. It implies that $\kappa = |f''|/s^2 \geq 1$ where κ is the curvature of f and $s := |f'(\lambda)| > 0$ for any $\lambda \in [0,\, 1]$ and thus $f'' \neq 0$. We have already known that $f' \cdot f'' = 0$ by Lemma 1. Hence, we have $f' \times f'' \neq 0$. It implies that I_1 and I_2 are well-defined and also smooth. Note that $S^2 \setminus B(\zeta)$ is decreasing as $n \to \infty$ by construction and that $S^2 \setminus B(0) \subseteq I_1([0,\, 1]) \bigcup I_2([0,\, 1])$, which completes the assertion. □

Proof of Lemma 5. By Proposition 5 and the assumptions that x is a non-ambiguity point of f, and $\lambda_f(x) \neq 0,\, 1$, we obtain $\lambda_{f+\epsilon g}(x) \neq 0,\, 1$ for sufficiently small $|\epsilon|$. On an ϵ near 0, $\lambda(\epsilon)$ is characterized by orthogonality between $f'_\epsilon(\lambda(\epsilon))$ and the geodesic through x and $f_\epsilon(\lambda(\epsilon))$; that is, $f'_\epsilon(\lambda) \cdot (x - f_\epsilon(\lambda)) = f'_\epsilon(\lambda) \cdot x = 0$ by the same argument in Lemma 1. Then, we define a map $F : [-1,\, 1] \times [0,\, 1] \to \mathbb{R}$ as

$F(\epsilon, \lambda) = f'_\epsilon(\lambda) \cdot x$. The map F is smooth by Proposition 1. By the definition of $B(\zeta)$,

$$\frac{\partial}{\partial \lambda} F(\epsilon, \lambda)\big|_{(0, \lambda_f(x))} \;\; = \;\; f''(\lambda_f(x)) \cdot x \neq 0.$$

By implicit function theorem, for each $x \in A^c \cap B(\zeta)$, $\lambda(\epsilon) = \lambda_{f_\epsilon}(x)$ is a smooth function for ϵ and $F(\epsilon, \lambda(\epsilon)) = 0$ in an open interval containing zero. Next, so as to prove uniform boundedness of $\frac{\partial \lambda(\epsilon)}{\partial \epsilon}$, we should show that $f''_\epsilon(\lambda_{f_\epsilon}(x))$ uniformly converges to $f''(\lambda_f(x))$ as $\epsilon \to 0$ on $x \in A^c \cap B(\zeta)$. First of all, for any λ and ϵ_0,

$$\begin{aligned} f_{\epsilon_0}(\lambda) \;\; &= \;\; f(\lambda) + \int_0^{\epsilon_0} g(\epsilon, \lambda)\, d\epsilon \\ &\Rightarrow \;\; f''_{\epsilon_0}(\lambda) = f''(\lambda) + \int_0^{\epsilon_0} g''(\epsilon, \lambda)\, d\epsilon \\ &\Rightarrow \;\; \big\| f''_{\epsilon_0}(\lambda) - f''(\lambda) \big\| \leq \int_0^{\epsilon_0} \big\| g''(\epsilon, \lambda) \big\|\, d\epsilon \leq \epsilon_0 M, \end{aligned} \tag{A.2}$$

for some $M > 0$. Note that the above derivatives are differentiation by λ. Also, the second equation holds true since $g(\epsilon, \cdot)$ is a twice continuously differentiable function for any ϵ; thus, it is able to change the order of derivative and the integration by dominated convergence theorem. The last inequality holds because $g''(\epsilon, \lambda)\big(= \frac{\partial^2 g(\epsilon, \lambda)}{\partial \lambda^2}\big)$ is continuous on $[-1, 1] \times [0, 1]$. Hence,

$$\begin{aligned} \big\| f''_\epsilon(\lambda_{f_\epsilon}(x)\big) - f''(\lambda_f(x)) \big\| \;\; &\leq \;\; \big\| f''_\epsilon(\lambda_{f_\epsilon}(x)) - f''(\lambda_{f_\epsilon}(x)) \big\| \\ &\quad + \big\| f''(\lambda_{f_\epsilon}(x)) - f''(\lambda_f(x)) \big\| \\ &\quad \to 0, \end{aligned} \tag{A.3}$$

as $\epsilon \to 0$ uniformly on $x \in A^c \cap B(\zeta)$, because the first term uniformly converges to 0 by (A.2) and the last one uniformly converges to 0 by Proposition 5 and the boundedness of f''' (assumption (A1)). We obtain that $|x \cdot f''\big(\lambda_f(x)\big)| > \zeta$ owing to $x \in B(\zeta)$. By (A.3), there exists a $\delta > 0$ such that $|\epsilon| < \delta \Rightarrow |x \cdot f''_\epsilon\big(\lambda_{f_\epsilon}(x)\big)| \geq \zeta/2$. Since $f_\epsilon(\lambda) = f(\epsilon, \lambda)$ has continuous second partial derivatives, it is able to change the order of partial derivatives by Schwarz's theorem, as

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} f'_\epsilon(\lambda(\epsilon_0)) &= \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} \frac{\partial}{\partial \lambda}\Big|_{\lambda=\lambda(\epsilon_0)} f_\epsilon(\lambda) \\
&= \frac{\partial}{\partial \lambda}\Big|_{\lambda=\lambda(\epsilon_0)} \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} f_\epsilon(\lambda) \\
&= g'(\epsilon_0,\, \lambda(\epsilon_0)),
\end{aligned}$$

for all $|\epsilon_0| < \delta$. By applying the implicit function theorem to F again, it follows that $\lambda(\epsilon)$ is differentiable at $\epsilon = \epsilon_0$ and

$$\begin{aligned}
|\lambda'(\epsilon_0)| = \Big|\frac{\partial \lambda_{f_\epsilon}(x)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0}\Big| &= \Big|\frac{-\partial F(\epsilon,\, \lambda)/\partial \epsilon}{\partial F(\epsilon,\, \lambda)/\partial \lambda}\Big|_{(\epsilon_0,\, \lambda(\epsilon_0))}\Big| \\
&= \frac{\Big|x \cdot \frac{\partial}{\partial \epsilon}\big|_{\epsilon=\epsilon_0} f'_\epsilon(\lambda(\epsilon_0))\Big|}{\Big|x \cdot f''_{\epsilon_0}(\lambda(\epsilon_0))\Big|} \le \frac{\|g'\|}{\zeta/2} \\
&\le 2/\zeta,
\end{aligned}$$

provided that $|\epsilon_0| < \delta$, which completes the proof. □

A.2 Appendix for Chapter 4

Proof of Theorem 3

Proof. For ease of notation, we denote $f+g$ as h. If $f = h$ then it is clear. For given $f \neq h$, we fix a point $X \in S^2$ satisfying $X \in A^c \cap B(\zeta)$ for some $\zeta > 0$ and $\lambda_f(X) \in (0,\, 1)$. (Note assumption (A3)). Denote $f_\epsilon := f + \epsilon g$ and $\lambda(\epsilon) := \lambda_{f_\epsilon}(X)$ for $|\epsilon| \le 1$. In particular, $\lambda(0) = \lambda_f(X)$. In addition, we sometimes omit X and respectively express λ_{f_ϵ} and λ_f, instead of $\lambda_{f_\epsilon}(X)$ and $\lambda_f(X)$. We apply Lebesgue's dominated convergence theorem so as to change the order of derivative and expectation in (4.4). To this end, we aim to show that $Z_\epsilon(X)$ is uniformly bounded on $X \in A^c \cap B(\zeta)$ for sufficiently small $|\epsilon|$. According to Lemma 5, there are constants $\eta > 0$ and $C > 0$ such that if $|\epsilon_0| < \eta$ then $\lambda(\epsilon)$ is differentiable at $\epsilon = \epsilon_0$ and $\big|\frac{\partial \lambda(\epsilon)}{\partial \epsilon}\big|_{\epsilon=\epsilon_0}\big| < C$, where $\lambda(\epsilon) := \lambda_{f_\epsilon}(X)$ for brevity. If $0 < |\epsilon_0| < \eta$, it follows from the triangular inequality that

$$\begin{aligned}
|Z_{\epsilon_0}(X)| &:= \left|\frac{d(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})) - d(X,\, f(\lambda_f))}{\epsilon_0}\right| \\
&= \left|\frac{d(X,\, f_{\epsilon_0}(\lambda(\epsilon_0))) - d(X,\, f(\lambda(0)))}{\epsilon_0}\right| \\
&\leq \left|\frac{d(f_{\epsilon_0}(\lambda(\epsilon_0)),\, f(\lambda(0)))}{\epsilon_0}\right| \\
&\leq \left|\frac{d(f(\lambda_f),\, f(\lambda(\epsilon_0))) + d(f(\lambda(\epsilon_0)),\, f_{\epsilon_0}(\lambda(\epsilon_0)))}{\epsilon_0}\right| \\
&\leq s \cdot \frac{|\lambda(0) - \lambda(\epsilon_0)|}{\epsilon_0} + \|g(\lambda(\epsilon_0))\|, \\
&\leq sC + \pi,
\end{aligned}$$

where $s := |f'(\lambda)|$ for all λ. The last inequality is done by mean value theorem. Hence, $Z_\epsilon(X)$ is uniformly bounded on $X \in A^c \cap B(\zeta)$ for $0 < |\epsilon| < \eta$.

Next, we will find the limit of $Z_\epsilon(X)$ as $\epsilon \to 0$. Let $\theta(\lambda, X)$ be the angle between two geodesic segments that connect $f(\lambda)$ with X and $f(\lambda)$ with $(f+g)(\lambda)$, respectively. Define $\text{ang}(\lambda) := |\cos(\theta(\lambda,\, X))|$ where X satisfies $\lambda_f(X) = \lambda$. The well-definedness of $\text{ang}(\lambda)$ can be easily proved. By Lemma 3,

$$\begin{aligned}
F(\epsilon) := \cos(d(X,\, f_\epsilon(\lambda_{f_\epsilon}))) = \cos(d(X,\, f(\lambda_{f_\epsilon}))) \cdot \cos(\epsilon\, \|g(\lambda_{f_\epsilon})\|) \\
+ \sin(d(X,\, f(\lambda_{f_\epsilon}))) \cdot \sin(\epsilon\, \|g(\lambda_{f_\epsilon})\|) \cdot \cos(\theta(\lambda_{f_\epsilon},\, X)),
\end{aligned}$$

where $\|g(\lambda)\| = d(f(\lambda),\, (f+g)(\lambda)) < \pi$. Using the chain rule to the derivative of F,

$$\begin{aligned}
\lim_{\epsilon_0 \to 0} \frac{\partial F(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} &= \lim_{\epsilon_0 \to 0} \big[\sin\big(d(X,\, f(\lambda(\epsilon_0)))\big) \cdot \cos(\theta(\lambda(\epsilon_0),\, X)) \\
&\qquad \cdot \big(\|g(\lambda(\epsilon_0))\| + \epsilon_0 \cdot \frac{\partial\, \|g(\lambda(\epsilon))\|}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0}\big)\big] \\
&\quad - \lim_{\epsilon_0 \to 0} \big[\sin\big(d(X,\, f(\lambda(\epsilon_0)))\big) \cdot \frac{\partial d(X,\, f(\lambda(\epsilon)))}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0}\big].
\end{aligned}$$

In addition,

$$\frac{\partial\, \|g(\lambda(\epsilon))\|}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0} = \frac{\partial\, \|g(\lambda)\|}{\partial \lambda}\Big|_{\lambda=\lambda(\epsilon_0)} \frac{\partial \lambda(\epsilon)}{\partial \epsilon}\Big|_{\epsilon=\epsilon_0},$$

which does not diverge as $\epsilon_0 \to 0$ and exists, since $\|g(\lambda)\| = d(f(\lambda),\, (f+g)(\lambda))$ is continuously differentiable for λ and $\frac{\partial\lambda(\epsilon)}{\partial\epsilon}\big|_{\epsilon=0}$ is bounded by Lemma 5. Moreover

$$\begin{aligned}\lim_{\epsilon_0\to 0}\frac{\partial d(X,\, f(\lambda(\epsilon)))}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0} &= \lim_{\epsilon_0\to 0}\frac{\partial d(X,\, f(\lambda))}{\partial\lambda}\Big|_{\lambda=\lambda(\epsilon_0)}\cdot\frac{\partial\lambda(\epsilon)}{\partial\epsilon}\Big|_{\epsilon=\epsilon_0}\\ &= \frac{\partial d(X,\, f(\lambda))}{\partial\lambda}\Big|_{\lambda=\lambda_f}\frac{\partial\lambda(\epsilon)}{\partial\epsilon}\Big|_{\epsilon=0} = 0,\end{aligned}$$

where $\lambda(\epsilon) = \lambda_{f_\epsilon}(X)$. The last equality is done by the definition of $\lambda_f(X)$. Accordingly, we obtain that

$$\lim_{\epsilon\to 0}\frac{\partial F(\epsilon)}{\partial\epsilon} = \|g(\lambda_f)\|\cdot\cos(\theta(\lambda_f,\, X))\cdot\sin(d(X,\, f(\lambda_f))). \tag{A.4}$$

By mean value theorem, for any $0 < |\epsilon_0| < \eta$, there exists a $0 < |\epsilon_1| < |\epsilon_0|$ such that

$$\begin{aligned}Z_{\epsilon_0}(X) &= \frac{d(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})) - d(X,\, f(\lambda_f))}{\epsilon_0}\\ &= \frac{\arccos(F(\epsilon_0)) - \arccos(F(0))}{\epsilon_0}\\ &= -\frac{1}{\sqrt{1-F^2(\epsilon_1)}}\cdot\frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1}.\end{aligned}$$

By (A.4),

$$\begin{aligned}\lim_{\epsilon_0\to 0} Z_{\epsilon_0}(X) &= \lim_{\epsilon_1\to 0}\Big[-\frac{1}{\sqrt{1-F^2(\epsilon_1)}}\cdot\frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1}\Big]\\ &= -\frac{1}{\sin(d(X,\, f(\lambda_f)))}\cdot\|g(\lambda_f)\|\cdot\cos(\theta(\lambda_f, X))\cdot\sin(d(X,\, f(\lambda_f)))\\ &= -\,\|g(\lambda_f)\|\cdot\cos(\theta(\lambda_f, X)),\end{aligned} \tag{A.5}$$

where (A.5) holds for any $X \in A^c \cap B(\zeta)$ satisfying $d(X,\, f(\lambda_f)) > 0$. Since the trajectory of f, $\{f(\lambda) \in S^2 \,|\, \lambda \in [0,\, 1]\}$, has measure zero by Lemma 4, (A.5) holds for a.e. $X \in B(\zeta)$ (note assumption (A3)). In addition, since $M_\lambda := \{x \in S^2 \,|\, \lambda_f(X) = \lambda\}$ is contained in a great circle on S^2, a proper connected subset of M_λ is isometric to a same-length interval in $\mathbb{R}$. Hence, it can be shown that f is a L_1-type principal curve of X if and only if

$$\mathbb{E}\big[\mathrm{sgn}\big(\cos(\theta(\lambda_f,\, X))\big) \;\big|\; \lambda_f(X) = \lambda\big] = 0 \;\text{ for a.e. } \lambda,$$

where $\mathrm{sgn}(x) = 1$ if $x \geq 0$ and otherwise -1.

Finally, by applying (A.5) and Lebesgue's dominated convergence theorem, we obtain

$$\begin{aligned}
&\frac{\partial \mathbb{E}\big[d(X,\, f+\epsilon g)\big]}{\partial \epsilon}\Big|_{\epsilon=0} = \lim_{\epsilon\to 0}\big[\frac{\mathbb{E}\big[d(X,\, f+\epsilon g)\big] - \mathbb{E}\big[d(X,\, f)\big]}{\epsilon}\big] \\
&= \mathbb{E}\big[\lim_{\epsilon\to 0}\frac{d(X,\, f+\epsilon g) - d(X,\, f)}{\epsilon}\big] = \mathbb{E}_\lambda\big[\mathbb{E}\big[\lim_{\epsilon\to 0} Z_\epsilon(X) \mid \lambda_f(X) = \lambda\big]\big] \\
&= -\mathbb{E}_\lambda\big[\mathbb{E}\big[\, \|g(\lambda_f(X))\| \cdot \cos(\theta(\lambda_f,\, X)) \mid \lambda_f(X) = \lambda\big]\big] \\
&= -\mathbb{E}_\lambda\big[\, \|g(\lambda)\| \cdot \mathrm{ang}(\lambda) \cdot \mathbb{E}\big[\mathrm{sgn}\big(\cos(\theta(\lambda_f,\, X))\big) \mid \lambda_f(X) = \lambda\big)\big]\big] = 0,
\end{aligned}$$

where $\cos(\theta(\lambda,\, X)) = |\cos(\theta(\lambda,\, X))| \cdot \mathrm{sgn}\big(\cos(\theta(\lambda,\, X))\big) = \mathrm{ang}(\lambda) \cdot \mathrm{sgn}\big(\cos(\theta(\lambda,\, X))\big)$ for $\lambda = \lambda_f(X)$. To prove the converse, we assume that

$$\mathbb{E}_\lambda\big[\, \|g(\lambda)\| \cdot \mathrm{ang}(\lambda) \cdot \mathbb{E}\big[\mathrm{sgn}\big(\cos(\theta(\lambda_f,\, X))\big) \mid \lambda_f(X) = \lambda\big]\big] = 0,$$

for any h $(= f + g)$ such that $\|g\| < \pi$ and $\|g'\| \leq 1$. It follows that

$$\mathbb{E}\big[\mathrm{sgn}\big(\cos(\theta(\lambda_f,\, X))\big) \mid \lambda_f(X) = \lambda\big] = 0 \text{ for a.e. } \lambda,$$

which is equivalent to that f is a L_1-type principal curve of X. □

Proof of Theorem 4

Proof. We use the same notations of proof in Theorem 3. Namely, denote $f + g$ as h. For a given $f \neq h$, we fix a point $X \in S^2$ satisfying $X \in A^c \cap B(\zeta)$ for an arbitrarily small $\zeta > 0$ and $\lambda_f(X) \in (0,\, 1)$. (Note assumption (A3)). As the proof of Theorem 3, we apply Lebesgue's dominated convergence theorem to change the order of derivative and expectation. To this end, we define

$$\begin{aligned}
Z(X,\, \epsilon) =& \frac{\rho(d(X,\, f+\epsilon g)) - \rho(d(X,\, f))}{\epsilon} \\
=& \frac{\rho\big(d(X,\, f_\epsilon(\lambda_{f_\epsilon}))\big) - \rho\big(d(X,\, f(\lambda_f))\big)}{\epsilon},
\end{aligned}$$

where a Huber loss ρ is defined by $\rho(x) = x^2$ if $|x| \leq c$, $c(2|t| - c)$ if $|x| > c$, for a constant $c > 0$.

Firstly, we aim to show that $Z(X,\,\epsilon)$ is uniformly bounded on $X \in A^c \cap B(\zeta)$ for small $|\epsilon|$. Note that the Huber loss ρ satisfies a Lipschitz condition by mean value theorem, i.e., $|\rho(x)-\rho(y)| \le 2c|x-y|$ for $x, y \in \mathbb{R}$. According to Lemma 5, there are positive constants η and C such that if $0 < |\epsilon_0| < \eta$, $\lambda(\epsilon)$ is differentiable at $\epsilon = \epsilon_0$ and $|\frac{\partial \lambda(\epsilon)}{\partial \epsilon}\big|_{\epsilon=\epsilon_0}| < C$. Thus, if $0 < |\epsilon| < \eta$, it follows by the Lipschitz condition and triangle inequality that

$$\begin{aligned}
|Z(X,\,\epsilon)| &\le 2c \cdot \frac{|d(X,\, f_\epsilon(\lambda(\epsilon))) - d(X,\, f(\lambda_f))|}{\epsilon} \\
&\le 2c \cdot \frac{d(f_\epsilon(\lambda(\epsilon)),\, f(\lambda_f))}{\epsilon} \\
&\le 2c \cdot \frac{d(f_\epsilon(\lambda(\epsilon)),\, f(\lambda(\epsilon))) + d(f(\lambda(\epsilon),\, f(\lambda_f)))}{\epsilon} \\
&\le 2c \cdot (\frac{s|\lambda(\epsilon) - \lambda(0)|}{\epsilon} + \|g(\lambda(\epsilon))\|) \\
&\le 2c \cdot (sC + \pi),
\end{aligned}$$

where $s = |f'(\lambda)|$ for any $\lambda \in [0,\,1]$. Thus $Z(X,\,\epsilon)$ is uniformly bounded on $X \in A^c \cap B(\zeta)$.

Secondly, we have to find the limit of $Z(X,\,\epsilon)$ as $\epsilon \to 0$. Owing to the equality,

$$\lim_{\epsilon \to 0} d(X,\, f_\epsilon(\lambda_{f_\epsilon})) = \lim_{\epsilon \to 0} d(X,\, f_\epsilon(\lambda(\epsilon))) = d(X,\, f(\lambda_f)).$$

Thus,

$$\begin{aligned}
\lim_{\epsilon \to 0} Z(X,\,\epsilon) = 2c \cdot \lim_{\epsilon \to 0} Z_1(X,\,\epsilon) I(d(X,\, f(\lambda_f)) > c) \\
+ \lim_{\epsilon \to 0} Z_2(X,\,\epsilon) I(d(X,\, f(\lambda_f)) \le c),
\end{aligned} \tag{A.6}$$

provided that the limits of Z_1 and Z_2 exist and these are same when $d(X,\, f(\lambda_f)) = c$, in which I denotes an indicator function and

$$\begin{aligned}
Z_1(X,\,\epsilon) &:= \frac{d(X,\, f_\epsilon(\lambda_{f_\epsilon})) - d(X,\, f(\lambda_f))}{\epsilon} \\
Z_2(X,\,\epsilon) &:= \frac{d^2(X,\, f_\epsilon(\lambda_{f_\epsilon})) - d^2(X,\, f(\lambda_f))}{\epsilon}.
\end{aligned}$$

We have already known that $\lim_{\epsilon \to 0} Z_1(X,\,\epsilon) = -\|g(\lambda_f)\| \cdot \cos(\theta(\lambda_f,\, X))$ holds for $X \in A^c \cap B(\zeta)$ by (A.5) in the proof of Theorem 3. We now aim to show that

$$\lim_{\epsilon \to 0} Z_2(X,\,\epsilon) = -2d(X,\, f(\lambda_f)) \cdot \|g(\lambda_f)\| \cdot \cos(\theta(\lambda_f,\, X)), \tag{A.7}$$

for $X \in A^c \cap B(\zeta)$. To this end, we define a map $u : (-1,\, 1] \to (1,\, \infty)$ by $u(x) = \arccos(x) \cdot \frac{1}{\sqrt{1-x^2}}$ if $x \in (-1,\, 1)$, and $u(1) = 1$. By simple calculations, u is a monotone decreasing continuous function on $(-1,\, 1]$. Note that $F(\epsilon)$ is differentiable for $|\epsilon| < \eta$. Applying mean value theorem to find the limit of $Z_2(X,\, \epsilon)$, we have

$$\begin{aligned} Z_2(X,\, \epsilon_0) &= \frac{d^2(X,\, f_{\epsilon_0}(\lambda_{f_{\epsilon_0}})) - d^2(X,\, f(\lambda_f))}{\epsilon_0} \\ &= \frac{\arccos^2(F(\epsilon_0)) - \arccos^2\big(F(0)\big)}{\epsilon_0} \\ &= -2\arccos(F(\epsilon_1)) \cdot \frac{1}{\sqrt{1-F^2(\epsilon_1)}} \cdot \frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1} \qquad \text{(A.8)} \end{aligned}$$

for $0 < |\epsilon_1| < |\epsilon_0| < \eta$. When $F(\epsilon_1) = 1$, the last equality is considered as a limit that is well-defined, because $\lim_{x\to 1} u(x) = 1$ and $u(x)$ is smoothly extended on an open interval containing 1 such that $u(x)$ is differentiable at $x = 1$. If $d(X,\, f(\lambda_f)) > 0$, it follows from (A.4) and (A.8) that

$$\begin{aligned} \lim_{\epsilon_0\to 0} Z_2(X,\, \epsilon_0) &= \lim_{\epsilon_1\to 0}\Big[-2\arccos F(\epsilon_1) \cdot \frac{1}{\sqrt{1-F^2(\epsilon_1)}} \cdot \frac{dF(\epsilon)}{d\epsilon}\Big|_{\epsilon=\epsilon_1}\Big] \\ &= -2u\big(\cos(d(X,\, f(\lambda_f)))\big) \cdot \|g(\lambda_f)\| \cdot \cos(\theta(\lambda_f,\, X)) \cdot \sin(d(X,\, f(\lambda_f))) \\ &\qquad \text{(A.9)} \\ &= \frac{-2d(X,\, f(\lambda_f))}{\sin(d(X,\, f(\lambda_f)))} \cdot \|g(\lambda_f)\| \cdot \cos(\theta(\lambda_f,\, X)) \cdot \sin(d(X,\, f(\lambda_f))) \\ &= -2d(X,\, f(\lambda_f)) \cdot \|g(\lambda_f)\| \cdot \cos(\theta(\lambda_f,\, X)), \qquad \text{(A.10)} \end{aligned}$$

In the case of $d(X,\, f(\lambda_f)) = 0$, the same result follows since both (A.9) and (A.10) are zero. That is, (A.10) is established for $X \in A^c \cap B(\zeta)$. From Lemma 4, (A.5) and (A.7) hold for a.e. $X \in B(\zeta)$ (by the assumption (A3)).

Thirdly, we aim to find an equivalent condition to be a Huber-type principal curve. Denote $M_\lambda = \big\{x \in S^2 \,|\, \lambda_f(x) = \lambda\big\}$. Since M_λ is contained in a great circle on S^2, a proper connected subset of M_λ is isometric to a same-length interval in $\mathbb{R}$. Consider a random variable Y compactly supported on $\mathbb{R}$ and let $\mu_0 \in \mathbb{R}$ be a minimizer of $m(\mu) = \mathbb{E}[\rho(Y-\mu)]$. The derivative of ρ is

$$\rho'(t) = \begin{cases} 2c & \text{if } t > c, \\ 2t & \text{if } |t| \le c, \\ -2c & \text{if } t < -c. \end{cases}$$

By Lebesgue's dominated convergence theorem,

$$\begin{aligned} 0 = &\frac{\partial m(\mu)}{\partial \mu}\Big|_{\mu=\mu_0} = \frac{\partial \mathbb{E}[\rho(Y-\mu)]}{\partial \mu}\Big|_{\mu=\mu_0} = \mathbb{E}[\frac{\partial \rho(Y-\mu)}{\partial \mu}\Big|_{\mu=\mu_0}] \\ = &-2\mathbb{E}[cI(Y-\mu_0 > c) + (Y-\mu_0)I(|Y-\mu_0| \le c) - cI(Y-\mu_0 < -c)] \\ = &-2\mathbb{E}[|Y-\mu_0| \cdot \operatorname{sgn}(Y-\mu_0)I(|Y-\mu_0| \le c) \\ &+ c \cdot \operatorname{sgn}(Y-\mu_0)I(|Y-\mu_0| > c)], \end{aligned} \tag{A.11}$$

where $\operatorname{sgn}(x) = 1$ if $x \ge 0$ and otherwise -1. In our case, on M_λ, because $f(\lambda)$ is a Huber-type measure on $X \mid \lambda_f(X) = \lambda$, the distance $Y - \mu_0$ in (A.11) and its sign, are replaced by $d(X, f(\lambda_f))$ and $\operatorname{sgn}\big(\cos(\theta(\lambda_f, X))\big)$ respectively. In this respect, f is a Huber-type principal curve of X if and only if

$$\begin{aligned} &\mathbb{E}\big[d(X, f(\lambda_f)) \cdot \operatorname{sgn}\big(\cos(\theta(\lambda_f, X))\big)I(d(X, f(\lambda_f)) \le c) \\ &+c \cdot \operatorname{sgn}\big(\cos(\theta(\lambda_f, x))\big)I(d(X, f(\lambda_f)) > c)\big|\ \lambda_f(X) = \lambda\big] = 0 \quad \text{for a.e. } \lambda. \end{aligned}$$

Finally, using (A.5), (A.6), (A.7), and Lebesgue's dominated convergence theorem again,

$$\begin{aligned} &\frac{\partial \mathbb{E}\big[\rho\big(d(X, f+\epsilon g)\big)\big]}{\partial \epsilon}\Big|_{\epsilon=0} = \lim_{\epsilon \to 0}\big[\frac{\mathbb{E}\big[\rho(d(X, f+\epsilon g))\big] - \mathbb{E}\big[\rho\big(d(X, f)\big)\big]}{\epsilon}\big)\big] \\ &= \mathbb{E}\big[\lim_{\epsilon \to 0}\frac{\rho\big(d(X, f+\epsilon g)\big) - \rho\big(d(X, f)\big)}{\epsilon}\big] = \mathbb{E}_\lambda\big[\mathbb{E}\big[\lim_{\epsilon \to 0} Z(X, \epsilon) \mid \lambda_f(X) = \lambda\big]\big] \\ &= \mathbb{E}_\lambda\big[\mathbb{E}\big[2c \lim_{\epsilon \to 0} Z_1(X, \epsilon) I(d(X, f(\lambda_f)) > c) \\ &\quad + \lim_{\epsilon \to 0} Z_2(X, \epsilon) I(d(X, f(\lambda_f)) \le c) \mid \lambda_f(X) = \lambda\big]\big] \\ &= -2\mathbb{E}_\lambda\big[\,\|g(\lambda)\| \cdot |\cos(\theta(\lambda, X))| \cdot \mathbb{E}\big[d(X, f(\lambda_f)) \cdot \operatorname{sgn}\big(\cos(\theta(\lambda_f, X))\big) \\ &\quad \cdot I(d(X, f(\lambda_f)) \le c) + c \cdot \operatorname{sgn}\big(\cos(\theta(\lambda_f, X))\big) I(d(X, f(\lambda_f)) > c) \mid \lambda_f(X) = \lambda\big]\big] = 0. \end{aligned}$$

To prove the converse, suppose that

$$\begin{aligned}\mathbb{E}_\lambda\big[\,\|g(\lambda)\|\cdot|\cos(\theta(\lambda,\,X))|\cdot\mathbb{E}\big[d(X,\,f(\lambda_f))\cdot\operatorname{sgn}\big(\cos(\theta(\lambda_f,\,X))\big)I(d(X,\,f(\lambda_f))\le c)\\ +c\cdot\operatorname{sgn}\big(\cos(\theta(\lambda_f,\,X))\big)I(d(X,\,f(\lambda_f))>c)\;\big|\;\lambda_f(X)=\lambda\big]\big]=0\end{aligned}$$

for any h $(= f+g)$ such that $\|g\|<\pi$ and $\|g'\|\le 1$. Then

$$\begin{aligned}\mathbb{E}\big[d(X,\,f(\lambda_f))\cdot\operatorname{sgn}\big(\cos(\theta(\lambda_f,\,X))\big)\cdot I(d(X,\,f(\lambda_f))\le c)+c\cdot\operatorname{sgn}\big(\cos(\theta(\lambda_f,\,X))\big)\\ \cdot I(d(X,\,f(\lambda_f))>c)\;\big|\;\lambda_f(X)=\lambda\big]=0\\ \text{for a.e. }\lambda,\end{aligned}$$

which is equivalent to that f is a Huber-type principal curve of X. □

A.3 Appendix for Chapter 6

Proof of Theorem 5. The proof follows the line of Proposition 5.1 in Panaretos et al. (2014). For a set of data $\mathcal{D}=\{x_1,\,x_2,\,\dots\,x_n\}\subset\mathbb{R}^D$ and $\overline{x}=\frac{1}{n}\sum_{i=1}^n x_i\in\mathbb{R}^D$. Since $h=\infty$, local principal geodesic procedure starts at $\overline{x}$. A h-scale local tangent covariance at $\overline{x}$ is defined as $\Sigma(\overline{x})=\frac{1}{n}\sum_{i=1}^n(x_i-\overline{x})(x_i-\overline{x})^T$. Also, sample covariance at $x\in\mathbb{R}^D$ is

$$\Sigma(x)=\frac{1}{n}\sum_{i=1}^n(x_i-\overline{x}+\overline{x}-x)\cdot(x_i-\overline{x}+\overline{x}-x)^T=\Sigma(\overline{x})+(x-\overline{x})(x-\overline{x})^T.$$

Let v be the leading eigenvector of $\Sigma(\overline{x})$ (i.e., $\Sigma v=\lambda v$ and λ is the largest eigenvalue). If we prove that, for any x on the first principal component line, then the leading eigenvector of $\Sigma(x)$ is also v. Inductively, because the local principal geodesic curve does not change its direction, it will be the first principal component line. From a simple calculation, the first eigenvector and eigenvalue of $(x-\overline{x})(x-\overline{x})^T$ are $\frac{x-\overline{x}}{\|x-\overline{x}\|}$ and $\|x-\overline{x}\|^2$ respectively. Therefore, by Weyl's matrix perturbation inequality, eigenvalue of $\Sigma(x)$ is not greater than $\lambda+\|x-\overline{x}\|^2$. For simplification of notations, denote $\Sigma(\overline{x})$ and $\Sigma(x)$ as Σ and Σ' respectively. Since $x-\overline{x}$ is parallel to

v, we obtain that

$$\Sigma^{'}\cdot(x-\overline{x}) = \big(\Sigma+(x-\overline{x})(x-\overline{x})^T\big)\cdot(x-\overline{x}) = \Sigma\cdot(x-\overline{x})+\|x-\overline{x}\|^2\cdot(x-\overline{x}) = (\lambda+\|x-\overline{x}\|^2)\cdot(x-\overline{x}).$$

Therefore, v is an eigenvector of $\Sigma^{'}(=\Sigma(x))$ corresponding to the largest eigenvalue $\lambda+\|x-\overline{x}\|^2$, thus, it is the first eigenvector. In the same way, it can be shown that the direction of backward step is always -v. □

Proof of Theorem 7. Let $\mathcal{G}_k \subset \mathcal{G}$ be the collection of continuous functions that have k-geodesic segments such that

$$d(f(\lambda_1),\, f(\lambda_2)) \leq \ell|\lambda_1-\lambda_2| \quad \text{for any } \lambda_1,\, \lambda_2 \in [0,\, 1]$$

and let $\mathcal{G}_\infty \subset \mathcal{G}$ be the collection of piecewise geodesic continuous functions with their vertices at most countable. We now wish to show that $\mathcal{G}_k$ is a closed set in $\mathcal{G}$. For a sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{G}_k$ satisfying $\text{dist}(f_n,\, f) \to 0$ as $n \to \infty$, clearly $f \in \mathcal{G}$ by the uniform convergence. Moreover it can be shown that $f \in \mathcal{G}_\infty$. Once $f \in \mathcal{G}_k$ is shown, $\mathcal{G}_k$ is accordingly closed. To this end, suppose that $f \in \mathcal{G}_\infty \setminus \mathcal{G}_k$. Then f has at least $(k+1)$ geodesic segments and there are some constants $0 \leq \lambda_0 < \lambda_1 < \lambda_2 \leq 1$, $N_1 \in \mathbb{N}$ such that $f(\lambda_1)$ is a vertex of f and for any $n \geq N_1$ $f_n(\lambda_0)$, $f_n(\lambda_1)$, and $f_n(\lambda_2)$ are contained in a single geodesic segment of f_n. Denote the geodesic distance between $f(\lambda_1)$ and $f(\lambda_2)$ by ℓ_0 $(= d(f(\lambda_1),\, f(\lambda_2) \geq 0)$ and the external angle between two geodesics $\overgroup{f(\lambda_1)f(\lambda_0)}$ and $\overgroup{f(\lambda_1)f(\lambda_2)}$ by θ $(0 \leq \theta < \pi)$. Extend the geodesic segment $\overgroup{f(\lambda_0)f(\lambda_1)}$ in the direction from $f(\lambda_0)$ to $f(\lambda_1)$ by the distance ℓ_0. The end point is then called q, i.e., $d(f(\lambda_1),\, q) = \ell_0$. Choose a $N_1 \in \mathbb{N}$ such that if $n \geq N_1$ then $\text{dist}(f_n,\, f) \leq (\ell_0 \sin\frac{\theta}{2})/2$. So, if $n \geq \max\{N_1, N_2\}$, then

$$\begin{aligned} \text{dist}(f_n,\, f) \geq d(f_n(\lambda_2),\, f(\lambda_2)) &\geq d(q,\, f(\lambda_2)) - d(q,\, f_n(\lambda_2)) \\ &> c\cdot\ell_0\sin(\frac{\theta}{2}) \end{aligned} \tag{A.12}$$

where c is some constant related to the assumption (B2), the sign of curvatures on M. Note that the above inequality holds by triangle inequality and same argument in Lemmas 8 and 9. (It involves a procedure using Toponogov and Rauch comparison theorems and then the computation of constant c is just tedious; hence, we omit the procedure.) (A.12) implies that $\mathrm{dist}(f_n, f) \nrightarrow 0$ as $n \to \infty$. It is a contradiction and then $\mathcal{G}_k$ is therefore closed. The remaining proof is same to that of Theorem 6. □

Proof of Theorem 8. The proof proceeds in the similar manner of Theorem 1 in Kégl (1999); Kégl et al. (2000), while there are some difficulties in an attempt to extend the theorem from Euclidean space to generic Riemannian manifold with same sign of curvature. To this end, it needs to prove the existence of *finite* ϵ-cover of $\mathcal{F}_k$ for any $k \geq 1$ and $\epsilon > 0$. Specifically, for any $k \geq 1$ and $\epsilon > 0$, $\mathcal{F}_{k,\,\epsilon}$ is a nonempty and finite collection of curves N, i.e. $\mathcal{F}_{k,\,\epsilon} \subset \mathcal{F}_k$, and is an ϵ-cover of $\mathcal{F}_k$. More precisely, to prove Theorem 8 we first prove the following Lemma which is concerned with convergence rate of the proposed procedure.

Lemma 7 (ϵ-covering property)**.** *Under $(B1) - (B2)$, for any $k \geq 1$ and $\epsilon > 0$, there exists a finite collection of curves $\mathcal{F}_{k,\,\epsilon}$ with $\phi \neq \mathcal{F}_{k,\,\epsilon} \subset \mathcal{F}_k$, satisfying the covering property: For any $f \in \mathcal{F}_k$, there exists $g \in \mathcal{F}_{k,\,\epsilon}$ such that*

$$\sup_{x \in N} |d^2(x,\, f) - d^2(x,\, g)| < \epsilon$$

The number of elements in $\mathcal{F}_{k,\,\epsilon}$ is bounded above by some constant which is not dependent on n; specifically,

$$|\mathcal{F}_{k,\,\epsilon}| \leq 2^{\frac{2c_0c_1r\ell}{\epsilon}+(c_1+2)k} V_D^{k+1} \Big(\frac{2c_0c_1\sqrt{D}r^2}{\epsilon}+\sqrt{D}\Big)^D \Big(\frac{2c_0^2c_1\sqrt{D}r\ell}{k\epsilon}+\big(c_0(c_1+1)+1\big)\sqrt{D}\Big)^{Dk}, \tag{A.13}$$

where r, D, V_D, c_0, and c_1 denote the diameter of N, dimension of M, the volume of the unit sphere in $\mathbb{R}^D$, the constant stated later in Lemma 8, and a Lipshitz constant of exp*, respectively.*

Proof of Lemma 7. Lemma 7 can be proved in the similar way in Lemma 2 of Kégl (1999) with some modifications. For the adaptation, the following Lemmas are inevitable to prove Lemma 7.

Lemma 8. *Under $(B1)-(B2)$, let $p \in M$ be a point in M and $v,\, w \in T_pM$ be tangent vectors with $\|v\|,\, \|w\| < \pi/\sqrt{\kappa}$. (In the case (a) of $(B2)$, $\kappa = 0$). There exists some constant $c_0 > 0$ such that*

$$\|v - w\| \leq c_0 d(\exp_p v,\, \exp_p w).$$

In the case (a) of $(B2)$, $c_0 = 1$. That is, $\log_p$ is Lipshitz and c_0 is the corresponding Lipshitz constant.

Proof of Lemma 8. *Case 1. Manifolds of nonpositive curvatures.* Suppose that M satisfies (a) of (B2). By applying Rauch comparison theorem (e.g., Chapter 11 of Lee (2006)) to the case that any sectional curvatures are not greater than zero, i.e. $K \leq 0$, we obtain that for any $v,\, w \in T_pM$

$$\|v - w\| \leq d(\exp_p v,\, \exp_p w).$$

In this case, $c_0 = 1$.

Case 2. Manifolds of strictly positive curvatures. Suppose that (b) of (B2) holds. On Riemannian manifolds of bounded sectional curvatures, comparison theorems of Rauch and Toponogov (e.g., see Chapter 11 of Lee (2006) for details) are useful tool to compare the deviations of geodesics emanating from a one point p. Specifically, denote $d(\exp_p vt,\, \exp_p wt)$ by $L_0(t)$, the angle between v and w by $\theta \in [0,\, \pi]$, and Riemannian distance between $\overline{\exp_p} vt$ and $\overline{\exp_p} wt$ in D-dimensional sphere with radius $1/\sqrt{\kappa}$ (that has constant sectional curvature κ) by $L_1(t)$, where $\overline{\exp}$ is the exponential map of the sphere. Specifically, by some calculation, we get

$$L_1(t) = 1/\sqrt{\kappa}\arccos\big(\sin(vt\sqrt{\kappa})\sin(wt\sqrt{\kappa})\cos(\theta) + \cos(vt\sqrt{\kappa})\cos(wt\sqrt{\kappa})\big).$$

By the Rauch- and Toponogov comparision theorems with $0 < K \leq \kappa$,

$$0 \leq L_1(t) \leq L_0(t) \leq t\,\|v - w\| \;\; (=: L_2(t)) \quad \text{for } 0 \leq t \leq 1.$$

By some calculation, it can be shown that

$$\frac{L_2(1)}{L_1(1)} \leq \sqrt{2}\pi/(\pi - \sqrt{\kappa}\cdot \mathrm{diam}(N)) \; (=: c_0),$$

which implying that $\frac{L_2(1)}{L_0(1)} \leq c_0$. Therefore,

$$\|v - w\| \leq c_0 d(\exp_p v, \exp_p w).$$

□

Lemma 9. *Under* $(B1)-(B2)$*, if* γ, $\eta : [0, 1] \to M$ *are two geodesics parametrized by constant speeds, then for any* $\lambda \in [0, 1]$

$$d(\gamma(\lambda), \eta(\lambda)) \leq 2c_0 \max\{d(\gamma(0), \eta(0)), d(\gamma(1), \eta(1))\},$$

where c_0 *is the constant stated in Lemma 8.*

Proof of Lemma 9. *Case 1. Manifolds of nonpositive curvatures.*
Suppose that M satisfies (a) of (B2). Since M is connected and complete, M is geodesically complete by Hopf-Rinow theorem (e.g., Chapter 6, pages 108–111, in Lee (2006)). So, there is a geodesic $\nu : [0, 1] \to M$ joining $\gamma(0)$, $\eta(1)$ parametrized by a constant speed. Note that any geodesics emanating from a one point spread out. In this regard, for any $\lambda \in [0, 1]$

$$\begin{aligned} d(\gamma(\lambda), \eta(\lambda)) &\leq d(\gamma(\lambda), \nu(\lambda)) + d(\nu(\lambda), \eta(\lambda)) \\ &\leq d(\gamma(0), \eta(0)) + d(\eta(1), \gamma(1)) \\ &= 2\max\{d(\gamma(0), \eta(0)), d(\gamma(1), \eta(1))\}. \end{aligned}$$

As mentioned in Lemma 8, $c_0 = 1$ in this case.
Case 2. Manifolds of strictly positive curvatures.
Suppose that M satisfies (b) of (B2). Following the notations in the proof of Lemma 8, we note that

$$\frac{L_2(1)}{L_1(1)} \leq c_0.$$

Due to the monotonicity of $L_2(\cdot)$, we get

$$L_0(t) \leq L_2(1) \leq c_0 L_1(1) \leq c_0 L_0(1) \quad \text{for any } 0 \leq t \leq 1. \tag{A.14}$$

Since M is geodesically complete by Hopf-Rinow theorem (e.g., Chapter 6, 108–111, in Lee (2006)), there is a geodesic $\nu : [0,\, 1] \to M$ joining $\gamma(0)$, $\eta(1)$ parametrized by a constant speed with $\nu(0) = \eta(0)$, $\nu(1) = \eta(1)$. Using (A.14), we obtain that for any $\lambda \in [0,\, 1]$

$$\begin{aligned} d(\gamma(\lambda),\, \eta(\lambda)) &\leq d(\gamma(\lambda),\, \nu(\lambda)) + d(\nu(\lambda),\, \eta(\lambda)) \\ &\leq c_0[d(\gamma(1),\, \eta(1)) + d(\gamma(0),\, \eta(0))] \\ &= 2c_0 \max\{d(\gamma(0),\, \eta(0)),\, d(\gamma(1),\, \eta(1))\}, \end{aligned}$$

as desired.

Remark 1. *If M is simply connected (meaning that roughly speaking M has no "holes") and is a manifold of nonpositive curvatures (Case 1.), it is a immediate result from geodesic comparison inequality (Corollary 2.5 and Proposition 3.1 in Sturm (2003)) that for any $\lambda \in [0,\, 1]$*

$$\begin{aligned} d(\gamma(\lambda),\, \eta(\lambda)) &\leq \lambda d(\gamma(0),\, \eta(0)) + (1-\lambda) d(\gamma(1),\, \eta(1)) \\ &\leq \max\{d(\gamma(0),\, \eta(0)),\, d(\gamma(1),\, \eta(1))\}. \end{aligned}$$

In this case, the constant in the right-hand side is thus halved.

□

We now enter into the main proof of Lemma 7. Notice that N is closed and bounded from the assumption (B1); thus, N is compact by Hopf-Rinow theorem. For any $p \in N$, $N \subset B_p(r)$ where $r = \text{diam}(N)$. One can show that $\exp_p$ is a Lipshitz function by using the similar argument in Lemma 9. Fix a point $p \in N$ and let $c_1 > 0$ be the Lipshitz constant of $\exp_p : \log_p[B_p(r)] \to B_p(r)$. It means that for any $x,\, y \in \log_p[B_p(r)] := \{z \in T_pM \mid \|z - p\| < r\}$

$$d(\exp_p x,\, \exp_p y) \leq c_1 \|x - y\|, \tag{A.15}$$

where $\|\cdot\|$ is the norm in $T_pM \simeq \mathbb{R}^D$. In the case (b) of (B2), $c_1 = 1$ by Toponogov comparison theorem. For a given $\epsilon > 0$, let $\delta = \epsilon/(2rc_0c_1\sqrt{D})$ and consider the rectangular grid that is centered at p and has side length δ in $T_pM \simeq \mathbb{R}^D$. An illustration

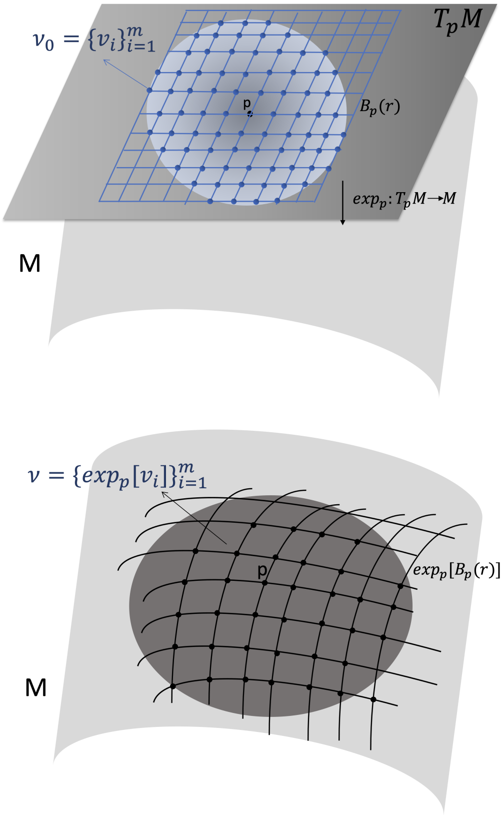

$\mathcal{V}_0 = \{v_i\}_{i=1}^m$
T_pM
p
$B_p(r)$
$exp_p: T_pM \rightarrow M$
M
$\mathcal{V} = \{exp_p[v_i]\}_{i=1}^m$
p
$exp_p[B_p(r)]$
M

Figure A.1: In the proof of Lemma 7, the configuration of M, T_pM, $\mathcal{V}_0$ (blue dots) and $\mathcal{V}$ (black dots) is illustrated.

of the configuration is given in Figure A.1. On this grid, let $\mathcal{V}_0 = \{v_i\}_{i=1}^m$ be the set of vertices of the grid whose distances to the set $\log_p[N]$ are not greater than $\delta\sqrt{D}$. Notice that the vertices inside $\log_p[N]$ is clearly included in $\mathcal{V}_0$ by construction. Denote $\mathcal{V} = \exp_p[\mathcal{V}_0] := \big\{\exp_p x \in M \,|\, x \in \mathcal{V}_0\big\} \subset B_p(r)$ in which c_0 is the constant stated in Lemma 8. For any $f \in \mathcal{F}_k$ and each $0 \leq i \leq k$, let y_i be the vertex of f in sequence and $\hat{y}_i$ be any closest point of y_i among $\mathcal{V}$, i.e., $\hat{y}_j = \operatorname{argmin}_{y \in \mathcal{V}} d(y_j, y)$. For each $0 \leq i \leq k$, there exists $\tilde{v}_i \in \mathcal{V}_0$ such that

$$\left\|\log_p y_i - \tilde{v}_i\right\| < \delta\sqrt{D}/2.$$

It follows that $\exp_p \tilde{v}_i \in \mathcal{V}$ by construction. From the definition of $\hat{y}_i$ and (A.15),

$$d(y_i,\, \hat{y}_i) \leq d(y_i,\, \exp_p \tilde{v}_i) \leq c_1 \left\| \log_p y_i - \tilde{v}_i \right\| < c_1 \delta \sqrt{D}/2 \quad \text{for any } 0 \leq i \leq k. \tag{A.16}$$

Let $\hat{f}$ be the piecewise-geodesic curve joining $\hat{y}_0,\, \hat{y}_1,\, \ldots,\, \hat{y}_k$ in sequence by geodesic segments. Denote the length of $\hat{f}$ by $L(\hat{f})$. Then by triangular inequality,

$$L(\hat{f}) \leq \ell + c_1 k \delta \sqrt{D}. \tag{A.17}$$

Meanwhile, for any finite-length continuous curves $g : I_1 \to M$ and $h : I_2 \to M$ with intervals $I_1,\, I_2 \subset \mathbb{R}$, (asymmetric) Hausdorff distance between the curves, say $d_H(g,\, h)$, is defined as

$$d_H(g,\, h) = \max_{\lambda \in I_1} \min_{\mu \in I_2} d(g(\lambda),\, h(\mu))$$

By the definition of d_H,

$$d_H(f,\, \hat{f}) \leq \max_{0 \leq i \leq k-1} d_H(\overline{y_i \hat{y}_i},\, \overline{y_{i+1} \hat{y}_{i+1}})$$

where $\overline{AB}$ denotes the geodesic segment joining A and B. Since $d_H(g,\, h)$ is invariant under parametrizations of g and h, without loss of generality we may assume that the geodesic segments $\overline{y_i \hat{y}_i}$, $\overline{y_{i+1} \hat{y}_{i+1}}$ are parametrized by respective constant speeds over $[0,\, 1]$. By (A.16) and Lemma 9, we obtain that

$$d_H(\overline{y_i \hat{y}_i},\, \overline{y_{i+1} \hat{y}_{i+1}}) \leq \operatorname{dist}(\overline{y_i \hat{y}_i},\, \overline{y_{i+1} \hat{y}_{i+1}}) < c_0 c_1 \delta \sqrt{D},$$

where $\operatorname{dist}(\cdot,\, \cdot)$ is defined in (6.6). We thus get $d_H(f,\, \hat{f}) < c_0 c_1 \delta \sqrt{D}$ and $d_H(\hat{f},\, f) < c_0 c_1 \delta \sqrt{D}$ also follows by symmetry. Using (A.26), we get

$$\sup_{x \in N} |d^2(x,\, f) - d^2(x,\, \hat{f})| \leq 2r \cdot \max \left\{ d_H(f,\, \hat{f}),\, d_H(\hat{f},\, f) \right\} < 2r c_0 c_1 \delta \sqrt{D} = \epsilon.$$

Define $\mathcal{F}_{k,\,\epsilon}$ as the family of piecewise-geodesic curves whose lengths are not greater than $\ell + c_1 k \delta \sqrt{D}$ with $(k+1)$-vertices (not necessarily different) being contained in $\mathcal{V}$. By (A.17), we get $\hat{f} \in \mathcal{F}_{k,\,\epsilon}$. We now wish to count the number of elements in $\mathcal{F}_{k,\,\epsilon}$, say $|\mathcal{F}_{k,\,\epsilon}|$, or to find its upper bound. To this end, choose a function $J \in \mathcal{F}_{k,\,\epsilon}$.

Then $L(J) \leq \ell + c_1 k\delta\sqrt{D}$ in which $l(J)$ denotes the length of J. For each $0 \leq j \leq k$, denote the j-th vertex of J by v_j and for each $0 \leq i \leq k-1$, denote the length of i-th geodesic segment by ℓ_i $(= d(v_i,\, v_{i+1}))$. After that, define

$$\hat{\ell}_i = \lceil \ell_i/(\delta\sqrt{D}) \rceil \delta\sqrt{D},$$

where $\lceil\cdot\rceil$ denotes the ceiling function; that is, $\lceil \ell \rceil$ is the least integer no smaller than ℓ and it holds that $\ell_i \leq \hat{\ell}_i < \ell_i + \delta\sqrt{D}$. Note that the number of vertices on the grid that are contained in any closed ball (in $\mathbb{R}^D$) centered at a vertex and with radius $L \geq 0$ is bounded by

$$V_D\big(\frac{L+\delta\sqrt{D}}{\delta}\big)^D \tag{A.18}$$

where V_D denotes the volume of the unit sphere in $\mathbb{R}^D$ and this is known as Gauss sphere problem. Fix the sequence $\hat{\ell}_0,\, \hat{\ell}_1,\, \ldots,\, \hat{\ell}_{k-1}$ and now count the number of functions $J \in \mathcal{F}_{k,\,\epsilon}$ generating $\hat{\ell}_0,\, \hat{\ell}_1,\, \ldots,\, \hat{\ell}_{k-1}$. The number of cases for choosing the first vertex v_0 equals to $|\mathcal{V}|$. According to (A.18),

$$|\mathcal{V}| \leq |\mathcal{V}_0| \leq V_D\big(\frac{r+\delta\sqrt{D}}{\delta}\big)^D. \tag{A.19}$$

We denote $\bar{B}_\cdot(\cdot)$ and $\bar{\bar{B}}_\cdot(\cdot)$ by the closed balls in M and $T_pM \simeq \mathbb{R}^D$, respectively. Inductively, for each $0 \leq i \leq k-1$, suppose that $v_0,\, v_1,\, \ldots,\, v_i$ are determined. By Lemma 8, for any $x \in \bar{B}_{v_i}(\hat{\ell}_i)$

$$\big\|\log_p x - \log_p v_i\big\| \leq c_0 d(x,\, v_i) \leq c_0\hat{\ell}_i,$$

thereby meaning that

$$\log_p[\bar{B}_{v_i}(\hat{\ell}_i)] \subset \bar{\bar{B}}_{\log_p v_i}(c_0\hat{\ell}_i). \tag{A.20}$$

By (A.18), (A.20), the number of cases for choosing v_{i+1} among $\mathcal{V} \cap \bar{B}_{v_i}(\hat{\ell}_i)$ is bounded by

$$V_D\big(\frac{c_0\hat{\ell}_i+\delta\sqrt{D}}{\delta}\big)^D \quad \text{for all } 0 \leq i \leq k-1. \tag{A.21}$$

Meanwhile, by arithmetic and geometric means inequality (AM-GM),

$$\prod_{i=0}^{k-1}[\frac{c_0\hat{\ell}_i}{\delta}+\sqrt{D}] \overset{\text{“AM-GM”}}{\leq} [\sum_{i=0}^{k-1}\frac{\frac{c_0\hat{\ell}_i}{\delta}+\sqrt{D}}{k}]^k \overset{\text{“}\hat{\ell}_i<\ell_i+\delta\sqrt{D}\text{”}}{<} [\sum_{i=0}^{k-1}\frac{\frac{c_0\ell_i}{\delta}+(c_0+1)\sqrt{D}}{k}]^k, \tag{A.22}$$

in which the last inequality holds by $\hat{\ell}_i < \ell_i + \delta\sqrt{D}$ for any $0 \leq i \leq k-1$. By combining (A.18), (A.19), (A.21) and (A.22), the number of such J with $\hat{\ell}_0, \hat{\ell}_1, \ldots, \hat{\ell}_{k-1}$ is bounded by

$$\begin{aligned} V_D\Big(\frac{r+\delta\sqrt{D}}{\delta}\Big)^D \prod_{i=0}^{k-1}\big[V_D\big(\frac{c_0\hat{\ell}_i+\delta\sqrt{D}}{\delta}\big)^D\big] \quad &= \quad V_D^{k+1}(\frac{r}{\delta}+\sqrt{D})^D\prod_{i=0}^{k-1}[\frac{c_0\hat{\ell}_i}{\delta}+\sqrt{D}]^D \\ &\overset{(A.22)}{\leq} \quad V_D^{k+1}(\frac{r}{\delta}+\sqrt{D})^D\big[\sum_{i=0}^{k-1}\frac{\frac{c_0\ell_i}{\delta}+(c_0+1)\sqrt{D}}{k}\big]^{Dk} \\ &\leq \quad V_D^{k+1}(\frac{r}{\delta}+\sqrt{D})^D\Big[\frac{c_0\ell}{k\delta}+(c_0+c_0c_1+1)\sqrt{D}\Big]^{Dk} \end{aligned} \tag{A.23}$$

where the first and second inequalities hold by (A.22) and $\sum_i \ell_i = L(J) \leq \ell + c_1 k\delta\sqrt{D}$, respectively.

The remaining part is to derive an upper bound of the number of the curves J generating $\hat{\ell}_0, \hat{\ell}_1, \ldots, \hat{\ell}_{k-1}$. From $\sum_i \hat{\ell}_i < \sum_i(\ell_i + \delta\sqrt{D}) = L(J) + k\delta\sqrt{D} \leq \ell + (c_1+1)k\delta\sqrt{D}$,

$$\sum_{i=0}^{k-1}\underbrace{\frac{\hat{\ell}_i}{\delta\sqrt{D}}}_{=:L_i\in\mathbb{N}} < \lceil\frac{\ell}{\delta\sqrt{D}}+c_1k\rceil + k.$$

Therefore, the number of the such J generating $\hat{\ell}_0, \hat{\ell}_1, \ldots, \hat{\ell}_{k-1}$ is same to that of nonnegative integer solutions of the following equation:

$$L_0 + L_1 + \ldots + L_{k-1} \leq \lceil\frac{\ell}{\delta\sqrt{D}}+c_1k\rceil + k - 1.$$

To count the number of nonnegative integer solutions of above, by means of combination with repetition we get the number of solutions and its upper bound, as

$$\sum_{m=k-1}^{\lceil\frac{\ell}{\delta\sqrt{D}}+c_1k\rceil+2k-2}\binom{n}{k} = \binom{\lceil\frac{\ell}{\delta\sqrt{D}}+c_1k\rceil+2k-1}{k} \leq 2^{\lceil\frac{\ell}{\delta\sqrt{D}}+c_1k\rceil+2k-1}. \tag{A.24}$$

Combining (A.23), (A.24) and putting $\delta = \epsilon/(2rc_0c_1\sqrt{D})$, we finally get

$$\begin{aligned}
|\mathcal{F}_{k,\,\epsilon}| &\leq 2^{\lceil \frac{\ell}{\delta\sqrt{D}} + c_1 k\rceil + 2k-1} \cdot V_D^{k+1}\Big(\frac{r}{\delta} + \sqrt{D}\Big)^D \Big(\frac{c_0\ell}{k\delta} + (c_0 + c_0c_1 + 1)\sqrt{D}\Big)^{Dk} \\
&= 2^{\lceil \frac{2c_0c_1r\ell}{\epsilon} + c_1 k\rceil + 2k-1} \cdot V_D^{k+1}\Big(\frac{2c_0c_1\sqrt{D}r^2}{\epsilon} + \sqrt{D}\Big)\Big(\frac{2c_0^2c_1\sqrt{D}r\ell}{k\epsilon} + \big(c_0(c_1+1)+1\big)\sqrt{D}\Big)^{Dk} \\
&\leq 2^{\frac{2c_0c_1r\ell}{\epsilon} + (c_1+2)k} \cdot V_D^{k+1}\Big(\frac{2c_0c_1\sqrt{D}r^2}{\epsilon} + \sqrt{D}\Big)\Big(\frac{2c_0^2c_1\sqrt{D}r\ell}{k\epsilon} + \big(c_0(c_1+1)+1\big)\sqrt{D}\Big)^{Dk},
\end{aligned}$$

as desired. So the proof of Lemma 7 ends.

Remark 2. *When $M = \mathbb{R}^D$, the Euclidean version of Lemma 9 can be easily obtained as $\|\gamma(\lambda) - \eta(\lambda)\| \leq \max\{\|\gamma(0) - \eta(0)\|, \|\gamma(1) - \eta(1)\|\}$ by some calculation. In this case, the consequence of Lemma 7 becomes to be that of Lemma 2 in Kégl (1999); Kégl et al. (2000). In this fashion, Lemma 7 is a generalization of Lemma 2 in Kégl (1999); Kégl et al. (2000) to generic Riemannian manifolds with same sign of curvatures.*

□

We now enter into the main body of proof in Theorem 8. Recall that

$$f_k^* = \mathrm{argmin}_{f\in\mathcal{F}_k} \mathbb{E}d^2(X,\, f) = \mathrm{argmin}_{f\in\mathcal{F}_k} R(f).$$

The excess risk of $f_{k,\,n}$, $\mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n})] - R(f^*)$, is separated into the two parts as follows:

$$\mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n})] - R(f^*) = [\overbrace{\mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n})] - R(f_k^*)}^{\text{estimation error}}] + [\overbrace{R(f_k^*) - R(f^*)}^{\text{approximation error}}].$$

We then wish to bound the approximation and estimation errors respectively. The latter is more technical than the former.

Step 1: Approximation error

To obtain an upper bound of the approximation error, define the (asymmetric) Hausdorff distance between two continuous functions $f,\, g : [0,\, 1] \to M$ as

$$d_H(f,\, g) = \max_{\lambda\in[0,\,1]} \min_{\mu\in[0,\,1]} d(f(\lambda),\, g(\mu)). \tag{A.25}$$

Suppose that f^* is parametrized by some constant speed with $L(f^*) \leq \ell$ and let $g : [0,\, 1] \to M$ be the piecewise-geodesic curve sequentially joining

$$\{f^*(0),\, f^*(1/k),\, f^*(2/k),\, \ldots,\, f^*((k-1)/k),\, f^*(1)\}$$

by geodesic segments. By construction, it follows from $L(g) \leq L(f^*) \leq \ell$ that $g \in \mathcal{F}_k$. For any $\lambda \in [0,\, 1]$, there is an index $1 \leq i \leq k$ such that $(i-1)/k \leq \lambda \leq i/k$. We thus get

$$\begin{aligned} \min_{\mu \in [0,\, 1]} d(f^*(\lambda),\, g(\mu)) &\leq d(f^*(\lambda),\, g((i-1)/k)) = d(f^*(\lambda),\, f^*((i-1)/k)) \\ &\leq L(f^*)/k \\ &\leq \ell/k, \end{aligned}$$

thereby leading to

$$d_H(f^*,\, g) = \max_{\lambda \in [0,\, 1]} \min_{\mu \in [0,\, 1]} d(f^*(\lambda),\, g(\mu)) \leq \ell/k.$$

For a given $x \in N$, denote $\lambda_{f^*}(x)$ by λ for simplicity and choose a $\mu \in [0,\, 1]$ such that $d(f^*(\lambda),\, g(\mu)) = \min_{\nu \in [0,\, 1]} d(f^*(\lambda),\, g(\nu))$ where it can be possible because $[0,\, 1]$ is compact. By the definition of d_H and triangle inequality, it holds from $\operatorname{diam}(N) = r$ that

$$\begin{aligned} d^2(x,\, g) - d^2(x,\, f^*) &\leq d^2(x,\, g(\mu)) - d^2(x,\, f^*(\lambda)) = 2r[d(x,\, g(\mu)) - d(x,\, f^*(\lambda))] \\ &\leq 2r \cdot d(f^*(\lambda),\, g(\mu)) \\ &\leq 2r \cdot d_H(f^*,\, g). \end{aligned} \tag{A.26}$$

Because (A.26) holds true for any $x \in N$, we get an upper bound of the approximation error as

$$\begin{aligned} R(f_k^*) - R(f^*) &\leq R(g) - R(f^*) = \mathbb{E}[d^2(X,\, g) - d^2(X,\, f^*)] \\ &\leq 2r \cdot d_H(f^*,\, g) \\ &\leq 2r\ell/k. \end{aligned} \tag{A.27}$$

Step 2: Estimation error

$\mathcal{X}_n = \{X_1,\, X_2,\, \ldots,\, X_n\}$ and X are identically distributed and independent (i.i.d.). Note that, for any real-valued random variable Z,

$$\mathbb{E}Z = \mathbb{E}Z_+ - \mathbb{E}Z_- \leq \mathbb{E}Z_+ = \int_0^\infty \mathbf{P}(Z_+ > u)du = \int_0^\infty \mathbf{P}(Z > u)du, \qquad (A.28)$$

where $Z_+ = \max\{Z,\, 0\} \geq 0$, $Z_- = -\min\{Z,\, 0\} \geq 0$ and $Z = Z_+ - Z_-$. Note also that $R(f_{k,\,n}) = \mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n]$ is random since $f_{k,\,n}$ is the function of $\mathcal{X}_n$. Then, by using (A.28),

$$\begin{aligned} \mathbb{E}_{\mathcal{X}_n}[R(f_{k,\,n})] - R(f_k^*) &= \mathbb{E}_{\mathcal{X}_n}\mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n] - \mathbb{E}d^2(X,\, f_k^*) \\ &= \mathbb{E}_{\mathcal{X}_n}\big[\mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n] - \mathbb{E}d^2(X,\, f_k^*)\big] \\ &\leq \int_0^\infty \mathbf{P}\big(\mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n] - \mathbb{E}d^2(X,\, f_k^*) > u\big)du. \end{aligned} \qquad (A.29)$$

We now wish to bound (A.29). For a given $\epsilon > 0$, via Lemma 7, there exists a $g \in \mathcal{F}_{k,\,\epsilon}$ such that $\sup_{x\in N} |d^2(x,\, f_{k,\,n}) - d^2(x,\, g)| < \epsilon$. It implies that $d^2(x,\, f_{k,\,n}) < d^2(x,\, g) + \epsilon$ for any $x \in N$ and that with probability one

$$\frac{1}{n}\sum_{i=1}^n d^2(X_i,\, g) - \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, f_{k,\,n}) < \epsilon$$

from $\mathbf{P}(\mathcal{X}_n \subset N) = 1$. From these facts, with probability one we get

$$\begin{aligned} & R(f_{k,n}) - R(f^*) \\ = \;& \mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n] - \mathbb{E}d^2(X,\, f_k^*) \\ = \;& \mathbb{E}[d^2(X,\, f_{k,\,n}) \,|\, \mathcal{X}_n] - \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, f_{k,\,n}) + \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, f_{k,\,n}) - \mathbb{E}d^2(X,\, f_k^*) \\ < \;& 2\epsilon + \mathbb{E}[d^2(X,\, g) \,|\, \mathcal{X}_n] - \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, g) + \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, f_{k,\,n}) - \mathbb{E}d^2(X,\, f_k^*) \\ \leq \;& 2\epsilon + \mathbb{E}d^2(X,\, g) - \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, g) + \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, f_k^*) - \mathbb{E}d^2(X,\, f_k^*) \\ \leq \;& 2\epsilon + 2 \max_{h\in\mathcal{F}_{k,\,\epsilon}\cup\{f_k^*\}} |\mathbb{E}d^2(X,\, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i,\, h)|, \end{aligned} \qquad (A.30)$$

where second inequality holds by $\mathbb{E}[d^2(X, g) \,|\, \mathcal{X}_n] = \mathbb{E}d^2(X, g)$ (since $\mathcal{X}_n$ and X are independent) and (6.11). To bound (A.29), Hoeffding's inequality should be applied. The Hoeffding inequality (e.g., Chapter 3.5 of Van de Geer (2000)) states that for any $t \geq 0$ and independent real-valued random variables $Z_1, Z_2, \ldots Z_n$ satisfying $a_i \leq Z_i \leq b_i$, $1 \leq i \leq n$ and $\sum_{i=1}^n (a_i - b_i)^2 > 0$

$$\mathbf{P}\big(|\mathbb{E}Z_1 - \frac{1}{n}\sum_{i=1}^n Z_i| > t\big) \leq 2\exp[-2n^2t^2/\sum_{i=1}^n (a_i - b_i)^2]. \tag{A.31}$$

For a given $h \in \mathcal{F}_k$, let $Z_i = d^2(X_i, h) \leq r^2$, $1 \leq i \leq n$. By (A.31), for any $t \geq 0$

$$\mathbf{P}\Big(|\mathbb{E}d^2(X, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i, h)| > t\Big) \leq 2\exp[-2nt^2/r^4]. \tag{A.32}$$

According to (A.30), for any $u \geq 2\epsilon$, it holds that

$$\begin{aligned}
&\mathbf{P}\big(\mathbb{E}[d^2(X, f_{k,n}) \,|\, \mathcal{X}_n] - \mathbb{E}d^2(X, f_k^*) > u\big) \\
&\quad \leq \quad \mathbf{P}\big(\max_{h \in \mathcal{F}_{k,\epsilon} \cup \{f_k^*\}} |\mathbb{E}d^2(X, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i, h)| > u/2 - \epsilon\big) \\
&\quad = \quad \mathbf{P}\Big(\bigcup_{h \in \mathcal{F}_{k,\epsilon} \cup \{f_k^*\}} \Big\{|\mathbb{E}d^2(X, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i, h)| > u/2 - \epsilon\Big\}\Big) \\
&\quad \leq \quad \sum_{h \in \mathcal{F}_{k,\epsilon} \cup \{f_k^*\}} \mathbf{P}\Big(|\mathbb{E}d^2(X, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i, h)| > u/2 - \epsilon\Big) \\
&\quad \leq \quad (|\mathcal{F}_{k,\epsilon}| + 1) \cdot \max_{h \in \mathcal{F}_{k,\epsilon} \cup \{f_k^*\}} \mathbf{P}\Big(|\mathbb{E}d^2(X, h) - \frac{1}{n}\sum_{i=1}^n d^2(X_i, h)| > u/2 - \epsilon\Big) \\
&\quad \overset{(A.32)}{\leq} \quad 2(|\mathcal{F}_{k,\epsilon}| + 1)\exp[-2n \cdot (u/2 - \epsilon)^2/r^4] \\
&\quad = \quad 2(|\mathcal{F}_{k,\epsilon}| + 1)\exp[-n(u - 2\epsilon)^2/(2r^4)],
\end{aligned} \tag{A.33}$$

where $|\mathcal{F}_{k,\epsilon}|$ denotes the number of elements in $\mathcal{F}_{k,\epsilon}$ and the last inequality holds true by (A.32). Meanwhile, according to (A.13) in Lemma 7 with choosing $\epsilon = 1/k$, it can be shown that there exists a constant $C(\ell, r)$ that is independent on both k and n such that

$$2r^4 \log 2(|\mathcal{F}_{k,\epsilon}| + 1) \leq C(\ell, r)k. \tag{A.34}$$

Due to the fact that $(-e^{-x^2}/2x)' = e^{-x^2} + e^{-x^2}/2x^2 > e^{-x^2}$ for $x > 0$, we obtain, by the integration from $x = t$ to $x = \infty$ with respect to x, that

$$\int_t^\infty e^{-x^2} dx < e^{-t^2}/2t \quad \text{for any } t > 0. \tag{A.35}$$

By plugging (A.33) into (A.29), letting $v = \sqrt{2r^4 \log[2(|\mathcal{F}_{k,\epsilon}| + 1)]/n} \geq 0$, and using (A.35), we get

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_n}[R(f_{k,n})] - R(f_k^*) &\leq v + 2\epsilon + 2(|\mathcal{F}_{k,\epsilon}| + 1) \int_{v+2\epsilon}^\infty \exp(-n(u-2\epsilon)^2/(2r^4))du \\
&\overset{x=\sqrt{n/2}(u-2\epsilon)/r^2}{=} v + 2\epsilon + 2(|\mathcal{F}_{k,\epsilon}| + 1)\sqrt{2/n}r^2 \int_{\sqrt{n/2}v/r^2}^\infty \exp(-x^2)dx \\
&\overset{\text{(A.35)}}{\leq} v + 2\epsilon + \underbrace{2(|\mathcal{F}_{k,\epsilon}| + 1) \cdot r^4/(nv) \cdot \exp(-nv^2/(2r^4))}_{=O(n^{-1/2})} \\
&= \sqrt{2r^4 \log[2(|\mathcal{F}_{k,\epsilon}| + 1)]/n} + 2\epsilon + O(n^{-1/2}) \quad \text{as } n \to \infty,
\end{aligned}$$

where the last equality holds by

$$r^4/(nv) = r^4/\sqrt{2nr^4 \log 2(|\mathcal{F}_{k,\epsilon}| + 1)} \leq r^2/(\sqrt{2n}\log 3) = O(n^{-1/2}) \quad \text{as } n \to \infty.$$

Finally, by the *Step 1* and *Step 2*

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}_n}[R(f_{k,n})] - R(f^*) &\leq \underbrace{\sqrt{C(\ell, r)k/n} + 2(r\ell + 1)/k}_{\text{leading term}} + O(n^{-1/2}) \\
&= O(n^{-1/3}) \quad \text{as } n \to \infty,
\end{aligned}$$

where the last equality holds by balancing the first two terms (approximation and estimation errors) with respect to $n \to \infty$ and the optimal asymptotic order of k, $k \asymp n^{1/3}$, is achieved. Therefore the proof of Theorem 8 finally ends. □

Proof of Theorem 9. In the proof of Theorem 8, by (A.33)

$$\mathbf{P}\big(R(f_{k,n}) - R(f^*) > u\big) \leq 2(|\mathcal{F}_{k,\epsilon}| + 1)\exp[-n(u-2\epsilon)^2/(2r^4)].$$

It suffices to find any $u \geq 2\epsilon$ such that

$$2(|\mathcal{F}_{k,\epsilon}| + 1)\exp[-n(u-2\epsilon)^2/(2r^4)] \leq \delta, \tag{A.36}$$

By letting $\epsilon = 1/k$, (A.36) is equivalent to

$$\sqrt{[2r^4 \log 2(|F_{k,\epsilon}| + 1) - 2r^4 \log \delta]/n} + 2/k \leq u.$$

By $2r^4 \log 2(|F_{k,\epsilon}| + 1) \leq C(\ell, r)k$, a sufficient condition for (A.36) is

$$\sqrt{[C(\ell, r)k - 2r^4 \log \delta]/n} + 2/k = u.$$

Using the approximation bound (A.27), the result follows as desired. □

Bibliography

Adler, D. and Murdoch, D. (2020). ***rgl**: 3D Visualization Using OpenGL*. R package version 0.100.50.

Afsari, B. (2011). Riemannian L^p center of mass: existence, uniqueness, and convexity. In *Proceedings of the American Mathematical Society*, volume 139, pages 655–673.

Arias-Castro, E. and Donoho, D. L. (2009). Does median filtering truly preserve edges better than linear filtering? *Annals of Statistics*, 37(3):1172–1206.

Banfield, J. D. and Raftery, A. E. (1992). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16.

Benner, P., Mehrmann, V., and Sorensen, D. C. (2005). *Dimension Reduction of Large-scale Systems*, volume 45. Springer.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics*, 31(1):1–29.

Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds: II. *Annals of statistics*, 33(3):1225–1259.

Bhattacharya, R. N., Ellingson, L., Liu, X., Patrangenaru, V., and Crane, M. (2012). Extrinsic analysis on manifolds is computationally faster than intrinsic analysis with applications to quality control by machine vision. *Applied Stochastic Models in Business and Industry*, 28(3):222–235.

Biau, G. and Fischer, A. (2011). Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58(3):1924–1939.

Bishop, R. L. and Crittenden, R. J. (2011). *Geometry of Manifolds.* Academic press.

Boothby, W. M. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press.

Buss, S. R. and Fillmore, J. P. (2001). Spherical averages and applications to spherical splines and interpolation. *ACM Transactions on Graphics (TOG)*, 20(2):95–126.

Chang, K.-Y. and Ghosh, J. (2001). A unified model for probabilistic principal surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):22–41.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American statistical association*, 91(434):862–872.

Chowdhury, J. and Chaudhuri, P. (2019). Nonparametric depth and quantile regression for functional data. *Bernoulli*, 25(1):395–423.

Cippitelli, E., Gasparrini, S., Gambi, E., and Spinsante, S. (2016). A human activity recognition system using skeleton data from rgbd sensors. *Computational Intelligence and Neuroscience*, 2016.

Coope, I. D. (1993). Circle fitting by linear and nonlinear least squares. *Journal of Optimization Theory and Applications*, 76(2):381–388.

Dai, X., Lin, Z., and Müller, H.-G. (2021). Modeling sparse longitudinal data on Riemannian manifolds. *Biometrics*, 77(4):1328–1341.

Dai, X. and Müller, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Annals of Statistics*, 46(6B):3334–3361.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77(1):84–116.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22.

Duchamp, T. and Stuetzle, W. (1996). Extremal properties of principal curves in the plane. *Annals of Statistics*, 24(4):1511–1520.

Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853.

Einbeck, J., Evers, L., and Einbeck, M. J. (2015). ***LPCM**: Local Principal Curve Method.* R package version 0.46-7.

Einbeck, J., Tutz, G., and Evers, L. (2005). Local principal curves. *Statistics and Computing*, 15(4):301–313.

Eltzner, B., Huckemann, S., and Mardia, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Annals of Applied Statistics*, 12(2):1332–1359.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, 105(2):171–185.

Fletcher, P. T. and Joshi, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262.

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

Fletcher, P. T., Venkatasubramanian, S., and Joshi, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152.

Flury, B. and Tarpey, T. (1996). Self-consistency: A fundamental concept in statistics. *Statistical Science*, 11(3):229–243.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310.

Goh, A. and Vidal, R. (2008). Clustering and dimensionality reduction on Riemannian manifolds. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7.

Gray, N. H., Geiser, P. A., and Geiser, J. R. (1980). On the least-squares fit of small and great circles to spherically projected orientation data. *Journal of the International Association for Mathematical Geology*, 12(3):173–184.

Hastie, T. (1984). *Principal curves and surfaces.* PhD thesis, Stanford University.

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

Hastie, T. and Weingessel, A. (2015). ***princurve**: Fits a Principal Curve in Arbitrary Dimension.* R package version 2.16.

Hauberg, S. (2016). Principal curves on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1915–1921.

He, X. and Simpson, D. G. (1992). Robust direction estimation. *Annals of Statistics*, 20(1):351–369.

Hijmans, R. J., Williams, E., and Vennes, C. (2017). ***geosphere**: Spherical Trigonometry*. R package version 1.5-10.

Huber, P. J. (2004). *Robust Statistics*, volume 523. John Wiley & Sons.

Huckemann, S., Hotz, T., and Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–58.

Huckemann, S. and Ziezold, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319.

Ilea, I., Bombrun, L., Terebes, R., Borda, M., and Germain, C. (2016). An m-estimator for robust centroid estimation on the manifold of covariance matrices. *IEEE Signal Processing Letters*, 23(9):1255–1259.

Ionescu, C., Li, F., and Sminchisescu, C. (2011). Latent structured models for human pose estimation. In *2011 International Conference on Computer Vision*, pages 2220–2227.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339.

Jung, S., Dryden, I. L., and Marron, J. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

Jung, S., Foskey, M., and Marron, J. (2011). Principal arc analysis on direct product manifolds. *Annals of Applied Statistics*, 5(1):578–603.

Jung, S., Park, K., and Kim, B. (2021). Clustering on the torus by conformal prediction. *Annals of Applied Statistics*, 15(4):1583–1603.

Justusson, B. (1981). Median filtering: Statistical properties. In *Two-Dimensional Digital Signal Prcessing II*, pages 161–196. Springer.

Kåsa, I. (1976). A circle fitting procedure and its error analysis. *IEEE Transactions on Instrumentation and Measurement*, IM-25(1):8–14.

Kégl, B. (1999). *Principal curves: learning, design, and applications.* PhD thesis, Concordia University.

Kégl, B., Krzyzak, A., Linder, T., and Zeger, K. (2000). Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):281–297.

Kelly, J. L. (1991). *General Topology.* Springer.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121.

Kirov, S. and Slepčev, D. (2017). Multiple penalized principal curves: Analysis and computation. *Journal of Mathematical Imaging and Vision*, 59(2):234–256.

Kohonen, T. (1990). The self-organizing map. In *Proceedings of the IEEE*, volume 78, pages 1464–1480. IEEE.

Lee, J., Kim, J.-H., and Oh, H.-S. (2020). Spherical principal curves. *arXiv preprint arXiv:2003.02578.*

Lee, J., Kim, J.-H., and Oh, H.-S. (2021a). Spherical principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2165–2171.

Lee, J., Kim, J.-H., and Oh, H.-S. (2021b). Supplementary material for spherical principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Available online.

Lee, J., Kim, J.-H., and Oh, H.-S. (2022a). spherepc: An R package for dimension reduction on a sphere. *R Journal*, 14(1):167–181.

Lee, J., Kim, J.-H., and Oh, H.-S. (2022b). ***spherepc**: Spherical Principal Curves.* R package version 0.1.7.

Lee, J. M. (2006). *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media.

Lin, Z. and Yao, F. (2019). Intrinsic Riemannian functional data analysis. *Annals of Statistics*, 47(6):3533–3577.

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

Mallasto, A. and Feragen, A. (2018). Wrapped Gaussian process regression on Riemannian manifolds. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5580–5588. IEEE.

Mardia, K. V. (2014). *Statistics of Directional Data.* Academic press.

Mardia, K. V. and Gadsden, R. J. (1977). A small circle of best fit for spherical data and areas of vulcanism. *Journal of the Royal Statistical Society: Series C*, 26(3):238–245.

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(505):424–436.

Pennec, X., Fillard, P., and Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66.

Petrus, P. (1999). Robust Huber adaptive filter. *IEEE Transactions on Signal Processing*, 47(4):1129–1133.

Robotham, A. (2013). ***sphereplot**: Spherical Plotting.* R package version 1.5.

Sard, A. (1965). Hausdorff measure of critical images on Banach manifolds. *American Journal of Mathematics*, 87(1):158–174.

Scales, L. (1985). *Introduction to Nonlinear Optimization.* Macmillan International Higher Education.

Shin, H.-Y. and Oh, H.-S. (2022). Robust geodesic regression. *International Journal of Computer Vision*, 130(2):478–503.

Siddiqi, K. and Pizer, S. (2008). *Medial Representations: Mathematics, Algorithms and Applications*, volume 37. Springer Science & Business Media.

Stanford, D. C. and Raftery, A. E. (2000). Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):601–609.

Sturm, K.-T. (2003). Probability measures on metric spaces of nonpositive. *Contemporary Mathematics*, 338:357–390.

Telschow, F. J., Pierrynowski, M. R., and Huckemann, S. F. (2019). Confidence tubes for curves on SO(3) and identification of subject-specific gait change after kneeling. *arXiv preprint arXiv:1909.06583.*

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing*, 2(4):183–190.

Tukey, J. W. (1977). *Exploratory Data Analysis*, volume 2. Reading, MA.

Umbach, D. and Jones, K. N. (2003). A few methods for fitting circles to data. *IEEE Transactions on Instrumentation and Measurement*, 52(6):1881–1885.

Van de Geer, S. (2000). *Empirical Processes in M-estimation.* Cambridge university press.

Wang, H. and Lee, T. C. (2008). Extraction of curvilinear features from noisy point patterns using principal curves. *Pattern Recognition Letters*, 29(16):2078–2084.

Yang, L. (2010). Riemannian median and its estimation. *LMS Journal of Computation and Mathematics*, 13:461–479.

Yao, Z., Xia, Y., and Fan, Z. (2019). Fixed boundary flows. *arXiv preprint arXiv:1904.11332*.

Yao, Z. and Zhang, Z. (2020). Principal boundary on Riemannian manifolds. *Journal of the American Statistical Association*, 115(531):1435–1448.

Zhang, H., Pedrycz, W., Miao, D., and Zhong, C. (2013). A global structure-based algorithm for detecting the principal graph from complex data. *Pattern Recognition*, 46(6):1638–1647.

Zhang, M. and Fletcher, T. (2013). Probabilistic principal geodesic analysis. *Advances in neural information processing systems*, 26.

국문초록

본 학위 논문은 다양체 자료의 변동성을 더욱 효과적으로 찾아내기 위해, 다양체 자료의 비모수적 차원축소방법론을 제시하였다. 구체적으로, 주곡선(principal curves) 방법을 일반적인 다양체 공간으로 확장하는 것이 주요 연구 주제이다. 주곡선은 주성분분석(PCA)의 비선형적 확장 중 하나이며, 본 학위논문은 크게 네 가지의 주제로 이루어져 있다.

첫 번째로, Hastie (1984); Hastie and Stuetzle (1989)의 방법을 임의의 차원의 구면으로 표준적인 방식으로 확장한다. 이 연구 주제의 공헌은 다음과 같다. (a) D차원 구면 S^D에서 내재적, 외재적인 방식의 주곡선 방법을 각각 제안한다. (b) 본 방법의 이론적 성질(정상성)을 규명한다. (c) 지질학적 자료 및 인간 움직임 자료와 같은 실제 자료와 2차원, 4차원 구면 위의 모의실험 자료에 본 방법을 적용하여, 그 유용성을 보인다.

두 번째로, 첫 번째 주제의 후속 연구 중 하나로서, 두꺼운 꼬리 분포를 가지는 자료에 대하여 강건한 비모수적 차원축소 방법을 제안한다. 이를 위해, L_2 손실함수 대신에 L_1- 및 휴버(Huber) 손실함수를 활용한다. 이 연구 주제의 공헌은 다음과 같다. (a) 이상치에 덜 민감한 강건화주곡선(robust principal curves)을 구면에서 정의한다. 구체적으로, 자료의 기하적 중심점을 지나는 L_1- 및 휴버 손실함수에 대응되는 새로운 주곡선을 제안한다. (b) 이론적인 측면에서, 강건화주곡선의 정상성을 규명한다. (c) 강건화주곡선을 구현하기 위해 계산이 빠른 실용적인 알고리즘을 제안한다.

세 번째로, 기존의 차원축소방법 및 본 방법론을 제공하는 R 패키지를 구현하였으며 이를 다양한 예제 및 설명과 함께 소개한다. 본 방법론의 강점은 다양체 위에서의 복잡한 최적화 방정식을 풀지않고, 직관적인 방식으로 구현 가능하다는 점이다. R 패키지로 구현되어 제공된다는 점이 이를 방증하며, 본 학위 논문의 연구를 재현가능하게

만든다.

마지막으로, 보다 복잡한 기저(underline) 구조를 갖는 다양체 자료의 효과적인 추정을 위해 국소주측지선분석(local principal geodesics) 방법을 우선 제안한다. 이 방법을 실제 지질학 자료 및 다양한 모의실험 자료에 적용하여 그 활용성을 보였다. 다음으로, 추정치의 분산안정화 및 이론적 정당화를 위하여 Kégl (1999); Kégl et al. (2000) 방법을 일반적인 리만다양체로 확장한다. 더 나아가 방법론의 일치성 및 수렴속도와 같은 점근적 성질을 비롯하여 비점근적 성질인 집중부등식(concentration inequality)을 통계적학습이론을 이용하여 규명한다.

주요어 : 곡선적합, 다양체 자료, 주곡선, 재현가능성, 차원축소방법, 통계적학습이론

학 번 : 2018–25763

감사의 글

설렘과 두려움을 안고 시작한 박사학위 과정도 이제 논문심사를 마치고 학위 논문 인쇄만을 남겨 두었습니다. 2018년 3월부터 오늘까지의 시간은 저에게는 학문의 길뿐 아니라 인내, 끈기 그리고 도전에 대한 매력을 선물해준 소중한 시간이었습니다. 동시에 연구자로서의 자격과 가치를 증명해야 한다는 압박감도 있었습니다. 그러한 과정에서 감사했던 분들께 이 기회를 빌려 감사의 말을 드리고자 합니다.

부족한 저를 제자로 받아들여, 하나에서 열까지 가르침을 주시고, 미래와 희망을 선물하여 주신 오희석 교수님께 깊이 감사드립니다. 부족함을 하나하나 채워나가는 모습으로 큰 가르침에 답하겠습니다. 바쁘신 와중에도 귀한 시간을 내시어 제 학위 논문의 심사와 소중한 조언을 해주신 이재용, 임채영, 이우주, 임예지 교수님께도 진심으로 감사드립니다. 그동안 직·간접적으로 수많은 가르침을 주신 통계학과 교수님께 감사의 인사를 올립니다.

대학원 생활동안 누구보다 가까이서 동거동락한 연구실 선·후배님에게 감사의 인사를 전하고 싶습니다. 특히 많은 걸 가르쳐주시고 저를 인정해주신 선철 형, 대학교 동기이자 연구실 선배로서 항상 생활의 지혜를 일깨워주고 현명한 조언과 격려를 해준 준현 형, 크고 작은 도움을 준 준표에게 감사하다는 말을 전합니다. 또한 연구실 방장할 때, 신경써준 연구실 후배 규순·수빈에게 고마움을 전합니다. 함께 지낸 통계학과 2018학번 대학원 동기들에게도 감사의 인사를 전하고 싶습니다. 특별히, 좋을 때나 그렇지 않을 때나 함께하고 학문적·생활적으로 교류할 수 있었던 윤호, 균형 잡힌 생활이 뭔지 알게 해주고 아낌없는 격려를 해준 태현 형, 재민, 민준 형 그리고 봉수, 자주 함께 산책한 창원이, 영준, 사람 좋은 게 느껴지고 배려심있는 몽주, 인간미 있는 규민이, 착한 창겸이에게 고맙다는 말을 전하고 싶습니다. 아울러, 지혜와 마음을 나누어주었던 한지혜

누나에게도 감사의 인사를 전합니다.

언제나 변함없이 저의 결정을 옆에서 믿어주시고 뒷받침 해주시는 아버지, 어머니께 감사의 인사를 드립니다. 어떤 자리에 있더라도 열심히 살아가며 부모님의 사랑에 보답하겠습니다. 스스로를 돌보지 못할 때에도, 저를 챙겨준 오랜 친구인 유지원에게도 진심으로 감사의 말을 전합니다. 아울러, 함께 대학에 입학하여 서로 다른 길을 가고 있는 대학교 동기들에게도 고마움의 메세지를 전하고 싶습니다. 특히 오랜 기간 친하게 지내고 긍정적인 기운을 주었던 황기훈, 대학원 생활에 직·간접적으로 조언을 주고 마주칠 때마다 웃어주는 종진 형, 인간적인 면에서 귀감이 된 중경, 박사 후 진로에 대해 유용한 정보를 주고 도와준 성수에게 감사하다는 말을 전합니다. 언제나 맛있는 음식으로 환영해주신 매형과 누나, 사랑스러운 조카 이솔·이찬에게 고마움을 전합니다. 마지막으로, 항상 곁에서 힘이 되어주는 어머니에게 큰 사랑과 감사를 전합니다. 어머니의 이해와 헌신 그리고 한없는 배려가 있었기에 지금의 결과를 만들 수 있었습니다.

2022년 8월 이종민