Probabilistic Models on Fibre Bundles

by

Shan Shan

Department of Mathematics
Duke University

Date: ____________________________
Approved:

____
Ingrid Daubechies, Supervisor

____
Sayan Mukherjee, Chair

____
Doug Boyer

____
Colleen Robles

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Mathematics
in the Graduate School of
Duke University

2019

ABSTRACT

Probabilistic Models on Fibre Bundles

by

Shan Shan

Department of Mathematics
Duke University

Date: ____________________________
Approved:

____
Ingrid Daubechies, Supervisor

____
Sayan Mukherjee, Chair

____
Doug Boyer

____
Colleen Robles

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Mathematics
in the Graduate School of
Duke University

2019

Abstract

In this thesis, we propose probabilistic models on fibre bundles for learning the generative process of data. The main tool we use is the diffusion kernel and we use it in two ways. First, we build from the diffusion kernel on a fibre bundle a *projected kernel* that generates robust representations of the data, and we test that it outperforms regular diffusion maps under noise. Second, this diffusion kernel gives rise to a natural covariance function when defining Gaussian processes (GP) on the fibre bundle. To demonstrate the uses of GP on a fibre bundle, we apply it to simulated data on a Möbius strip for the problem of prediction and regression. Parameter tuning can also be guided by a novel semi-group test arising from the geometric properties of diffusion kernel. For an example of real-world application, we use probabilistic models on fibre bundles to study evolutionary process on anatomical surfaces. In a separate chapter, we propose a robust algorithm (ariaDNE) for computing curvature on each individual surface. The proposed machinery, relating diffusion processes to probabilistic models on fibre bundles, provides a unified framework for ideas from a variety of different topics such as geometric operators, dimension reduction, regression and Bayesian statistics.

Acknowledgements

I am greatly indebted to three very important people. It was their vision that made this thesis be.

To work with someone like Ingrid Daubechies is a bliss. The pages here are peppered with her creativity, endeavor for simplicity, and good humor. She has brought an unparalleled clarity to this work, especially on the diffusion process side. I am also grateful for having the opportunities, as her student, to closely observe how she works as a mathematician and the way she lives her life as an artist. I am constantly in awe with the divine love she has for everyone around her and the efforts she makes for building a more supportive mathematics community. When I see her, I see who I want to be.

I thank Sayan Mukherjee for suggesting the probabilistic theory on fibre bundles in the first place and for the many enlightening conversations afterwards. It was Sayan who helped me refine the Gaussian process model each time and guided me to see my work in the broader context of geometry and statistical learning.

Doug Boyer nourished my interests in teeth, anatomy and biology in general. Without his encouragement, I would not have continued on the project of ancestral surface reconstruction, which later led to the ariaDNE paper and ultimately the fibre bundle approach for evolutionary studies.

Thanks to Colleen Robles, who serves on my committee and has been extremely supportive around the time of my defense.

I am happy and grateful to be part of the tooth and claw group and the Rhodes Information Initiative at Duke. I thank everyone for their good vibes and inspiring discussions. Thanks to Shahar Kovalsky and Julie Winchester for the collaboration on ariaDNE! Thank you Ethan Fulwood for gathering all the lemur teeth data.

Thank you Deborshee Sen for pointing out that marginal sampler! Thank you also to Tingran Gao, Rob Ravier, Henry Kirveslahti, Barak Sober, Nadav Dym, Chen-yun Lin, Wei Zhu, Panchali Nag. Particular thanks to Kathy Peterson who makes iiD the best place to work!

I would also like to express my gratitude to my mentors and role models over the years – Robert Calderbank, Lillian Pierce, Peter Diao and Tom Beale – for many words of wisdom.

Thanks also to Christine Markwalter, Vaishakhi Mayya and Caitlin Leverson for your friendship.

Lastly, I want to thank my parents Runbo Shan and Xuezhi Zhou, especially my mother who cooked every meal for me during the past six weeks to help me concentrate on my writing. I always feel blessed that I was never in doubt of their love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Data, such as genetic sequences, climate patterns, stock market prices, audio signals, images, shapes, video clips, etc., are often represented as points in a high dimensional Euclidean space. Nevertheless, the data may be concentrated near a subspace or subset of a much smaller dimension. For example, suppose we are given N images, where each image is represented by the gray-scale values at each pixel. If there were 256×256 pixels in each image, then each image yields an element in $\mathbb{R}^{256\times 256}$. Despite being high-dimensional, a collection of images may have an intrinsic structure that is of lower dimension. For instance, the collection could consist of images representing photographs of the same object taken under different angles with a moving camera. The intrinsic dimensionality of this collection should then be the degree of freedoms in rotating the camera; the images in this collection would then be labeled by these (relatively few) degrees of freedom. The data can thus be viewed in different ways: an image can be viewed as an element in $\mathbb{R}^{256\times 256}$, or as an array of 256×256 real values, or as one element (with its own internal structure, representing relations between the pixels) in a lower dimensional structure within $\mathbb{R}^{256\times 256}$. We shall use the name *data object* to distinguish this richer structure from, e.g., a single pixel value, which we could also call a *data point*.

The problem of finding the intrinsic low-dimensional structure is called dimension reduction. Dimension reduction in the Euclidean setting dates back to 1900s. The objective was to find a low-dimensional representation for data objects that improves storage, reduces noise and avoids over-fitting. Principal component analysis and Multidimensional scaling are two classical methods in this category.

The next stage of development in dimension reduction saw data as lying on a geometric object, e.g., manifold or fibre bundle, and the connection to geometry has been fertile for methods that capture the non-linear structure in the data. While these methods still compute a low-dimensional representation for data, the main concern was to define new distances that reflect the clustering or community structure within the data objects.

In this thesis, rather than try to find the explicit data representation or pairwise distances between data objects, we shall extend the phrase "dimension reduction" towards denoting aiming to understand the underlying "data generation process", that is to find a model which is most consistent with the given data. The reduction comes from summarizing the generative model with a small number of parameters. Our approach focuses on the diffusion kernel, a mathematical entity used in *diffusion maps* [CL06] to generate efficient representations of complex data structures.

There are two benefits brought by viewing the diffusion kernel as the primary object of dimension reduction.

Firstly, the diffusion kernel on a manifold is directly related to the Laplace-Beltrami operator, and therefore bears very useful geometric properties. The associated integral operators also form a semi-group. Later we shall see that this semi-group property can help us tune parameters in diffusion methods, which is often a difficult task in practice!

Secondly, the diffusion kernel gives rise to a natural choice for the covariance matrix or function in defining probabilistic models on the dataset.

1.2 Main results

We begin with assuming data lie on a fibre bundle E which consists of a base manifold M, a fibre manifold F and a projection map π. Intuitively, we can think of a fibre bundle as gluing a fibre to each point on the base manifold, in such a way that the fibre bundle looks like a product space locally. Formally, a fibre bundle E is a quadruple (E, M, F, π) that satisfies the following properties:

1. E, M, F are differential manifolds.
2. $\pi : E \to M$ is a projection with local trivialization, i.e., M admits an open cover $\mathcal{U}$ such that, for any $U \in \mathcal{U}$, $\pi^{-1}(U)$ is diffeomorphic to $U \times F$.

We think of our data set as consisting of data objects corresponding to elements of on the base manifold M, each labeled by its corresponding element of M; each data object x contains data points that belong to F_x, the fibre attached to $x \in M$. Between two data objects, there is also a correspondence map which can be interpreted as an approximation to the parallel-transport along the geodesic connecting the two elements on the base manifold. The task we set ourselves is to extract, as much and/or accurate as possible, information on the base manifold from data points lying on the fibre and the maps between them.

Our main investigation tool consists of diffusion maps on fibre bundles [Gao16]. We further develop the *projected kernel method* which computes a denoised version of the diffusion kernel on the base manifold. We shall see that the performance of the projected method is better than regular diffusion maps under noise.

The main theme of this thesis is to develop probabilistic theory on fibre bundles so that standard data learning machinery can be carried out.

The probabilistic models we develop are:

1. Gaussian processes on a fibre bundle, to wit diffusion kernels of fibre bundles yielding geometrically motivated covariance functions.
2. Regression on fibre bundle, where we deal with the inference problem and parameter tuning.

We emphasize a distinct interest in computing parameters on the base manifold. Oftentimes, we hope to understand how the data object, as a whole, is created or how several are related to each other, more than what's happening to each individual data point as a separate entity. To obtain the base manifold parameters, we marginalize out the parameters on the fibre. This can be viewed as an analogue to the projected kernel method in the non-probabilistic setting.

1.3 Applications to biology

Besides results in the general learning theory, we also made advancement in applications to morphological data, especially molar surfaces of primates.

Understanding the evolutionary process of the primate group is one of the most fundamental questions in biology. Surfaces of teeth are important study objects in relating extinct and extant species. We consider the shape space of teeth as a fibre bundle. A shape S is an element s in the base manifold; the coordinates of points on S are elements of the fibre attached to s. We describe the shape evolution as a Gaussian process on the fibre bundle of teeth. Determining for which evolution model we have more evidence is then equivalent to a parameter learning or model selection problem on the base manifold in the fibre bundle framework.

1.4 Organization

The plan of this thesis is as follows: Chapter 2 is an introduction to the dimension reduction problem in the Euclidean setting. Chapter 3 is a review of diffusion maps. It also includes a discussion on the semi-group test to guide parameter tuning.

Chapters 4 through 8 develop one main theme. Chapter 4 introduces the fibre bundle assumption in data analysis, reviews diffusion maps on fibre bundles and proposes the projected kernel method. The problem of regression is studied in Chapter 5. We use it to motivate our viewing the kernel as a primary object in dimension reduction. Chapter 6 deals with Gaussian process on fibre bundle and the problem of prediction. Chapter 7 studies regression on a fibre bundle. The link between the diffusion kernel and dimension reduction is capped in this chapter. In Chapter 8, we apply regression on a fibre bundle to evolutionary studies on lemur teeth.

Chapter 9 looks at morphology at a different level. In this chapter, we zoom in onto each individual fibre and look at features on the tooth surface. We propose a robustly implemented algorithm for computing the Dirichlet energy of the normal maps (ariaDNE). This work was done in collaboration with Shahar Z. Kovalsky, Julie M. Winchester, Doug M. Boyer and Ingrid Daubechies. It was previously published in *Methods in Ecology and Evolution*.

Finally, Chapter 10 sketches extensions and applications of probabilistic models on a fibre bundle.

Chapter 2

Dimension reduction on Euclidean space

This chapter introduces the dimension reduction problem in the Euclidean setting. The main goal is a review of classic dimension reduction techniques, which serves as motivation for what follows in the manifold and fibre bundle setting. Section 2.1 illustrates the benefits of dimensionality reduction with Principal Component Analysis (PCA) [Jol11] and Multidimensional Scaling (MDS) [Tor52]. We shall also discuss in Section 2.2 the limitations of these methods that lead us to viewing data as lying on a more interesting non-linear geometric structure.

2.1 The problem of dimension reduction

Formally, we are given a dataset consisting of N data objects $\{\mathbf{y}_i \mid 1 \leq i \leq N\}$, where each data object $\mathbf{y}_i \in \mathbb{R}^d$ has d data points. The task we set ourselves is to represent $\mathbf{y}_i$ in a reduced Euclidean subspace as $\mathbf{x}_i \in \mathbb{R}^k$, where $k \ll d$. We term the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^k$ a k-dimensional *embedding*, where ϕ is defined as $\phi(\mathbf{y}_i) = \mathbf{x}_i$.

2.1.1 Principal component analysis

Principal Component Analysis (PCA) is one of the most widely used dimension reduction methods. The low-dimensional embedding via PCA is easy to compute. Suppose we are given $\{\mathbf{y}_i \in \mathbb{R}^d \mid 1 \leq i \leq N\}$ as before. We first normalize the data to have zero mean and unit variance by

1. replacing each $\mathbf{y}_i$ with $\mathbf{y}_i - \bar{\mathbf{y}}$, where $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i$.

2. replacing each y_{ij} with y_{ij}/σ_j, where $\sigma_j^2 = \sum_{i=1}^{N} y_{ij}^2$.

We calculate the $d \times d$ sample covariance matrix,

$$S = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^{\intercal}.$$

The eigenvectors $\{\mathbf{w}_i \in \mathbb{R}^d \mid 1 \leq i \leq k\}$ corresponding to the k largest eigenvalues of S are found, i.e., for $i = 1, \ldots, k$, $S\mathbf{w}_i = \lambda_i \mathbf{w}_i$, $||\mathbf{w}_i|| = 1$; $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$, and $||S\,|_{\{\mathbf{w}_1,\ldots,\mathbf{w}_k\}^{\perp}}\,|| \leq \lambda_k$. We then define the low-dimensional representation of $\mathbf{y}_n$ by

$$\mathbf{x}_i = (\mathbf{w}_1^{\intercal}\mathbf{y}_i, \mathbf{w}_2^{\intercal}\mathbf{y}_i, \ldots, \mathbf{w}_k^{\intercal}\mathbf{y}_i) \in \mathbb{R}^k.$$

For simplicity, we write

$$\mathbf{x}_i = \phi(\mathbf{y}_i) = \mathbf{W}^{\intercal}\mathbf{y}_i, \tag{2.1}$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d \times k}$. We remark here that PCA is a *linear* dimension reduction method. Indeed, the embedding ϕ is a projection onto the eigenvectors.

Besides the obvious data compression and dimension reduction, PCA has many applications. First, PCA assists data visualization. The data $\mathbf{y}_n \in \mathbb{R}^d$ is reduced to $\mathbf{x}_n \in \mathbb{R}^k$, where $k \ll d$. If $k = 2$ or 3, we can plot the data and use this to gain insights about the structure, e.g., similarity and clustering, of the data. Second, PCA improves over-fitting. Before running a supervised-learning algorithm, reducing the data dimensionality can reduce the complexity of the hypothesis class (e.g., the number of parameters in the parametrized function space) and thus help reduce over-fitting. Third, Hotelling [Hot33] showed that the projection as in (2.1) maximizes the variance of the data in the projected space. Hence, PCA extracts the systematic and interesting variations among data, and at the same time gets rid of the noise from measurement, scanning devices, and so on. A classic example is the *eigenface* method [TP91]. Another application of this idea is to estimate curvature on a discrete surface. Details can be found in Chapter 9, where we discuss its application and use in biological studies [SKW+19].

2.1.2 Multidimensional scaling

In many situations, the sample covariance matrix $\mathbf{S}$ can not be obtained. One such example is when not all $\mathbf{y}_n$'s are of the same dimensionality d. In this case, we will take advantage of the proximities (e.g., pairwise distance between data), if available. In this section, we review the classical Multidimensional Scaling (MDS) method, which reduces the data dimensionality from the proximity inputs.

Suppose we are given data $\{\mathbf{y}_i \in \mathbb{R}^{d_i} \mid 1 \leq i \leq N\}$[1]. Note that d_i may differ for different data objects $\mathbf{y}_i$. Suppose we are also given their proximity $D = [D_{ij}] \in \mathbb{R}^{N \times N}$. The proximity D can be computed by taking the pairwise distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ if $d_i = d_j$, or if $\mathbf{y}_i$ and $\mathbf{y}_j$ are viewed as approximations to $\tilde{\mathbf{y}}_i$, $\tilde{\mathbf{y}}_j \in \mathbb{R}^d$ respectively, between $\tilde{\mathbf{y}}_i$ and $\tilde{\mathbf{y}}_j$; D can also be given a priori. MDS finds low-dimensional representation $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^k$ respectively such that the Euclidean distance between $\mathbf{x}_i, \mathbf{x}_j$ is close to D_{ij}, i.e.,

$$||\mathbf{x}_i, \mathbf{x}_j|| := \left(\sum_{p=1}^{k} (x_{ip} - x_{jp})^2 \right)^{1/2} \approx D_{ij}.$$

---

[1] The dataset $\{\mathbf{y}_i \in \mathbb{R}^{d_i} \mid 1 \leq i \leq N\}$ can also be unknown. MDS only requires the proximity data.

To compute MDS, we first compute the matrix $A = [A_{ij}]$, with entries

$$A_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2),$$

where $d_{i\cdot}^2 = \frac{1}{N}\sum_{p=1}^{N} d_{ip}^2$, $d_{\cdot j}^2 = \frac{1}{N}\sum_{p=1}^{N} d_{pj}^2$, and $d_{\cdot\cdot}^2 = \frac{1}{N}\sum_{p=1}^{N}\sum_{q=1}^{N} d_{pq}^2$. We then compute the eigenvectors of A,

$$A\mathbf{w}_i = \lambda_i \mathbf{w}_i$$

and obtain the k-largest eigenvalues and their corresponding normalized eigenvectors, $\{\mathbf{w}_i \in \mathbb{R}^N \mid 1 \leq i \leq k\}$. Finally, we define the low-dimensional representation $\mathbf{x}_i$ as

$$\mathbf{x}_i = (\lambda_i w_{1i}, \lambda_i w_{2i}, \ldots, \lambda_i w_{ki}) \in \mathbb{R}^k.$$

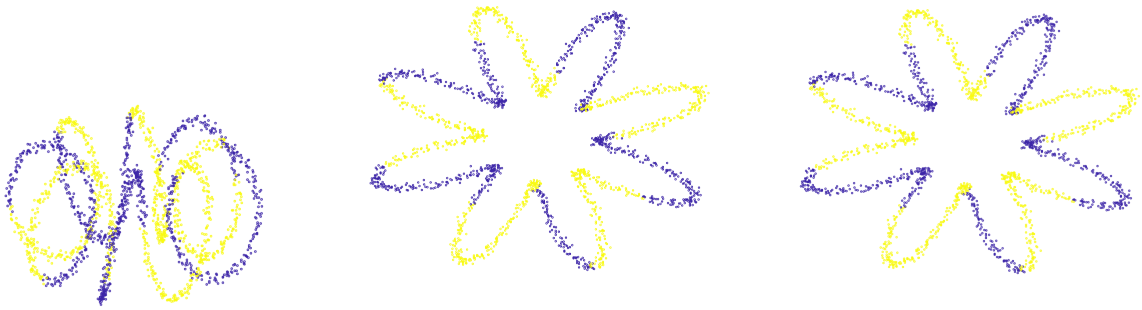
To see that MDS is also a linear dimension reduction method, we write $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{R}^{k \times N}$. Then

$$\mathbf{X} = \mathbf{\Lambda}\mathbf{W}, \tag{2.2}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries $\Lambda_{ii} = \lambda_i$ and $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k)$.

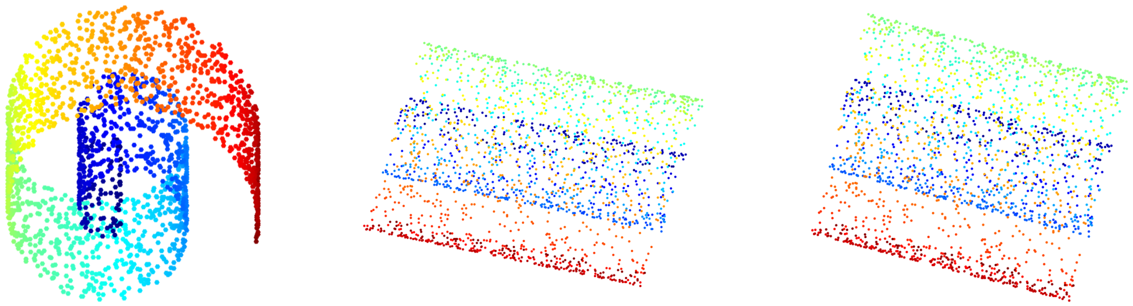
2.2 Limitation of PCA and MDS

Due to their linearity, neither PCA nor MDS can adequately handle complex nonlinear data. As an illustration, consider a set of points $\{\mathbf{y}_i\}$ in $\mathbb{R}^3$ sampled from a helix. We computed and plotted the $\mathbb{R}^2$ embedding obtained using PCA and MDS. The results are shown in Figure 2.1. The arc-length parametrization should be a perfect circle. However, neither PCA nor MDS found this intrinsic representation. Consider another data set, again in $\mathbb{R}^3$, sampled from a Swiss roll. Figure 2.2 shows



(a) Original data (b) PCA embedding (c) MDS embedding

Figure 2.1: PCA and MDS embedding of points lying on a helix.



(a) Original data (b) PCA embedding (c) MDS embedding

Figure 2.2: PCA and MDS embedding of points lying on a Swiss roll.

that both PCA and MDS flattened the roll by simply squashing it from its side wall instead of "unrolling" it from one of its ends. A similar flattening effect by PCA was also noticed in [PM02] for the analysis of an ecological dataset. Strongly non-linear data often leads to horseshoe (or arch)-shaped configurations in the first two principal components. To address this flattening issue on complex data in real world applications, we resort to non-linear dimensionality reduction methods, as discussed in the next chapter.

Chapter 3

Diffusion on manifold

Perhaps the most heard sentence in machine learning these days is that "data lies on (or near) a low-dimensional manifold". This entails that the given data set $\{y_i \in M \mid 1 \leq i \leq N\}$ are sampled from a Riemannian manifold M whose intrinsic dimension is low.

This chapter discusses why making the data space into a Riemannian manifold is useful. We illustrate with the method of *diffusion maps*, which is a general framework that encapsulates a variety of kernel eigenmaps methods, including Laplacian eigenmaps [BN08], local linear embedding [RS00], Hessian eigenmaps [DG03] and local tangent space alignment [ZZ04], etc.

We begin in Section 3.1 with a review on the Laplace-Beltrami operator, whose eigenfunctions give rise to an optimal embedding. We further discuss its connection to the diffusion operator via the heat equation. Section 3.2 describes the construction of the family of diffusion operators as in [CL06], the resulting diffusion maps and its organization power according to the intrinsic structure of the underlying manifold. Finally in Section 3.3 we describe the diffusion semi-groups, which gives rise to (1) a computationally efficient way of constructing diffusion maps and (2) a useful criterion for parameter tuning in practical applications.

3.1 Laplace-Beltrami operator

3.1.1 Definition and optimal embedding

We begin by defining the Laplace-Beltrami operator applied to a scalar function $f : M \to \mathbb{R}$ on a Riemannian manifold M; for simplicity we shall assume M to be compact, without boundary.

Definition (Laplace-Beltrami) Let M be a Riemannian manifold with a metric g. The Laplace-Beltrami operator on M is defined by

$$\Delta_M f(x) = -\mathrm{Trace} \nabla^M \nabla f(x)$$

where ∇^M denotes the canonical *Levi-Civita connection* on M.

Note in the case where M is a compact manifold with a smooth boundary, one has to consider appropriate boundary conditions in order to define ∇_M as a self-adjoint operator.

The spectrum of Δ_M on a compact manifold M is known to be discrete. Let the eigenvalues be $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$ and let f_i be the eigenfunction corresponding to eigenvalue λ_i. The eigenfunctions of the Laplace-Beltrami operator give rise to an embedding operation with certain optimality properties.

The crucial observation, due to [BN08], is that $||\nabla f||$ provides us with an estimate of how far apart f maps nearby points. Let $y_i, y_j \in M$; then

$$|f(y_i) - f(y_j)| \leq \mathrm{dist}_M(y_i, y_j)||\nabla f|| + o(\mathrm{dist}_M(y_i, y_j)). \tag{3.1}$$

Points that are close together on the manifold should be mapped close together in $\mathbb{R}$. A map that best preserves locality across the manifold, as measured by $||\nabla f||$ is given by

$$\underset{||f||_{L^2(M)}}{\arg\min} \int_M ||\nabla f(y)|| \tag{3.2}$$

It turns out that minimizing the objective function of (3.2) is equivalent to finding eigenfunctions of the Laplace Beltrami Δ_M. It follows that f_1 is the optimal embedding map to the real line[1]. The optimal embedding in $\mathbb{R}^k$ is then defined by

$$\mathbf{g} := (f_1(y), \ldots, f_k(y))$$

3.1.2 Diffusion operator

The Laplace-Beltrami operator Δ gives rise to the diffusion operator $e^{-t\Delta}$, in a sense that Δ is the infinitesimal generator of $e^{-t\Delta}$, i.e.,

$$\lim_{t \to 0} \frac{I - e^{-t\Delta}}{t} f = -\Delta f,$$

whenever f belongs to a suitable dense subset of $C(M)$. The diffusion operators $\{e^{-t\Delta}\}_{t>0}$ share the same eigenfunctions as Δ_M, and the eigenvalues of $e^{-t\Delta}$ are bounded: $1 = e^{-\lambda_0 t} \geq e^{-\lambda_1 t} \geq e^{-\lambda_2 t} \geq \cdots > 0$.

The Laplace-Beltrami operator Δ and the diffusion operator $e^{-t\Delta}$ are also related through the heat equation on the manifold.

Definition (Heat Equation). Let $f : M \to \mathbb{R}$ be the initial temperature distribution on a manifold M embedded in $\mathbb{R}^d$. The *heat equation* is the partial differential

---

[1] It is easily seen that f_0 is a constant function that maps the entire manifold to a single point.

equation

$$\frac{\partial u}{\partial t} + \Delta_M u = 0$$
$$u(x,0) = f(x)$$

The solution of the heat equation is given by the diffusion operator $e^{-t\Delta}$,

$$\begin{aligned} u(x,t) &= e^{-t\Delta} f(x) && (3.3) \\ &= \int_M h_t(x,y) f(y) d_{vol_M} y && (3.4) \end{aligned}$$

where h_t is the heat kernel. When the density of data is uniform and x, y are close and t is small, h_t can be approximated by the Gaussian

$$h_t(x,y) = (4\pi t)^{-\frac{d}{2}} e^{-\frac{||x-y||^2}{4t}}, \tag{3.5}$$

where d is the dimension of M.

On discrete data samples of N data objects, the diffusion operator $e^{-t\Delta}$ is approximated by a left multiplication with $D^{-1}W$ where W is an N by N matrix defined by

$$W_{ij} = \begin{cases} e^{-\frac{||y_i-y_j||^2}{4t}} & \text{if } ||y_i - y_j|| < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

and $D = [D_{ii}]$ is the diagonal matrix, whose entry is given by $D_{ii} = \sum_j W_{ij}$. We then compute eigenvalues and eigenvectors for

$$D^{-1}W\mathbf{f} = \lambda\mathbf{f}.$$

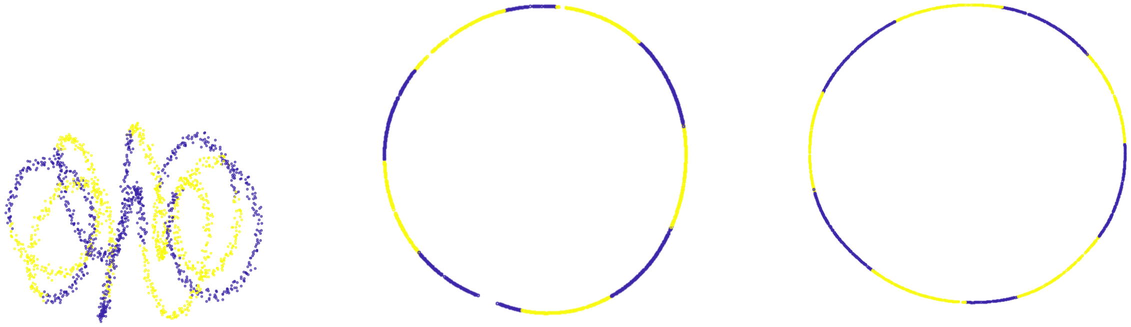

(a) Original data (b) PCA embedding (c) MDS embedding

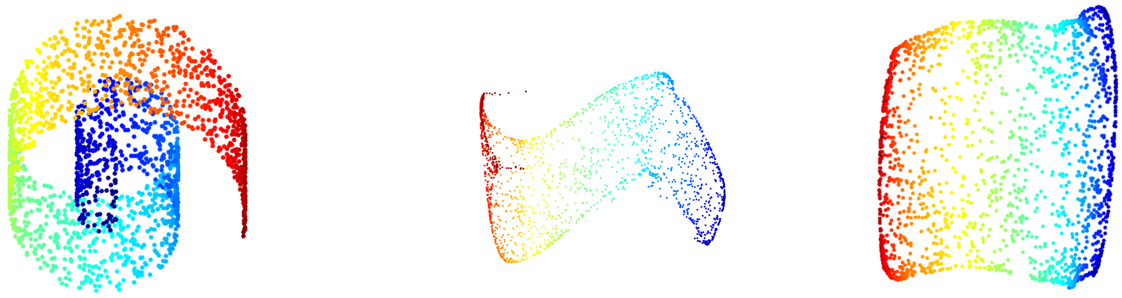
Figure 3.1: Laplacian eigenmaps and diffusion maps embedding of points lying on a helix.

The k-dimensional embedding is then defined to be

$$\mathbf{g}_i = (\mathbf{f}_1^{[i]}, \ldots, \mathbf{f}_k^{[i]}),$$

where $\mathbf{f}_1^{[i]}$ is the i-th entry in the N-dimensional eigenvector.

We further remark that if we use the Laplace-Beltrami operator Δ rather than $e^{-t\Delta}$ and approximate it with the normalized graph Laplacian $D^{-1}L$, where $L = D - W$, we recover the Laplacian eigenmaps. Figure 3.1 shows the embedding of points residing on a helix by Laplacian eigenmaps. Compared to PCA and MDS, Laplacian eigenmaps recovers the intrinsic data structure. However, when the density of data is not uniform, as shown in the Swiss roll example (Figure 3.2), the embedding produced by Laplacian eigenmaps is not optimal. In particular, due to sampling density, there seems to be leaks in the embedded image. To capture the geometry of a given manifold, regardless of the density, we discuss in the next section a different normalization that asymptotically recovers the eigenfunctions of the Laplace-Beltrami operator on the manifold.



(a) Original data (b) Laplacian (c) Diffusion maps

Figure 3.2: Laplacian eigenmaps and diffusion maps embedding of points lying on a Swiss roll.

3.2 Diffusion maps

3.2.1 Family of diffusion operators

We begin with the re-normalization of the heat kernel[2] h_t defined as in (3.5) for the case of non-uniform data.

Assume that the dataset Y is the entire manifold M, where M is a manifold embedded in $\mathbb{R}^d$. Let $q(x)$ denote the density of points on M. The re-normalization as in [CL06] consists of two steps:

1. Fix α. We form the new kernel

   $$h_{\alpha,t}(x,y) = \frac{h_t(x,y)}{p_t^\alpha(x)p_t^\alpha(y)},$$

   where

   $$p_t(x) = \int_M h_t(x,y)q(y)\,dy.$$

---

[2] In fact, diffusion maps assume an isotropic kernel which is invariant under rotations, $r(||x-y||^2/t)$, of which the heat kernel is an example.

2. Apply the weighted graph Laplacian normalization to this kernel, and we get

$$k_{\alpha,t}(x,y) = \frac{h_{\alpha,t}(x,y)}{d_{\alpha,t}(x)}, \tag{3.7}$$

   where

$$d_{\alpha,t}(x) = \int_M h_{\alpha,t}(x,y) q(y)\, dy.$$

We can define a family of operators $\{A_\alpha^t\}_{t>0}$ from the kernel (3.7),

$$A_\alpha^t f(x) = \int_M k_{\alpha,t}(x,y) f(y) q(y)\, dy. \tag{3.8}$$

Under mild additional assumptions on k, A_α^t has discrete eigenvalues $1 = \mu_0^t > \mu_1^t \geq \mu_2^t \geq \cdots \geq 0$, with eigenfunctions f_i^t corresponding to eigenvalue μ_i^t. We shall see later that, when t is small, A_α^t has the same eigenfunctions f_i for all t, and their corresponding eigenvalues μ_i^t are indeed powers of the eigenvalues μ_i of $A_{t=1}^\alpha$.

The spectral properties of the operators $\{A_\alpha^t\}$ relate to the geometry of the data again through the Laplace-Beltrami operator. Coifman and Lafon [CL06] showed that the infinitesimal generator defined by

$$L_{\alpha,t} = \frac{I - A_\alpha^t}{t}$$

can be used to approximate a specific symmetric Schrodinger operator. In particular, when $\alpha = 1$, $L_{\alpha,t}$ approximates the Laplace-Beltrami operator. Equivalently, the operator $e^{-t\Delta}$ defined via the heat kernel (3.4) can be approximated on $L^2(M)$ by

$A^t_{\alpha=1}$

$$\lim_{t\to 0} A^t_{\alpha=1} = e^{-t\Delta}. \tag{3.9}$$

Therefore, the global characterization of the geometry of the data set can be obtained from propagating and accumulating the local geometric information encoded in $\{A^t_\alpha\}$ in the same way the local transition of a diffusion/heat flow can be knitted together.

3.2.2 Diffusion distances and diffusion maps

The diffusion distance is defined via spectral properties of the diffusion operator $\{A^t_\alpha\}$. Fix t, α and let f^t_l, $l \leq 1$ be eigenfunctions of A^t_α and μ^t_i the corresponding eigenvalues, where $1 = \mu^t_0 > \mu^t_1 \geq \mu^t_2 \geq \cdots \geq 0$. The diffusion distance D^t_α is given by

$$D^t_\alpha(x,y) := \left(\sum_{l\geq 1} \left(\lambda^t_l\right)^2 \left(f^t_l(x) - f^t_l(y)\right)^2 \right)^{1/2}. \tag{3.10}$$

Since the eigenvalues monotonically decrease to zero, we can also truncate the sum up to k and get,

$$D^t_\alpha(x,y) \approx \left(\sum_{l=1}^{k} \left(\lambda^t_l\right)^2 \left(f^t_l(x) - f^t_l(y)\right)^2 \right)^{1/2}.$$

We further note that if we define $\phi^t_\alpha : M \to \mathbb{R}^k$ by

$$\phi^t_\alpha(y) := \left(\lambda^t_1 f^t_1(y), \ldots, \lambda^t_k f^t_k(y)\right),$$

the diffusion distance is equivalent to the Euclidean distance between data in the

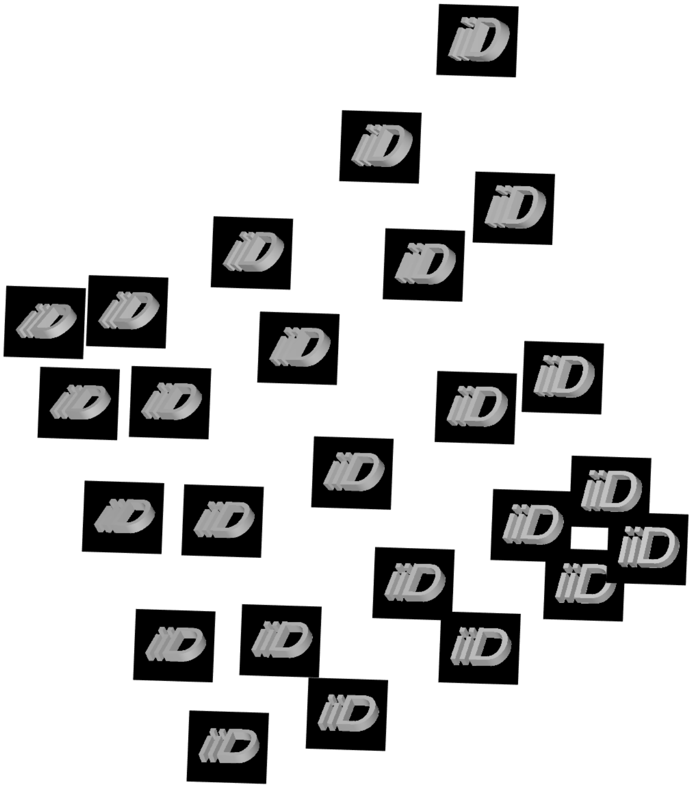


Figure 3.3: Diffusion maps organizes data by its intrinsic structure. The images of "iiD" letters are embedded in $\mathbb{R}^2$ with the first two non-trivial eigenvectors of the diffusion operator.

embedded space by ϕ_α^t.

$$D_\alpha^t(x, y) = ||\phi_\alpha^t(x) - \phi_\alpha^t(y)||.$$

The family of embedding functions $\{\phi_\alpha^t\}$ are called the *diffusion maps*.

The diffusion maps reorganize data according to their mutual diffusion distances. We illustrate the organizational power of diffusion maps in Figure 3.3. A collection of 25 images of the word "iiD" is generated by rotation with different angles. Each image consists of 25 by 25 gray-scale pixel values. The figure plots the diffusion embedding with the first two eigenvectors. Diffusion maps recover the natural parametrization dictated by the two angles of variation.

We must also not overlook the conservation property of the kernel $k_{\alpha,t}$,

$$\int_M k_{\alpha,t}(x,y)dy = 1. \tag{3.11}$$

This suggests that $k_{\alpha,t}$ can be viewed as the transition probability of a Markov process on M.

Indeed, the family of *diffusion distance* $\{D_\alpha^t\}$ between $x, y \in M$ can also be defined by summing the transition probability from x to y over all possible short paths connecting them.

$$\begin{aligned} D_\alpha^t(x,y) :&= ||k_{\alpha,t}(x,\cdot) - k_{\alpha,t}(y,\cdot)||^2_{L^2(M)} \\ &= \int_M |k_{\alpha,t}(x,u) - k_{\alpha,t}(y,u)|^2 \, du. \end{aligned}$$

The interpretation by transition probability suggests that the diffusion distance emphasizes the notion of a cluster at a time scale t. By definition, $D_\alpha^t(x,y)$ is small, if there is a large number of short paths between x and y, that is, there is a large probability of transition from x to y. Hence, D_α^t reflects the connectivity and community structure of data. Furthermore, since diffusion distance considers all possible paths in the graph, it will also be robust to noise. If the given pairwise distances between data objects are noisy/inaccurate, the diffusion distance can be considered as denoising/improving distances between data objects, and thus give a better view of how data are related on the object level.

With diffusion maps, we obtain, in addition to a low-dimensional data representation/parametrization, an improvement to the pairwise distances that reflect the intrinsic structure of the underlying space.

3.3 Diffusion semi-group

A family of operators $\{T^t\}$ where $0 \leq t < \infty$ is said to be a *semi-group*, if

$$T^{t_1+t_2} = T^{t_1}T^{t_2}.$$

Furthermore, T^0 is the *identity operator*.

We recognize that the family of operators $\{e^{-t\Delta}\}_{t>0}$ form a semi-group. Since the operator A_α^t approximates $e^{-t\Delta}$ as $t \to 0$, we know $\{A_\alpha^t\}$ must have the semi-group property when t is small.

The advantages of the semi-group property can be most seen when we apply diffusion methods to discrete data samples. Suppose our data set Y is given as a collection of data objects $\{y_i \in M \mid 1 \leq i \leq N\}$. If we think of the data objects as nodes of a symmetric graph whose weight function is specified by $k_{\alpha,t}$, then $k_{\alpha,t}$ can be viewed as the transition kernel of a Markov chain associated to the symmetric graph on Y. The operators $\{A_\alpha^t\}$ will be approximated by the discrete sum,

$$A_\alpha^t f(y_i) = \sum_{j=1}^{N} k_{\alpha,t}(y_i, y_j) f(y_j) q(y_j).$$

Furthermore, if $K^{(t)}$ is an $N \times N$ matrix, whose entries are $K_{ij}^{(t)} = k_{\alpha,t}(y_i, y_j)$, we have

$$K^{(t)} \begin{pmatrix} f(y_1) \\ \vdots \\ f(y_N) \end{pmatrix} = \begin{pmatrix} A_\alpha^t f(y_1) \\ \vdots \\ A_\alpha^t f(y_N) \end{pmatrix}.$$

In practice, we obtain $K^{(t)}$ by first constructing an N by N matrix W with entries

$$W_{ij} = h_t(y_i, y_j),$$

where h_t is the heat kernel defined as in (3.5). We then obtain

$$H = P^{-\alpha} W P^{-\alpha}, \tag{3.12}$$

where P is a diagonal matrix with $P_{ii} = \sum_{j=1}^{N} W_{ij}$. Next, we apply the weighted graph Laplacian normalization to H, and we get

$$K^{(t)} = D^{-\frac{1}{2}} H D^{-\frac{1}{2}}$$

In this regime, where $K^{(t)}$ is an approximation to $e^{-t\Delta}$ via the link of the continuous operator A_{α}^{t}, we expect that $K^{(t)}$ inherits the semi-group properties:

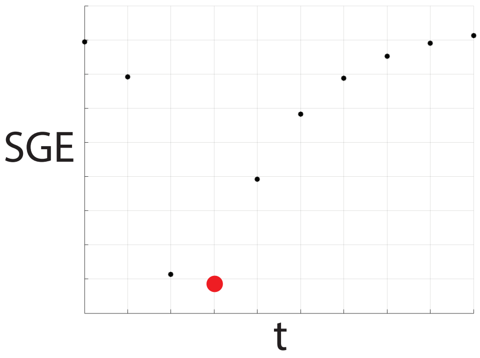
$$K^{(t_1+t_2)} \approx K^{(t_1)} K^{(t_2)}. \tag{3.13}$$

This means that when t is small, the matrix $K^{(t)}$ is indeed the t-th power of $K^{(1)}$.

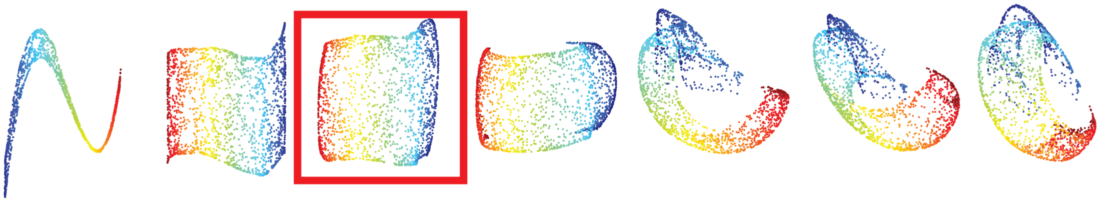
The following sections 3.3.2 - 3.3.1 describe two improvements enabled by the approximate identity (3.13): (1) a semi-group test for choosing the “right” or “best” value for the diffusion time parameter t; (2) an efficient way of computing the diffusion maps and distances on discrete samples.

3.3.1 Semi-group test

The semi-group property gives rise to a semi-group test for choosing the “right” t. The value of t cannot be too large in order for the approximate identity (3.13) to



(a) Semi-group error against different values of t. The red dot is the optimal value of t.



(b) Embedding of the Swiss roll by diffusion maps with increasing values of t. The red box indicates the optimal embedding.

Figure 3.4: Semi-group test for choosing the optimal t.

hold. On the other hand, t cannot be less than the smallest distance between the data objects. Since $K^{(t)}$ is a discretized approximation to A_α^t, the discretization dictates the minimum value of t.

To find the optimal t, we begin with initializing a wide range of discrete values for t. Call the initial set T. For each value t_i in T, we construct the diffusion matrices $\left(K^{(t_i)}\right)^2$ and $K^{(2t_i)}$. Because of the normalization step (3.12), both $\left(K^{(t_i)}\right)^2$ and $K^{(2t_i)}$ have norm one. We call the difference between $\left(K^{(t_i)}\right)^2$ and $K^{(2t_i)}$ as the *semi-group error* (SGE),

$$\mathrm{SGE} = \|\left(K^{(t_i)}\right)^2 - K^{(2t_i)}\|.$$

The suitable value for t should give a reasonably small SGE. Figure 3.4 demonstrates for points lying on a Swiss roll how to apply the semi-group test to obtain an optimal $\mathbf{R}^3$ embedding with the first three eigenvectors of $K^{(t)}$. The top figure

(a) shows the SGE values for t ranging from 2^{-10} to 2^{-1}. The optimal value of t is around 2^{-7}. The top figure (b) shows the embedded image of the Swiss roll by diffusion maps with increasing values of t. When t is too large, the embedded image was distorted by squashing along the side wall. As t decreases, the SGE also decreases. The embedded data points in unroll themselves to achieve the expected flattening effect. When t drops below the discretization threshold, the embedded image is only a curl with no thickness retained.

The semi-group test provides a good guiding principle for how to choose parameters in diffusion methods, which is often a difficult task in practice. In the next section, we shall see that the semi-group properties also give rise to a computationally efficient way of computing diffusion maps.

3.3.2 An alternative way of computing diffusion maps

Recall from Section 3.2.2 that the construction of a family of diffusion maps/distances consists of eigen-decomposing the continuous operatr A^t for each given diffusion time t. On discrete data, we first build the diffusion matrix $K^{(t)}$, followed by a matrix eigen-decomposition. However, this approach of computing diffusion maps/distances soon becomes computationally intractable when the number of data samples is large. A data set with N data objects will result in an $N \times N$ diffusion matrix $K^{(t)}$. When N is large, computing the eigenvalues and eigenvectors on $K^{(t)}$ is known to be slow. Although great progress have been made on improving time and accuracy for eigen-decomposition methods of large matrices, the problem still remains for the procedure is repeated at each t!

To address this issue, the key observation is that when t is small the diffusion matrix $K^{(t)}$ at time t is close to the t-th power of $K^{(1)} = K$ according to (3.13). The spectral properties of $K^{(t)}$ and K are related. Suppose the matrix K has eigenvectors $\mathbf{f}_l$ and the corresponding eigenvalues λ_l. Then the matrix $K^{(t)}$ have the same eigenvectors $\mathbf{f}_l$ and the eigenvalues λ_l^t are powers of the eigenvalues of K. Therefore, to compute the diffusion methods for all possible values of t, we only need one eigen-decompistion. Given $\lambda_l, \mathbf{f}_l$ as above, the diffusion distance D_α^t is computed by

$$D_\alpha^t(x,y) = \left(\sum_{l=1}^{N} \lambda_l^{2t} \left(\mathbf{f}_l(x) - \mathbf{f}_l(y)\right)^2\right)^{1/2}. \tag{3.14}$$

Similar to the original construction of diffusion maps, we can also truncate (3.14) up to k and get

$$D_\alpha^t(x,y) = \left(\sum_{l=1}^{k} \lambda_l^{2t} \left(\mathbf{f}_l(x) - \mathbf{f}_l(y)\right)^2\right)^{1/2}.$$

The diffusion map is then given as

$$f_\alpha^t(y_i) = (\lambda_1^t \mathbf{f}_1^{[i]}, \ldots, \lambda_k^t \mathbf{f}_k^{[i])}).$$

As before, $\mathbf{f}_1^{[i])}$ denotes the i-th entry of the eigenvector $\mathbf{f}_1$.

In practice, the semi-group approach for computing diffusion maps/distances consists of: (1) pick a good t_{init} with the semi-group test; (2) build the diffusion matrix $K^{(t_{\text{init}})}$; (3) compute eigenvalues and eigenvectors of $K^{(t_{\text{init}})}$, $K^{(t_{\text{init}})}\mathbf{f}_l = \lambda_l \mathbf{f}_l$, for $l = 1, 2, \ldots, N$; (4) build the diffusion maps/distances by λ_i^r, where $t = r \cdot t_{\text{init}}$.

The benefit of this alternative approach comes from the fact that $K^{(t)}$ is a good approximation to $e^{-t\Delta}$ only when t is small. When t is large, the link between $K^{(t)}$ and $e^{-t\Delta}$ no longer holds, and therefore the resulting diffusion maps/distances computed with $K^{(t)}$ do not indicate the underlying data structure. However, by raising the powers of the eigenvalues, meaning that we can diffuse longer, we bring more information into the horizon by propagating and accumulating more, and consequently, improve the overall characterization of the underlying manifold.

Chapter 4

Diffusion on fibre bundle

This chapter introduces the what and why of fibre bundle in data analysis. Section 4.1 deals with the assumption of viewing data as lying on a fibre bundle. Section 4.2 reviews diffusion maps on fibre bundle. In Section 4.3, we propose a projected kernel method for computing the low-dimensional representation of the data under noise.

4.1 Fibre bundle assumption

So far we have considered those data sets consisting of *data objects* (e.g., images, point clouds, triangular meshes, etc.), where each data object contains *data points*[1](e.g., RGB values, *xyz* coordinates, etc.). In addition, between similar data objects there exist pairwise structural correspondences; each correspondence is a map defined from a source data object (collection of all source data points) to a target data object (collection of all target data points).

Fibre bundle provides a natural description of a geometric model that puts together the data objects, data points and pairwise structural correspondences. Our basic assumption is that the data points lie approximately on a fibre bundle. In this section, we discuss the fibre bundle assumption and review diffusion maps defined on fibre bundles.

A fibre bundle is a manifold that locally looks like the product space. Formally,

---

[1] Note that the number of data points in each data object may differ.

the definition is given as follows.

Definition (Fibre Bundle). A *fibre bundle* is a quadruple (E, M, F, π) where E, M, F are topological spaces and $\pi : E \to M$ is a projection satisfying the local triviality condition, i.e., for every $p \in E$, there is an open neighborhood $U \subset M$ of $\pi(x)$ and a diffeomorphism

$$\phi : \pi^{-1}(U) \to U \times F$$

called a *local trivialization* of E such that the following diagram commutes:

$$\begin{tikzcd} \pi^{-1}(U) \arrow[r, "\phi"] \arrow[d, "\pi"] & U \times F \arrow[dl, "\pi_1"] \\ U & \end{tikzcd} .$$

We call E the *total space*, M the *base space* and F the *fibre space*. For simplicity, we use E to denote the fibre bundle quadruple (E, M, F, π). Unless otherwise stated, we assume throughout this paper that M and F are orientable Riemannian manifolds[2].

It follows from this definition that fibre bundle is the union of copies of F "glued" onto each point in M. The way in which the fibers are "glued" together is determined by the the structure of the total space E and the projection map $\pi : E \to M$. For any $x \in M$, $\pi^{-1}(x)$ is diffeomorphic to F. We denote F_x for $\pi^{-1}(x)$ and call it the *fibre over* $x \in M$. A local trivialization gives a way of locally "unwinding" the fibers into the topologically-simpler product space. For our fibre bundle assumption, we may think the space of data objects as M and for each data object $x \in M$, the data points of x lie approximately on F_x.

In addition to data objects and data points, the data sets to which the fibre bundle assumption applies, are equipped with pairwise correspondence maps between data

---

[2] Volume form and integration are well-defined for orientable Riemannian manifolds.

objects. A map between two data objects $x, y \in M$ can be seen as an approximation of the parallel transport between F_x and F_y. Fix an Ehresmann connection on the fibre bundle E. A parallel transport $P_{yx} : F_x \to F_y$ to y from x maps a point u in F_x to a point v in F_y. Intuitively, P_{yx} identifies elements in the fibre spaces F_x and F_y. In some sense, this is what a correspondence map does. From now on, we shall use P_{yx} to denote the correspondence map to y from x.

The correspondence map P_{yx} can be obtained by matching algorithms, e.g., the *Hungarian algorithm* in graph theory, the *Continuous Procrustes* in geometry processing, the *transport plans* in optimal transportation. However, the computed P_{xy} only exists (or are of high fidelity) if x, y are nearby on M. For x, y that are distant on M, P_{yx} can be defined via composing together the correspondences between nearby points along "small hops",

$$P_{yx} = P_{yx_n} \circ \ldots P_{x_2 x_1} \circ P_{x_1 x},$$

provided that the base manifold is well-populated. Although the correspondences defined this way will in general not be consistent (as composing along different paths lead to different correspondences), this inconsistency would reflect the *curvature* and *holonomy* of the connection on the fibre bundle.

In practice, we are usually given: (1) within each data object, abundant and clean data points; (2) between data objects, noisy and inaccurate distances generated by their correspondence maps. The number of data objects is in general way less than the number of data points lying on them. Our goal is to extract, as much and/or accurate as possible, information on the base manifold from data points lying on the fibre and the maps between them. The main tool that we used is diffusion maps on fibre bundles.

4.2 Diffusion maps on fibre bundle

We shall generalize diffusion maps on the machinery we have just built. As before, the key ingredient is the diffusion kernel. Here, we present two developments, *horizontal* and *coupled* diffusion kernels, by [Gao16].

4.2.1 Horizontal diffusion

The first approach is motivated by a *horizontal lift* of random walk on the base manifold to the total space of the fibre bundle. By local trivialization, a point $e \in E$ can be written as (x, v), where $x \in M$ and $v \in F_x$. Similarly, $e' \in E$ can be written as (y, w) where $y \in M$ and $u \in F_y$. For one step, e is allowed to jump to e' only when they can be identified with the parallel transport P_{xy}. In the continuous limit, the diffusion process on the fibre bundle can be viewed as the horizontal lift of the limit diffusion process on the base manifold.

More precisely, we define a kernel on the fibre bundle by a kernel on the base manifold. Let $e = (x, v)$ and $e' = (y, w)$ as before,

$$k_t(e, e') = k_t(x, y) = \begin{cases} \exp\left(-\frac{d_M^2(x,y)}{t}\right) & \text{if } w = P_{yx}v \\ 0 & \text{otherwise} \end{cases}$$

where $d_M(x, y)$ is the geodesic distance between x and y on M.

For any $f \in C^\infty(E)$, we define the horizontal diffusion operator $H_t : C^\infty(E) \to C^\infty(E)$ as

$$H_t f(e) = H_t(x, v) = \int_M k_t(x, y) f(y, P_{yx}v) dvol_M(y), \tag{4.1}$$

where $e = (x, v)$ by local trivialization and $x \in M, v \in F_x$.

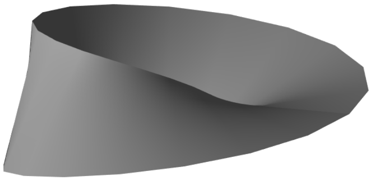


Figure 4.1: A Mobius strip of half-width 0.5 and of radius 1 at height $z = 0$.

If the density of data is non-uniform on E, we can apply the normalization trick as in 3.2 to obtain an anisotropic kernel $k_{\alpha,t}$ and its corresponding diffusion operator H^t_α, for any given normalization parameter $\alpha \in [0, 1]$. When $\alpha = 1$, the infinitesimal generator of H^t_α is Δ_H the *rough horizontal Laplacian* defined as in [Gao16] on E,

$$\lim_{t \to 0} H^t_\alpha = Ce^{-t\Delta_H}, \tag{4.2}$$

where C is positive constant depending only on the base manifold M.

We now illustrate the horizontal diffusion kernel on a Möbius strip, the simplest non-trivial fibre bundle. The Mobius strip in Figure 4.1 is parametrized by

$$\begin{aligned} x &= 1 + r\cos\left(\frac{a}{2}\right)\cos(a) \\ y &= 1 + r\cos\left(\frac{a}{2}\right)\sin(a) \\ z &= r\sin\left(\frac{a}{2}\right) \end{aligned}$$

for r in $[-0.5, 0.5]$ and a in $[0, 2\pi)$. We denote $\mathcal{X}$ for the Mobius strip,

$$\mathcal{X} = \{(x(a,r), y(a,r), z(a,r)) \mid -0.5 \le r \le 0.5, 0 \le a \le 2\pi\}.$$

Let $e, e' \in \mathcal{X}$. Their extrinsic coordinates are given by the xyz coordinates above, while their intrinsic coordinates are given by local trivialization $e = (a, r)$ and $e' =$

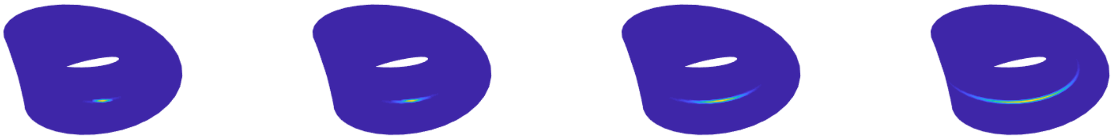


Figure 4.2: Diffusion effects of horizontal kernel of of increasing values of t.

(a', r'), where $a \in [0, 2\pi)$ and $r \in [-0.5, 0.5]$. The geodesic distance on the base circle is given by $d_M^2(x, y) = (a - a')^2$. The horizontal diffusion kernel on the Möbius strip is then defined as

$$k_t(e, e') = \begin{cases} \exp\left(-\frac{(a-a')^2}{t}\right) & \text{if } r' = P_{a'a} r \\ 0 & \text{otherwise} \end{cases}$$

where $P_{a'a}$ is the parallel transport to a' from a. Figure 4.2 illustrates the function values of $k_t(e, \cdot)$ on the Mobius strip $\mathcal{X}$, where e is a point on the side wall of $\mathcal{X}$. As t increases, the non-trivial values of $k_t(e, \cdot)$ spread and wrap around the Mobius strip horizontally.

4.2.2 Coupled diffusion

The coupled diffusion operator is derived from viewing the total fibre bundle space as a manifold. Let $e = (x, v)$ and $e' = (y, w)$ be given as before. Define

$$k_{t,\gamma}(e, e') = k_{t,\gamma}(x, v; y, w) = \exp\left(-\frac{d_M^2(x, y)}{t}\right) \exp\left(-\frac{d_{F_y}^2(P_{yx} v, w)}{\gamma t}\right).$$

We denote by $d_M(\cdot, \cdot)$ the geodesic distance on M and $d_{F_y}(\cdot, \cdot)$ the geodesic distance on F_y. As before, t is the diffusion time parameter. The parameter γ takes into consideration of the different sampling rate in the base manifold and the fibre manifold.

For any $f \in C^\infty(E)$, we define the coupled diffusion operator $H_{t,\gamma} : C^\infty(E) \to$

$C^\infty(E)$ as

$$H_{t,\gamma}f(x,v) = \int_M \int_{F_y} k_{t,\gamma}(x,v;y,w) f(y,w) dvol_{F_y}(w) dvol_M(y).$$

Again, we can obtain the normalized kernel $H^t_{\alpha,\gamma}$, for $\alpha \in [0,1]$. If $\alpha = 1$ and γ is finite, then the infinitesimal generator of $H^t_{\alpha,\gamma}$ is Δ_E the Laplace-Beltrami operator on the fibre bundle,

$$\lim_{t\to 0} H^t_{\alpha,\gamma} = Ce^{-t\Delta_E}, \tag{4.3}$$

where C is a positive constant.

Now we illustrate the coupled diffusion kernel on the Möbius strip $\mathcal{X}$ that we defined in the previous section. The coupled diffusion kernel on $\mathcal{X}$ with $e = (a, r)$ and $e' = (a', r')$ is given by

$$k_{t,\gamma}(e,e') = \exp\left(-\frac{(a-a')^2}{t}\right) \exp\left(-\frac{(P_{aa'}(r) - r')^2}{\gamma t}\right).$$

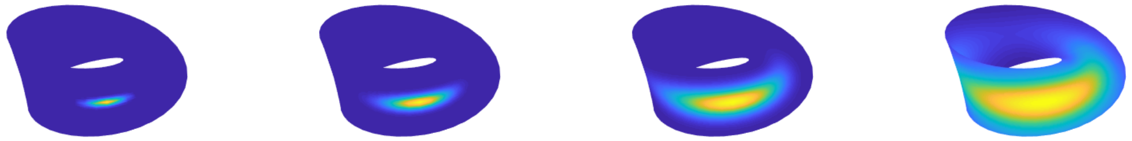
Fix e, the values of $k_{t,\gamma}(e,\cdot)$ on $\mathcal{X}$ is shown in Figure 4.3. First, as t increases, the non-trivial values of $k_{t,\gamma}$ spread away from e along all directions on the Mobius strip $\mathcal{X}$. Second, a different choice of γ suggests a different diffusion rate on $\mathcal{X}$.

4.2.3 Diffusion maps

With no surprise, diffusion maps on fibre bundle can be defined via eigenvalues and eigenfunctions of the diffusion operators we introduced above.

Let $E = (E, M, F, \pi)$ denote a fibre bundle as usual. Let $H^t_{\alpha,\gamma}$ be a diffusion operator (horizontal or coupled) on E. Fix α, t and γ, and suppose that $H^t_{\alpha,\gamma}$ has

(a) t increases →



(b) a smaller γ

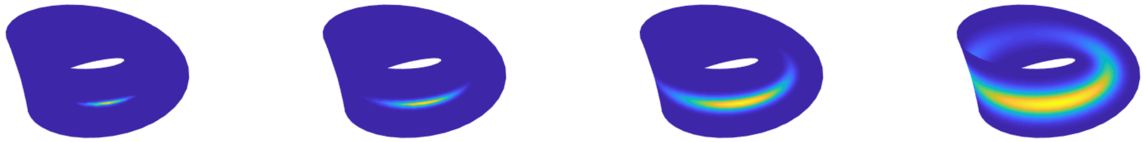


Figure 4.3: Diffusion effects of coupled diffusion kernel of of increasing values of t.

eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and the corresponding eigenvectors f_i for eigenvalue λ_i, the k dimensional *diffusion map* on E is defined as

$$\phi^t_{\alpha,\gamma}(e) = (\lambda_1 f_1(e), \ldots, \lambda_k f_k(e)), \quad e \in E.$$

From a similar argument as in 2.1.2, the diffusion distance on fibre bundle is defined as

$$D^t_{\alpha,\gamma}(e, e') = ||\phi^t_{\alpha,\gamma}(e) - \phi^t_{\alpha,\gamma}(e')||.$$

On discrete dataset sampled from E, we are given N data objects on the base manifold, $\{x_i \in M \mid 1 \leq i \leq N\}$, and each x_i has n_i data points, $x_i = \{v_{i,l}\}_{l=1}^{n_i}$. The diffusion map is computed by a sequence of re-normalization on the diffusion weight matrix W, an $N \times N$ block matrix. Let W_{ij} denote the i, j-th block and its l, m-th entry is given by

$$[W_{ij}]_{l,m} = k_{t,\gamma}(x_i, v_{i,l}; x_j, v_{j,m}),$$

where $k_{t,\gamma}$ can be the horizontal or coupled diffusion kernel as before. Applying the normalization trick, we define

$$W^t_{\alpha,\gamma} = D^{-\alpha} W D^{-\alpha},$$

where D is a diagonal matrix, whose entries are the row sum of W. The diffusion matrix is defined by

$$H^t_{\alpha,\gamma} = D^{-1/2} W^t_{\alpha,\gamma} D^{-1/2},$$

where D is the row sum of $W^\alpha_{t,\gamma}$. Let λ_l, f_l be eigenvalues and eigenvectors of $H^t_{\alpha,\gamma}$,

$$H^t_{\alpha,\gamma} f_l = \lambda_l f_l,$$

for $l = 1, 2, \ldots, n_1 + n_2 + \cdots + n_N$. We remark here that f_l can be thought as a concatenation of N segments of length $n_1, n_2, \ldots, n_N$ respectively.

The diffusion map provides for each data point $v_{i,l}$ a new parametrization in $\mathbb{R}^k$,

$$\phi^t_{\alpha,\gamma}(v_{i,l}) = \left(\lambda_1 f_1^{[i,l]}, \ldots, \lambda_k f_k^{[i,l]}\right),$$

where $f_k^{[i,l]}$ denotes the l-th entry in the i-th segment of f_k. The diffusion distance between two data point $v_{i,l}$ and $v_{j,m}$ is given by the Euclidean distance in the embedded space,

$$d^t_{\alpha,\gamma}(v_{i,l}, v_{j,m}) = ||\phi^t_{\alpha,\gamma}(v_{i,l}) - \phi^t_{\alpha,\gamma}(v_{j,m})||.$$

Again, the diffusion distance is equipped with a random walk interpretation. By the conservation property of the kernel, diffusion distance computes the transition probability from $v_{i,l}$ to $v_{j,m}$ over all possible short paths connecting them, and therefore can be considered as an improved/denoised distance that reflects the connectivity of the points. In consequence, the diffusion maps, which re-parametrize the data points according to their mutual diffusion distance, provide a global registration for all fibres. More precisely, one can reconstruct the pairwise correspondence maps in the embedded space, and obtain a denoised version of the original correspondence maps.

As with the traditional diffusion method on a manifold, the semi-group properties of $e^{-t\Delta_E}$ and $e^{-t\Delta_H}$ also provide an alternative way of computing diffusion maps on a fibre bundle. The construction procedure for $\phi^t_{\alpha,\gamma}$ with α, t, γ given is as follows. We begin with an initial value t_{init}. Note that a similar semi-group test can be implemented here for the right value of t_{init}. We then construct the $N \times N$ block diffusion kernel matrix $H^t_{\alpha,\gamma}$ and compute its eigenvalues and eigenvectors λ_l, f_l for $l = 1, \ldots, n_1 + n_2 + \cdots + n_N$. Let $t = rt_{\text{init}}$. Finally, we build the diffusion map $\phi^t_{\alpha,\gamma}$ by raising powers of the eigenvalues,

$$\phi^t_{\phi,\gamma}(v_{i,l}) = \left(\lambda_1^r f_1^{[i,l]}, \ldots, \lambda_k^r f_k^{[i,l]}\right).$$

4.3 Diffusion on the base manifold

As we have already seen, diffusion maps on fibre bundle improve the correspondence maps. We must point out that we are not mainly concerned with the correspondence maps, but rather, the data objects on the base manifold. This section deals with the diffusion process on the base manifold. We begin with a review on the base diffusion maps and distances defined by [Gao16]. In Section 4.3.2, we introduce the projected kernel method which derives the diffusion kernel on the base manifold from the diffusion kernel on the entire fibre bundle. In Section 4.3.3, we test the performance of the projected kernel method under noise on simulated examples.

4.3.1 Base diffusion maps and distances

As before, let $\{x_i \mid 1 \le i \le N\}$ denote a set of N data objects on the base manifold. Let $H^t_{\alpha,\gamma}$ denote the diffusion kernel matrix defined as above. Let λ_k be the k-th largest eigenvalue of $H^t_{\alpha,\gamma}$ and f_k be its corresponding eigenvector.

Recall that $H^t_{\alpha,\gamma}$ is an $N \times N$ block matrix. For each block $H_{ij} = \left[H^t_{\alpha,\gamma}\right]_{ij}$, we compute its Frobenius norm as

$$\begin{aligned} ||H_{ij}||_F^2 &= \mathrm{Tr}\left[(H_{ij})(H_{ij})^\intercal\right] \\ &= \mathrm{Tr}\left[\sum_{k,k'} \lambda_k \lambda_{k'} f_k^{(i)} f_k^{(i)\intercal} f_{k'}^{(j)} f_k^{(i)\intercal}\right] \\ &= \mathrm{Tr}\left[\sum_{k,k'} \lambda_k \lambda_{k'} f_k^{(i)\intercal} f_k^{(i)} f_k^{(i)\intercal} f_{k'}^{(j)}\right] \\ &= \sum_{k,k'} \lambda_k \lambda_{k'} f_k^{(i)\intercal} f_k^{(i)} f_k^{(i)\intercal} f_{k'}^{(j)}. \end{aligned}$$

Here, $f_k^{(i)}$ denotes the i-th segment of f_k. The *base diffusion maps* (BDM) is then defined as

$$\phi^{\alpha,t,\gamma}_{\text{base}}(x_i) = \left(\lambda_k^{1/2} \lambda_{k'}^{1/2} f_k^{(i)\intercal} f_{k'}^{(i)}\right)_{0 \le k,k'}.$$

The *base diffusion distance* (BDD) is given by

$$d^{\alpha,t,\gamma}_{\text{base}}(x_i, x_j) = ||\phi^{\alpha,t,\gamma}_{\text{base}}(x_i) - \phi^{\alpha,t,\gamma}_{\text{base}}(x_j)||.$$

Base diffusion maps outperform standard diffusion maps on data samples with high spatial complexity, e.g. molar teeth of primates in [Gao19]. However, we came to realize that the diffusion kernel is indeed the primary object of dimension reduction. We shall see in later chapters that our shift of attention from the embedding maps to kernels has far-reaching implications. For now, let us postpone the discussion on why we should care about kernels and introduce the projected kernel method.

4.3.2 Base diffusion kernel

Following the same notation as in 4.3.1, we note that a new diffusion kernel H_{base} on the base manifold can be constructed from taking the Frobenius norm of each block in $H^t_{\alpha,\gamma}$. To write it more explicitly, we set

$$[H_{\text{base}}]_{ij} = ||H_{ij}||_F = \sqrt{\sum_{m=1}^{n_j} \sum_{l=1}^{n_i} ([H_{ij}]_{l,m})^2}, \tag{4.4}$$

where $[H_{ij}]_{l,m}$ denotes the l,m-th entry in the i,j-th block of $H^t_{\alpha,\gamma}$.

The conservation property says that $[H_{ij}]_{l,m}$ is the transition probability of a Markov process on E. Eq. (4.4) indicates that $||H_{ij}||_F$ sums the probability from all data points in object i to all data points in object j over all possible short paths connecting them. Since we now consider all possible paths *and* all possible points in the data objects, the new base kernel will be even more robust to noise. To make H_{base} a transition kernel, we apply the normalization trick as before. Once we obtain the normalized H_{base}, we apply the regular diffusion map, i.e, compute an embedding for x_i in $\mathbb{R}^k$ by a spectral decomposition of the normalized H_{base},

$$\phi_{\text{base}}(x_i) = (\lambda_1 f_1^{(i)}, \ldots, \lambda_k f_k^{(i)}).$$

This procedure can also be considered as using fibre bundle diffusion to denoise the diffusion kernel on the base manifold, followed by a regular diffusion map on the improved kernel. For notational convenience, the procedure is called the *projected kernel method*, as the diffusion kernel on the base manifold is indeed taking an average of the blocks in the fibre bundle kernel. In the next section, we test the performance of the projected diffusion kernel method against noise.

4.3.3 Experiments on simulated examples

As indicated above, the projected diffusion kernel method can be more robust to noisy and inaccurate distances on the base manifold. To compare the performance of the traditional diffusion maps (DM) and the projected kernel method, we simulate a scenario that is close to what we have in practice, that is, there are clean and abundant data points within each data object, yet the distances between the data objects are noisy due to errors in their correspondence maps.

For simplicity, we begin with a cylindrical surface. The base manifold is a unit circle centered at the origin; to each point on the base circle we attach a unit interval. We generate a discrete dataset from the cylindrical surface with the following steps. We first pick 25 equally spaced points on $[0, 2\pi)$. We denote U as the set of the chosen points. The points on the base circle are then computed by $(\cos(u_i), \sin(u_i))$, where $u_i \in U$. On the fibre space, i.e., the unit interval $[0, 1]$, we generate 10 equally spaced points, and denote the set by V. In total, the points on the cylinder surface are given as

$$\{(\cos(u_i), \sin(u_i), v_l) \in \mathbb{R}^3 \mid i = 1, \ldots, 25, l = 1, \ldots, 10\}.$$

The sampled points on a continuous cylindrical surface is shown as colored dots in Figure 4.4.

Note that a data object is a strip of vertical points on the cylinder. For con-

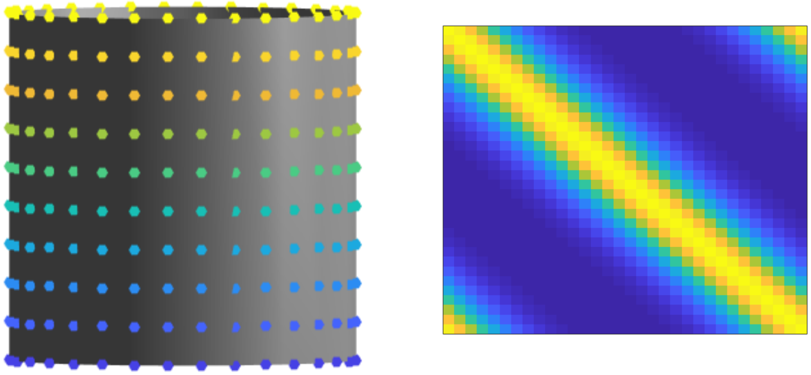


Figure 4.4: Left: Illustration of 25×10 points sampled on a simple cylinder surface. Right: diffusion kernel defined on the base circle with no noise.

venience, let us denote x_i as the i-th data object on the base circle, where $x_i = (\cos(u_i), \sin(u_i))$. Let Y_i denote the set of data points lying on the unit interval attached to x_i, where

$$Y_i = \{(\cos(u_i), \sin(u_i), v_l) \in \mathbb{R}^3 \mid l = 1, \ldots, 10\}.$$

Let x_i, x_j be two data objects and Y_i, Y_j denote the set of data points attached to x_i, x_j respectively. The true geodesic distance between x_i and x_j is given by the arc length on the circle,

$$d_M^2(x_i, x_j) = (u_i - u_j)^2$$

In practice, we do not know a priori the parametrization of the data objects on the base manifold. In fact, the base manifold is what we are interested to learn! We often estimate d_M with the Euclidean distance between Y_i and the transported Y_j under P_{ij}. In the cylindrical example, we compute

$$d_M^2(x_i, x_j) \approx \sum_{l=1}^{10} (\cos(u_i) - \cos(u_j))^2 + (\sin(u_i) - \sin(u_j))^2 + (v_l - P_{ij} v_l)^2.$$

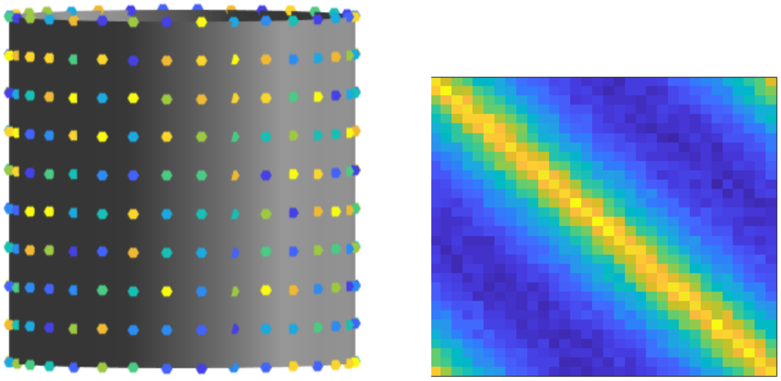


Figure 4.5: Left: Illustration of randomly permuted points on cylinder surface. Right: noisy diffusion kernel defined on the base circle.

Since the cylinder $S^1 \times [0,1]$ is a trivial fibre bundle, the parallel transport is indeed the identitiy map. We therefore have

$$d_M^2(x_i, x_j) \approx \sum_{l=1}^{10} (\cos(u_i) - \cos(u_j))^2 + (\sin(u_i) - \sin(u_j))^2.$$

To simulate noise on $d_M^2(x_i, x_j)$ from the correspondence maps, we generate a random permutation on $\{v_l \mid l = 1, \ldots, 10\}$, call it P_{ij}^{noise}. Then set

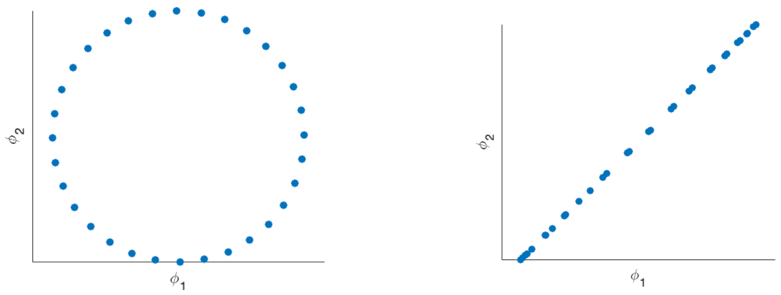
$$d_{\text{noise}}^2(x_i, x_j) = \sum_{l=1}^{10} (\cos(u_i) - \cos(u_j))^2 + (\sin(u_i) - \sin(u_j))^2 + (v_l - P_{ij}^{\text{noise}} v_l)^2.$$

The random permutation is illustrated in Figure 4.5.

We apply DM and the projected kernel method to the dataset we described above with d_M and d_{noise}. We shall only illustrate the procedure with d_{noise} due to its similarity with d_M. For DM, we construct the kernel matrix by

$$W_{ij} = \exp\left(-\frac{d_{\text{noise}}(x_i, x_j)^2}{t}\right).$$

Apply the normalization trick with $\alpha = 1$ and obtain the diffusion matrix H_{DM}. For



(a) DM (b) Projected kernel method

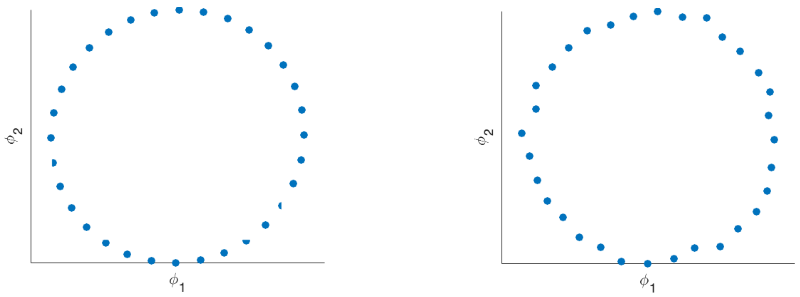
Figure 4.6: Embed points on cylinder by DM ($t = 1$) and projected method kernel ($t = \gamma = 1$) without noise.

the projected diffusion method, we first construct the $N \times N$ block diffusion weight matrix W. Let $[W_{ij}]_{lm}$ denote the l, m-th entry in the i, j-th block. Define

$$[W_{ij}]_{lm} = \exp\left(-\frac{d_{\text{noise}}(x_i, x_j)^2}{t}\right) \exp\left(-\frac{||v_l - P_{ij}^{\text{noise}}(v_m)||^2}{\gamma t}\right).$$

Apply the normalization trick with $\alpha = 1$ and obtain the $N \times N$ block diffusion matrix. For each block, we compute its Frobenius norm and acquire the base diffusion matrix H_{proj}. Now we compute the eigenvalues and eigenvectors, and plot the data parametrized by the first two eigenvectors in Figures 4.6, 4.7. Where there is no noise, both DM and projected kernel recovers the instrinsic data object structure, which is the base cirlce. When we add noise, DM fails to recognize the base circle in all experimenting parameters.

A similar experiment can be done on a torus (Figure 4.8) $S^1 \times S^1$, where we attach a unit circle to each point on the base unit circle. We sampled 16 uniformly spaced points on the base circle and 64 points on the fibre circle and added noise to the pairwise correspondence maps by randomly permuting points on the fibre circle. Results with the optimal diffusion time parameters are shown in Figure 4.8. Again, the projected kernel method fully recovers the perfect circle on the base manifold,



(a) DM (b) Projected kernel method

Figure 4.7: Embed points on cylinder by DM ($t = 1$) and projected method kernel ($t = \gamma = 1$) under noisy correspondence maps.

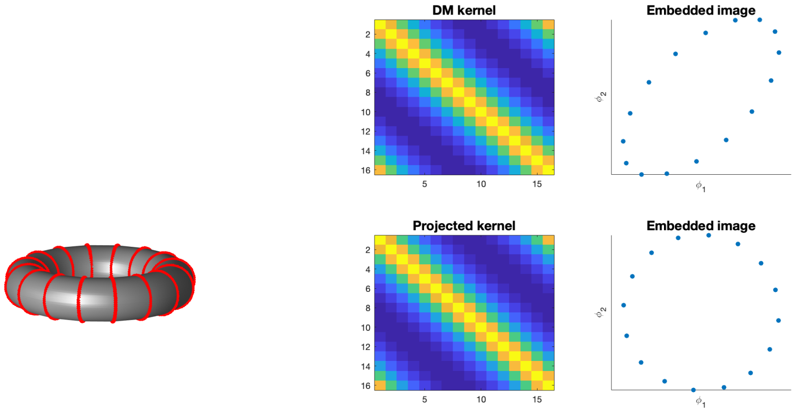


Figure 4.8: Embedding with DM and projected kernel method a torus

while the diffusion kernel only outputs a skewed circle at its optimized parameter.

Chapter 5

Kernel in dimension reduction

In this chapter, we explain in the context of regression why kernel plays an important role in dimension reduction. Section 1 introduces the problem of regression in general. Section 2 formulates the dimension reduction problem as a problem of regression. We illustrate the benefits of thinking the dimension reduction problem in a probabilistic way by PCA. The final section discusses Gaussian process, the best linear predictor for regression, from which we establish the connection between kernel and dimension reduction.

5.1 The Problem of Regression

The main concern of regression is predicting continuous quantities from a discrete set of observations. Suppose we are given a data set with N observations

$$\{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}, 1 \leq i \leq N\},$$

where $\mathcal{X}$ and $\mathcal{Y}$ are two data spaces. The variables x_i's are usually called input variables (or covariates), and hence the space $\mathcal{X}$ is named the input space. The variables y_i's are called the output or dependent variables, and hence the space $\mathcal{Y}$ is called the output space. Recall from previous chapters, the data spaces come in varieties: Euclidean space, manifold or fibre bundle. For simplicity, let us assume for now that $\mathcal{Y}$ is the real line $\mathbb{R}$. We are interested in making inference about the relationships between the input and output variables, i.e., we would like to learn a function $f : \mathcal{X} \to \mathbb{R}$ so that $f(x_i)$ is as close to y_i as possible.

In the following sections, we provide three ways of formulating the regression problem.

5.1.1 Deterministic approach

The deterministic approach deals with the regression problem by optimization. Let H denote a set of functions mapping $\mathcal{X}$ to $\mathbb{R}$. The regression problem is to find

$$\arg\min_{f \in H} \sum_{i=1}^{N} L(f(x_i), y_i).$$

We denote by $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ the *loss function* used to measure the difference between $f(x_i)$ and y_i. The most common choice of L is the *squared error loss* defined as,

$$L(y_i, y_j) = (y_i - y_j)^2, \; y_i, y_j \in \mathbb{R}.$$

5.1.2 Generative model approach

An alternative way to formulate the regression problem is via a generative model, from which the squared error loss function arises naturally.

Let us assume that the input values and output values are related by the following statistical model

$$y_i = f(x_i) + \epsilon_i, \; \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{5.1}$$

where ϵ_i is an error term that describes either random noise or un-modeled effects. Furthermore, we assume that ϵ_i is distributed i.i.d (independently and identically distributed) according to a normal distribution. We can also write (5.1) as

$$y_i \sim \mathcal{N}(f(x_i), \sigma^2).$$

The regression problem concerns finding an appropriate f. Suppose the function f is parametrized by θ, denoted as f_θ. The principle of *maximum likelihood* leads to a natural criterion for this. We should choose θ to make y_i as high probability as possible. By the independence assumption of ϵ_i, we compute the *likelihood function* of θ as,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} p(y_i \mid x_i, \theta) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(f_\theta(x_i) - y_i)^2}{2\sigma^2}\right) \end{aligned}$$

Equivalently, we can maximize the *log likelihood* given as,

$$l(\theta) = N \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (f_\theta(x_i) - y_i)^2.$$

This is the same as minimizing the squared error loss,

$$\sum_{i=1}^{N} (f_\theta(x_i) - y_i)^2.$$

5.1.3 Random variable approach

The third way to define the regression problem resorts to random variables and probability spaces.

The following set-up is defined in [FHT01]. Let $X \in \mathcal{X}$, $Y \in \mathbb{R}$ be random variables with joint probability distribution $p(x, y)$. The regression problem consists of choosing f such that the *expected prediction error* is minimized,

$$\text{EPE}(f) = \mathbb{E}(L(f(X), Y),$$

where L denotes the loss function as before.

If L is the squared error loss, we have

$$\begin{aligned}\text{EPE}(f) &= \mathbb{E}(f(X) - Y)^2 \\ &= \int (y - f(x))^2 p(x, y) dx dy\end{aligned}$$

The solution is the conditional expectation,

$$f = \mathbb{E}(Y \mid X = x).$$

We remark that the function f is non-parametric, because f does not take a predetermined from and is constructed from scratch based on information from the data. This view, which abandons an explicit form of f, but rather studies f under a probability distribution on the function space, in fact opens doors to new methods in dimension reduction.

5.2 Dimension Reduction as Regression

In this section we establish the dimension reduction problem as a problem of regression. This allows us to reexamine the previous methods and develop further techniques from a probabilistic view.

Our introductory example is Probabilistic PCA [TB99], which is formulated as a *linear regression* model: Given a dataset $\{y_i \in \mathbb{R}^d \mid 1 \leq i \leq N\}$ and latent variables

$x_i \in \mathbb{R}^k$ distributed i.i.d with a multivariate normal distribution,

$$\begin{aligned}
y_i &= \mu + W x_i + \epsilon_i, \\
x_i &\sim \mathcal{N}(0, I_k), \\
\epsilon_i &\sim \mathcal{N}(0, \sigma^2 I_d).
\end{aligned}$$

Our main objective is to use the observation y_i to improve our knowledge on the unknown parameters μ, W, σ^2.

The following observations are in order. The conditional distribution of y_i is given by,

$$p(y_i \mid \mu, W, \sigma^2) = \mathcal{N}(\mu + W x_i, \sigma^2 I_d).$$

The marginal distribution of y_i is given by,

$$p(y_i \mid \mu, W, \sigma^2) = \mathcal{N}(\mu, W W^\intercal + \sigma^2 I_d).$$

Let $C = W W^\intercal + \sigma^2 I_d$. The log-likelihood is then computed by

$$l = -\frac{N}{2} \left(d \ln(2\pi) + \ln |C| + \operatorname{Tr}(C^{-1} S) \right).$$

Here S denote the sample covariance matrix as defined in PCA. The maximum likelihood (MLE) solution for μ, W, σ^2 is summarized below.

$$\begin{aligned}
\mu_{\text{MLE}} &= \frac{1}{N} \sum_{i=1}^{N} y_i, \\
W_{\text{MLE}} &= U_k (\Lambda_k - \sigma^2 I_k)^{1/2}, \\
\sigma_{\text{MLE}} &= \frac{1}{d-k} \sum_{i=k+1}^{d} \lambda_i.
\end{aligned}$$

The columns of U_k are the eigenvectors corresponding to the k largest eigenvalues of S defined as before. The eigenvalues are stored in the diagonal matrix Λ_k. The deterministic PCA is recovered in the limit $\sigma^2 \to 0$.

Probabilistic PCA has several advantages as outlined in [TB99]:

1. flexibility to be combined as a probabilistic mixture model, thus extending the scope of traditional PCA;
2. ability to deal with missing data values via projection PCA;
3. free control on the degree of freedom as a covariance model of data;
4. potential to be solved with Bayesian methods.

5.3 Gaussian Process Regression

We have now seen the connection between dimension reduction and the problem of regression with an example of PCA. It is natural to ask why not use non-parametric regression to derive new dimension reduction methods? The precise answer will be postponed in Chapter 6. Before that, a crucial ingredient must be present, namely the Gaussian process. As mentioned before, Gaussian process is the best linear solution to a non parametric regression problem [Ras03]. We begin with a review on the definition of Gaussian process.

5.3.1 Gaussian process

Definition (Gaussian Process) A random function $f : \mathcal{X} \to \mathbb{R}$ with *mean function* $m : \mathcal{X} \to \mathbb{R}$ and *covariance function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *Gaussian Process* if any finite realization $\{f(x_i) \mid 1 \leq i \leq n\}$ is jointly normal, i.e.,

$$\{f(x_i) \mid 1 \leq i \leq n\} \sim \mathcal{N}(m_n, K_n),$$

with mean vector

$$m_n = m(x_1), \dots, m(x_n),$$

and covariance matrix

$$K_n = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ \vdots & & & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix}.$$

We use the notation $f \sim \mathcal{GP}(m, k)$. The domain $\mathcal{X}$ is called the *index set* of f.

Recall the single output regression problem is given by

$$f = \mathbb{E}(Y \mid X = x),$$

where X, Y are random input and output variables, defined as in 5.1.3. Under the Gaussian process model $f \sim \mathcal{GP}(m, k)$ with m and k given, the solution to the single-output regression problem (also known as the *best linear predictor*) is

$$\begin{aligned} f(x) &= \mathbb{E}(f(x) \mid f(x_1) = r_1, \dots, f(x_n) = r_n) \\ &= m(x) + k_n(x){K_n}^{-1}(R_n - m_n) \end{aligned}$$

where $k_n = (k(x, x_1), \dots, k(x, x_n))^\intercal$ and $R_n = (r_1, \dots, r_n)^\intercal$.

5.3.2 Gaussian process as the best linear predictor

The general problem of prediction using Gaussian process begins with a set of observed data $\{x_i, y_i\}_{i=1}^n$, where $\mathbf{x} = \{x_i\}_{i=1}^n$ are the input points and $\mathbf{y} = \{y_i\}_{i=1}^n$ are the output values. We would like to make a prediction about the value on the new point x^*.

From the Gaussian process assumption, $f \sim \mathcal{GP}(m, k)$, the $n+1$ random variables $(Y_1, Y_2, \ldots, Y_n, Y^*)$ have a joint multivariate distribution with mean vector

$$\mu = \left[m(x_1), m(x_2), \ldots, m(x^*)\right],$$

and covariance matrix K_+, where the entries in K_+ are given as

$$[K_+]_{ij} = k(x_i, x_j).$$

The $(n+1) \times (n+1)$ matrix K_+ consists of a $n \times n$ matrix K, a length n vector $\mathbf{k} = [k(x^*, x_1), \ldots, k(x^*, x_n)]$ and a scalar $k^* = k(x^*, x^*)$,

$$K_+ = \begin{bmatrix} K & \mathbf{k} \\ \mathbf{k}^\intercal & k^* \end{bmatrix}.$$

If $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, then the conditional distribution for Y^* is also normal,

$$Y^* \mid x^*, \mathbf{x}, \mathbf{y} \sim \mathcal{N}(\mu^*, k^*),$$

with

$$\begin{aligned} \mu^* &= \mathbf{k} K^{-1} \mathbf{y}, \\ k^* &= k^* - \mathbf{k} K^{-1} \mathbf{k}^\intercal. \end{aligned}$$

We often use μ^* as the predicted value for the new point x^* and k^* to indicate the uncertainty score or give information about the confidence region.

If we assume a Gaussian noise model as in (5.1), the mean and variance for the predicted value on x^* are given by

$$\begin{aligned} \mu^* &= \mathbf{k}(K + \sigma^2 I)^{-1}\mathbf{y}, \\ k^* &= k^* + \sigma^2 - \mathbf{k}(K + \sigma^2 I)^{-1}\mathbf{k}^\intercal. \end{aligned}$$

It is not difficult to see where the name of "best linear predictor" comes from, as the predicted value μ^* is a weighted average of the given values $\mathbf{y}$ and the weights are determined by the covariance function.

5.3.3 Gaussian process as dimension reduction

We end this chapter with a brief discussion on Gaussian process as dimension reduction. While classical dimension reduction techniques, like PCA, MDS and diffusion maps characterize the behavior of data objects by finding their low-dimensional representation, methods like Probabilistic PCA and Gaussian process aim at retrieving causes, the probabilistic generating mechanism from, effects, the observed data points. The duality – one finds the outer form/ representation of data, and the other finds the inner model/ generative scheme that is most consistent with the observed data – is fundamentally the same, with the same objective to understand the intrinsic structure that data underlies. Yet, the latter does not adhere to the external manifestation, and therefore retains more flexibility and opens doors to tools from probability, statistics, etc.

Gaussian processes are uniquely determined by a mean function m and a covariance function k. In practice, m is often chosen to be the sample mean and k can be described by a few parameters. The dimension reduction of Gaussian processes comes from summarizing the data generative process with parameters in the covariance function. In the next chapter, we shall see how to define Gaussian process on fibre bundles and how diffusion kernels arise as natural covariance functions.

Chapter 6

Gaussian process on fibre bundle

We shall now develop the main theme of this thesis. Our purpose is to examine basic probabilistic tools such as Gaussian process on fibre bundle.

Previously we defined Gaussian process on an arbitrary index set $\mathcal{X}$. We have seen that a Gaussian process $f \sim \mathcal{GP}(m, k)$ is uniquely determined by its mean m and covariance function k. We assume now $\mathcal{X} = E$ is a fibre bundle. The question we ask to ourselves is how to choose m and k on a fibre bundle? More specifically, how is the geometry of fibre bundle useful?

6.1 Diffusion kernel as covariance function

As mentioned in the previous chapter, the mean function is usually chosen to be the sample mean. The covariance function is indeed a special type of kernel. A criterion for a general kernel to be the covariance function of a Gaussian Process is given as follows,

If a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a covariance function, then k must satisfy:

1. symmetric: $k(\omega, \omega') = k(\omega', \omega)$.
2. positive semi-definite: for all $f \in L^2(\Omega)$

$$\int \int k(\omega, \omega') f(\omega) f(\omega') d\mu(\omega) d\mu(\omega') \geq 0.$$

Diffusion kernel arises as a natural candidate for covariance function on a fibre bundle. Given E a fibre bundle defined as before, let $e, e' \in E$. By local trivialization, we write $e = (x, v) \in M \times F_x$ and $e' = (y, w) \in M \times F_y$. Recall in Section 4.2, the diffusion kernel on a fibre bundle is defined in two different ways, which we abbreviate as follows.

1. horizontal diffusion kernel

$$k_t(e, e') = k_t(x, y) = \begin{cases} \exp\left(-\frac{d_M^2(x,y)}{t}\right) & \text{if } w = P_{yx}v \\ 0 & \text{otherwise} \end{cases}$$

   where $t > 0$. Here, P_{yx} is the parallel transport to y from x and $d_M(x, y)$ is the geodesic distance between x and y on M.

2. coupled diffusion kernel

$$k_{t,\gamma}(e, e') = k_{t,\gamma}(x, v; y, w) = \exp\left(-\frac{d_M^2(x, y)}{t}\right) \exp\left(-\frac{d_{F_y}^2(P_{yx}v, w)}{\gamma t}\right)$$

   where $t > 0$ and $\gamma > 0$. Let d_{F_y} denote the geodesic distance on the fibre attached to y, and P_{yx}, d_M be given as before.

For convenience, let $k_{t,\gamma}$ denote either the horizontal or coupled diffusion kernel. We make a small modification to the normalization procedure to make $k_{t,\gamma}$ symmetric. Fix $\alpha \in [0, 1]$, we then define the re-normalized kernel $k^t_{\alpha,\gamma}(e, e')$ by the following two steps:

1. Let

$$h^\alpha_{t,\gamma}(e, e') = \frac{k_{t,\gamma}(e, e')}{p^\alpha(e)p^\alpha(e')},$$

where

$$p(e) = \int k_{t,\gamma}(e, e')q(e')de'.$$

2. Let

$$k^t_{\alpha,\gamma}(e, e') = \frac{h^{\alpha}_{t,\gamma}(e, e')}{(d^{\alpha}(e))^{1/2}\,(d^{\alpha}(e'))^{1/2}},$$

where

$$d^{\alpha}(e) = \int h^{\alpha}_{t,\gamma}(e, e')q(e')de'.$$

It is clear from the re-normalization step that $k^t_{\alpha,\gamma}$ is symmetric. Now, to show that $k^t_{\alpha,\gamma}$ is positive definite, we resort to $H^t_{\alpha,\gamma}$ the integral operator associated with $k^t_{\alpha,\gamma}$. From eq.(4.2) and (4.3), we have $H^t_{\alpha,\gamma}$ is close to $e^{-t\Delta}$ when $\alpha = 1$, $\gamma < \infty$ and t is small. Since $e^{-t\Delta}$ is a positive operator, we have the positive definiteness of $k^t_{\alpha,\gamma}$.

In practice, the eigenvalues of the diffusion operator are not always positive due to numerical issues. We can enforce the positive-definiteness by an approximate kernel to the original k by simply replacing its zero and negative eigenvalues with very small positive values and construct the approximate kernel as,

$$\tilde{k}(e, e') = \sum_{\lambda_i > 0} \lambda_i \phi_i(e)\phi_i(e') + \sum_{\lambda_i \leq 0} c\phi_i(e)\phi_i(e'),$$

where c is a positive constant that is very close to 0.

This is analogous to applying a high-pass filter in imaging to the diffusion kernel. By taking only the large eigenvalues, we not only ensure the positive definiteness of the matrix, but also iron out the noise on the geometry of the input space.

6.2 Why diffusion kernel

We have shown that the diffusion kernel satisfies the two conditions for being a covariance function of Gaussian process on fibre bundles. Now we provide some intuition about why diffusion kernel is a good choice.

6.2.1 Geometric prior

First, the covariance function expresses our prior beliefs about the correlation and similarities of data. To see this, we recognize the interpolation in the Gaussian process solution

$$f(x) = m(x) + k_n(x)k_n^{-1}(R_n - m_n)$$

is a weighted average of all observed values of f, where the weights are determined by the covariance function. Choosing the diffusion kernel that gives rise to the Laplace-Beltrami operator as the covariance function hence encodes the geometric information of the underlying fibre bundle to the regression problem we wish to solve.

6.2.2 Stochastic process

Diffusion kernel is related to the covariance of a stochastic model on data. For the moment, let us return to the discrete world. Given a data set $\{e_i \in E \mid 1 \leq i \leq N\}$, consider the random field obtained by attaching independent, zero mean, σ^2 variance random variable $Z_i(0)$ to each data point e_i. Now let each of these random variable sends a fraction $t < 1$ of their values to their neighboring points. Two points $e_i \sim e_j$ are said to be neighboring points if $P_{xy}(v) = w$ and $x \neq y$, where $e_i = (x, v)$ and

$e_j = (y, w)$. We have at discrete times $s = 1, 2, \ldots$

$$Z_i(s+1) = Z_i(s) + t \sum_{e \sim e'} \left(Z_j(s) - Z_j(s) \right).$$

To see the connection with the diffusion kernel, we first define a weight matrix W by

$$W = \begin{bmatrix} h(e_1, e_1) & \ldots & h(e_1, e_N) \\ \vdots & & \vdots \\ h(e_N, e_1) & \ldots & h(e_N, e_N) \end{bmatrix},$$

where h is defined as

$$h(e, e') = \begin{cases} 1 & \text{if } w = P_{yx} v \\ 0 & \text{otherwise} \end{cases}.$$

We recognize its similarities with the horizontal and coupled diffusion kernel.

Next, we define $H = D - W$, hwere D is a diagonal matrix whose entries are $D_{ii} = \sum_{j=1}^{N} W_{ij}$. From H, we introduce the following operator,

$$T(s) = (1 + tH)^s.$$

Then $\mathbf{Z}(s) = [Z_1(s), \ldots, Z_n(s)]^\intercal$ can be written as

$$\mathbf{Z}(s) = T(s)\mathbf{Z}(0).$$

The covariance of the random field at time s is given by

$$\mathrm{Cov}_{ij}(s) = \sigma^2 T_{ij}(2s).$$

Now we decrease the time step in the operator T from 1 to Δs,

$$T(s) = \left(1 + \frac{tH}{1/\Delta s}\right)^{s/\Delta s}.$$

As Δs approaches 0, $T(s)$ becomes e^{tH}. In particular, the covariance becomes $\mathrm{Cov}_{ij}(t) = \sigma^2 e^{2tH}$.

6.2.3 Semi-group properties

As mentioned before, when t is small, the integral operator $H^t_{\alpha,\gamma}$ associated with $k^t_{\alpha,\gamma}$ is close to $e^{-t\Delta}$. Since $\{e^{-t\Delta}\}$ forms a semi-group, we expect that in this regime where $H^t_{\alpha,\gamma}$ is close to $e^{-t\Delta}$, we have advantages of the semi-group properties as in Chapter 3.

First, the semi-group properties allow us to obtain positive definiteness for larger t. Given t_{init} small, we spectral decompose $k^{t_{\text{init}}}_{\alpha,\gamma}$ as

$$k^{t_{\text{init}}}_{\alpha,\gamma}(e, e') = \sum_i \lambda_i \phi_i(e)\phi_i(e').$$

Let $t = rt_{\text{init}}$. We can build $k^t_{\alpha,\gamma}$ by

$$k^t_{\alpha,\gamma}(e, e') = \sum_i \lambda_i^r \phi_i(e)\phi_i(e'),$$

and it is not difficult to see that $k^t_{\alpha,\gamma}$ is symmetric and positive definite.

Second, we obtain computational savings from the semi-group properties. Computations with Gaussian process involves finding inverse of the covariance matrix K_n, which can often be time-consuming when the size of the matrix is large. Since the covariance function can be written so nicely by the eigenfunctions, we can write the inverse covariance matrix as

$$K_n^{-1} = \sum_i \lambda_i^{-1} \phi_i \phi_i^\intercal .$$

6.3 Example: Möbius strip

As an example of Gaussian process on fibre bundle, we illustrate how to predict colors on the Möbius strip. We begin with $\mathcal{X}$, the Möbius strip defined in . Suppose we are given scattered information about the colors on the Möbius strip as in Figure 6.1. We would like to know what color is the rest of the Möbius strip. Mathematically, this can be stated as a prediction problem. Given a set of observed values $\{(e_i, y_i) \in \mathcal{X} \times \mathbb{R} \mid 1 \leq i \leq n\}$, The output values y_i's can be thought as assigning colors to the input points e_i's. We want to find the values for e^*, where $e^* \in \mathcal{X}$.

Figure 6.1 (top) illustrates the predicted color for $\mathcal{X}$ by the coupled diffusion kernel with $t = 0.0625$ and $\gamma t = 2$. The observed values were in fact generated from the function

$$y = \sin\left(\frac{1}{2}a + \frac{1}{2}\pi\right) \cos\left(\pi r + \frac{1}{2}\pi\right). \tag{6.1}$$

The prediction with Gaussian process indeed returns the true values.

We then apply Gaussian process prediction to data that are from an unknown function. The predicted values, as shown in Figure 6.1 (bottom right), are averages of observed points nearby. This echoes the fact that Gaussian process is the best

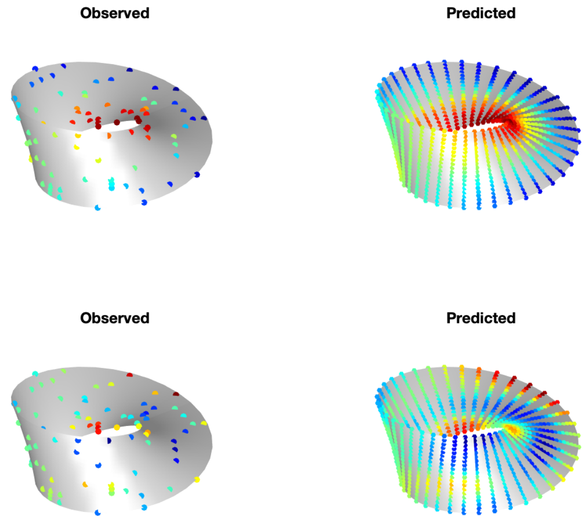


Figure 6.1: Prediction with Gaussian process on a Möbius strip, given observed values generated from a known trig function (top) and an unknown function (bottom).

linear prediction.

How did we know what parameter values to use? Parameter tuning is in general a difficult task in machine learning. In the next chapter, we shall discuss how to train a Gaussian process on fibre bundles.

Chapter 7

Regression on fibre bundle

In previous chapters, we have come to the link between the diffusion kernel and dimension reduction via Gaussian process. We have defined Gaussian process on fibre bundles and suggested diffusion kernel as a canonical choice of covariance function on fibre bundles. In this chapter, we shall discuss how to choose models and tune parameters for Gaussian process on fibre bundles.

7.1 Gaussian process regression

We begin with a revisit to Gaussian process regression. Let E be a fibre bundle as before. Given the observed data set, $\mathcal{D} = \{(e_i, y_i) \in E \times \mathbb{R} \mid 1 \le i \le N\}$, the Gaussian process regression on fibre bundle is modeled by,

$$y_i = f(e_i) + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{7.1}$$

Here, ϵ_i is i.i.d noise distributed according to a normal distribution. For f, we assume a Gaussian process prior with zero mean and some covariance function $k : E \times E \to \mathbb{R}$,

$$f \sim \mathcal{GP}(0, k).$$

The goal of Gaussian process regression is to obtain a posterior of f after observing the dataset $\mathcal{D}$, i.e., $p(f|\mathcal{D})$.

Recall that a Gaussian process is uniquely determined by its mean and covariance function. In our case, the mean is assumed to be zero for illustration purpose. If the covariance function k is parametrized by a set of parameters θ_k, then our parameter space consists of $\theta = \theta_k \cup \{\sigma^2\}$. To find $p(f|\mathcal{D})$ is equivalent to finding the parameter posterior, $p(\theta|\mathcal{D})$. With Bayes's Rule, we obtain

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{\int_{\theta'} p(\theta')p(\mathcal{D}|\theta')d\theta'}$$

The integral above is usually difficult to attain in practice, and we often resort to sampling schemes like MCMC.

To be more precise in the context of fibre bundle, we assume the covariance function k, defined as in Section 7.1, to be of the form

$$k(e, e') = \tau^2 k_{\text{diff}}(e, e'),$$

where $\tau^2 > 0$ and k_{diff} is the horizontal or coupled diffusion kernel defined as in Chapter 6. If k_{diff} is the horizontal diffusion kernel, then total parameter space θ is $\theta = \{t, \tau^2, \sigma^2\}$; if k_{diff} is the coupled diffusion kernel, then $\theta = \{t, \gamma, \tau^2, \sigma^2\}$. Our goal is to learn, after seeing the data points $\mathcal{D}$, the parameter posterior. More formally, we would like to compute the joint distribution of the posterior $p(\theta|\mathcal{D})$.

7.2 Parameter learning

This section deals with how to compute the posterior $p(\theta|\mathcal{D})$ that arises from the regression problem.

The frequentist's approach to determining the parameters in a Gaussian process usually consists of maximizing the log marginal likelihood.

Let $\mathcal{D} = \{(e_i, y_i) \in E \times \mathbb{R} \mid 1 \leq i \leq N\}$ be given as before. Define $\mathbf{y} = [y_1, \ldots, y_N]$ to be the vector of the N observed target values. Let θ denote the hyperparameters as before. We derive the log marginal likelihood $l(\theta)$ by

$$l(\theta) = \log p(\mathbf{y}|\theta) = -\frac{1}{2} \log \det C(\theta) - \frac{1}{2} \mathbf{y}^\intercal C^{-1}(\theta) \mathbf{y} - \frac{N}{2} \log(2\pi),$$

where $C = \tau^2 K + \sigma^2 I$ and K is the covariance matrix obtained by

$$K_{ij} = k_{\text{diff}}(e_i, e_j).$$

The three terms of the marginal likelihood balance between the model complexity indicated by $\log \det C(\theta)$, and the data fit indicated by $\mathbf{y}^\intercal C^{-1}(\theta) \mathbf{y}$.

To set the hyperparameters, we maximize the marginal likelihood by applying standard gradient based methods, such as conjugate gradient or quasi-Newton. We derive the gradients for $\theta_i \in \theta$ as

$$\frac{\partial l}{\partial \theta_i} = \frac{1}{2} \mathbf{y}^\intercal C^{-1} \frac{\partial C}{\partial \theta_i} C^{-1} \mathbf{y} - \frac{1}{2} \text{Tr} \left[C^{-1} \frac{\partial C}{\partial \theta_i} \right]$$

The frequentist's regime will output one optimal choice for each parameter in θ. However, it does not give us the confidence region or the uncertainty score about that optimal value. The Bayesian approach applies an MCMC sampler to compute $p(\theta|\mathbf{y})$. The prior structure is given by,

$$\begin{aligned} \tau^{-2} \sim \text{Gamma}(a_\tau, b_\tau), \quad \sigma^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma), \\ t, \gamma \sim \text{DiscUnif}\left([\phi_1, \ldots, \phi_q] \times [\kappa_1, \ldots, \kappa_p]\right). \end{aligned}$$

We iterate sampling the following,

1. $\sigma^{-2} \mid y, \tau^2, \phi, \kappa$ with a Metropolis random walk on $\log(\sigma^2)$;
2. $\tau^{-2} \mid y, \sigma^2, \phi, \kappa$ with a Metropolis random walk on $\log(\tau^2)$;
3. $\operatorname{pr}(\phi_l, \kappa_m \mid -) = c \det(\tau^2 K^{\phi_l,\kappa_m} + \sigma^2 I_n)^{-1/2} \exp(-y'(\tau^2 K^{\phi_l,\kappa_m} + \sigma^2 I_n)^{-1} y)$.

Here, c is a constant such that the sum of probability on the discrete space is 1. K^{ϕ_l,κ_m} denotes the covariance matrix induced by the diffusion kernel with parameters $t = \phi_l$ and $\gamma = \kappa_m$.

After burn-in, the sampled values for $\sigma^2, \tau^2, \theta_M, \theta_F$ will be distributed, approximately, according to $p(\theta|\mathbf{y})$. The uncertainty score can therefore be computed as the variance of the sampled values and the confidence region can also be acquired conveniently.

The sampler above helps us choose the optimal parameters when a particular choice of covariance function is given. However, there may be many possibilities of covariance functions. We already have two diffusion kernels, not to mention the variations we can have (e.g., by changing the rotation invariant function, adding or subtracting two kernels, adapting to data, etc.)! How do we choose between different covariance functions? More generally, how do we do model selection?

Consider the rather restricted setting with two models M_1, M_2 parametrized by θ_1, θ_2 respectively. To learn which model is more consistent with the observed data, we compute the *Bayes factor* based on the observed data $\mathcal{D}$, which is defined by,

$$B_{12} = \frac{P(\mathcal{D}|M_1)}{P(\mathcal{D}|M_2)} = \frac{\int P(\mathcal{D}|\theta_1)P(\theta_1)d\theta_1}{\int P(\mathcal{D}|\theta_2)P(\theta_2)d\theta_2}.$$

It is obvious that the ratios of integrals add some additional difficulties in computing the integrals. Nevertheless solutions can be obtained by usual MCMC simulations. A full treatment on model selection with Bayes factor can be found in [Rob07].

Chapter 8

Understanding evolutionary process with fibre bundle

An important application of the machinery we have developed so far is to study evolutionary process on anatomical surfaces. Evolutionary biologists study shapes (e.g., teeth and bones) of extinct and extant species to understand the evolutionary process on the phylogenetic tree. With the advent of easily accessible and widely available 3D digital models of anatomical surfaces, devloping methods that directly utilize the surface data to model and understand the tempo and mode of macroevolution is timely. In this chapter, we apply Gaussian process regression on fibre bundles to surfaces that represent primate teeth.

8.1 Evolution of continuous traits

Traditional statistical methods model variations in surfaces by defining traits that quantify certain aspect or the overall geometry of the surface. Examples of such traits include size, surface area, height, etc. More complicated shape characterziers include Relief Index, OPC and DNE which all describe the surface complexity.

The evolution of a continuous biological trait is usually described by three models: Brownian motion (BM), Ornstein-Uhlenbeck (OU) and Early Burst (EB).

The BM model is

$$dZ(t) = \tau dW(t),$$

and it induces the kernel,

$$k_{\text{evol}}(z_i, z_j) = \tau^2 t_{ij}, \tag{8.1}$$

where t_{ij} is the shared path length from the root to the common ancestor of taxa i and j and τ^2 is the evolutionary rate.

The OU model is

$$dZ(t) = -\alpha(Z(t) - \mu) + \tau dW(t),$$

and it induces the kernel

$$k_{\text{evol}}(z_i, z_j) = \frac{\tau^2}{2\alpha} e^{-2\alpha(T - t_{ij})}(1 - e^{-2\alpha_{ij}}), \tag{8.2}$$

were α is the constraint parameter and T is the length of the deepest split on the tree. The parameter τ^2 and t_{ij} are defined as before.

Finally, the EB model is

$$dZ(t) = \gamma(t) dW(t),$$

where $\gamma^2(t) = \tau^2 e^{rt}$ and $r < 0$ is a parameter describing the rate of change. The EB model describes an evolutionary process with slowing rate of change over time. When r is close to 0, the EB model returns the BM model. Under the EB model, the kernel is defined by

$$k_{\text{evol}}(z_i, z_j) = \tau^2 \left(\frac{e^{rt_{ij}} - 1}{r} \right). \tag{8.3}$$

While traits on surface has been very helpful in understanding aspects of the evolutionary process in the clade, if we analyse wrong or poorly justified traits, they can be misleading about which processes meaningfully describe a clade's evolution as a whole. Sometimes, different traits can have different evolutionary patterns in the same clade. For example, [MR10] found that tooth size and carnassial angle variables followed very different evolutionary patterns within Carnivora. Carnassial angle, arguably the more directly functional variable, followed an adaptive radiation model, while m1 size followed a simpler brownian motion model.

It would be better, if possible, to evaluate evolutionary models from the points themselves that consitute the shapes rather than using such proxies. The next section presents the fibre bundle approach for evolutionary process.

8.2 Evolution on fibre bundle

The fibre bundle approach models shape variations on the data point level. We begin with viewing the shape evolution as a Gaussian process on fibre bundles. Within this framework, each point on a shape can be thought as a realization of the Gaussian process. The shape as a whole is an element of the base manifold, where the evolution undertakes. Between any pair of two shapes, we compute a correspondence map that match their structural similarites. The correspondence map is an approximation to the parallel transport on the fibre bundle. Our goal is to retrive, from variation of geometry in the data points and their correspondence maps, information about the evolutionary process on the base manifold. To write this more precisely, let y_i be a point on a shape in the data set,

$$y_i = f(e_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where ϵ is a noise random variable as before. We assume that f, defined on E, has a GP prior

$$f \sim \mathcal{GP}(0, k).$$

Again, we assume the mean function is zero for simplicity. For the covariance function, we can define a kernel that jointly models the evolutionary process and stochasticity on the fibre given $e_i, e_j \in E$. By the trivialization property $e_i \in E$ is equivalent to $(z_i, u_i) \in (M, F_{z_i})$. The covariance function k is given by the following product,

$$k(e_i, e_j) = k((z_i, u_i), (z_j, u_j)) = k_{\text{evol}}(z_i, z_j) \cdot k_{\text{fibre}}(u_i, u_j). \tag{8.4}$$

where k_{evol} can be defined as (8.1 - 8.3) in previous section. We remark here that in practice the explicit form of (z_i, u_i) is usually not given a priori. However, this does not mean we cannot compute (8.4). For $k_{\text{evol}}(z_i, z_j)$, since we know to which species each shape belongs, we can read off from the phylogenetic tree, t_{ij} the shared path length from root to common ancestor between the two species. For $k_{\text{fibre}}(u_i, u_j)$, we can approximate it with the Euclidean distance between y_i, y_j after propagating y_i to the space of y_j with the map $P_{z_j z_i}$,

$$k_{\text{fibre}}(u_i, u_j) = \exp\left(-\frac{d^2(P_{z_j z_i}(y_i), y_j)}{t_F}\right). \tag{8.5}$$

In (8.4) and (8.5), we recoganize the coupled diffusion kernel. The geodeisic distance on the base manifold is now replaced with evolution distance on the phylogenetic tree.

To study evolutionary process now becomes solving a training problem for f. Let

k_{evol} be parametrized with θ_{evol}, and k_{fibre} with θ_{fibre}. The parameters are summarized in $\theta = \{\theta_{\text{evol}}, \theta_{\text{fibre}}, \sigma^2\}$. We define $\mathbf{y} = [y_1, \ldots, y_N]$ to denote the vector of the N observations. Note that $\mathbf{y}$ consists of all points on all shapes in the dataset. As outlined in the previous chapter, we can assess the fit of different evolutionary models to $\mathbf{y}$ by computing the Bayes factor. To estimate the rate or parameters in each of the competing evolution models, we can compute the posterior $p(\theta|\mathbf{y})$. We are particularly interested in parameters on the base manifold, as they encode information about the evolutionary process. Towards this goal, we compute $p(\theta_{\text{evol}}|\mathbf{y})$.

$$p(\theta_{\text{evol}}|\mathcal{D}) = \int p(\theta_{\text{evol}}|\theta^*, \mathcal{D})p(\theta^*|\mathcal{D})d\theta^*,$$

where θ^* is $\theta \backslash \theta_{\text{evol}}$, the complement of θ_{evol} in θ, As before, the customary difficulty is the computation of the integral. To solve this, we first compute the joint posterior $p(\theta|\mathbf{y})$ with the discrete sampler as outlined in Section 7.2. Then we take the distribution of the sampled values of θ_M to obtain an approximation of the marginal posterior $p(\theta_{\text{evol}}|\mathbf{y})$.

In the following section, we shall apply the fibre bundle approach to test the adaptive radiation hypothesis for the lemurs of Madagascar.

8.3 Lemurs of Madagascar

Lemurs of Madagascar are considered a classic example of adaptive radiation [Her17]. Lemurs are the most diverse clade in the strepsirrhini primates, while the sister clade to lemurs in strepsirrhini, the lorises and galogos of Africa and Asia are of fewer species than lemurs. Based on this pattern of speciation, early researchers believed that the strepsirrhini ancestor dispersed to Madagascar and diversified with an "early burst" signal, meaning that the evolution rate was high at the beginning and gradually declined over time. This hypothesis has been only partially tested before. In this study, we will use the Gaussian process regression on fibre bunlde to infer the evolution model of lemurs on Madagasgar.

Our data set consists of 40 second mandibular molars from 40 different strepsirrhini primates. One tooth represents one species. Each molar surface was scanned and discretized into a triangular mesh with approximately 10,000 faces and 5,000 vertices. As a pre-processing step, all surfaces are centered at the origin and scaled to be tightly bounded by a unit cube. Furthermore, all surfaces are aligned to have the same orientation with auto3dgm [BLC$^+$11], a fully automated approach for aligning shapes. For each pair of two shapes, we compute their correspondence map with *Continuous Procrustes Distance* introduced by [BLC$^+$11]. We denote the map to tooth T_j from T_i as P_{ji}.

To test the adaptive radiation hypothesis, we model the evolution with the EB model as described in Section 8.1. The key ingredient is the covariance matrix K in $C = K + \sigma^2 I$, where K is a 40×40 block matrix and each block is a matrix of size approximately 5000×5000. To unload the computational burden, we first downsample each surface to be represented by 100 landmarks with *GP landmarking* [GKBD19], [GKD19]. Let y_{il} denote the l-th landmark on the i-th tooth. The covariance matrix K, now of size 4000×4000, is constructed by defining $[K_{ij}]_{l,m}$, its l, m-th entry of the i, j-th block as

$$[K_{ij}]_{l,m} = \tau^2 \left(\frac{e^{rt_{ij}} - 1}{r} \right) \exp \left(- \frac{||P_{ij}(y_{il}) - y_{jm}||^2}{t_F} \right).$$

The parameter space is $\theta = \{\tau^2, \sigma^2, r, t_F\}$. To compute $p(\theta|\mathbf{y})$, we adopt the following

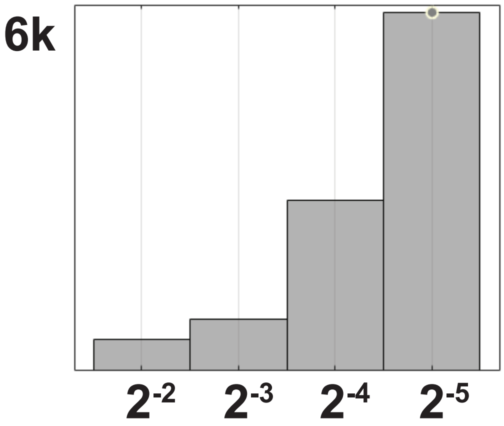


Figure 8.1: Marginal posterior of $-r$.

discrete prior structure,

$$\tau^2 \sim \text{DiscUnif}(\tau_1^2, \ldots, \tau_{n_1}^2),$$
$$\sigma^2 \sim \text{DiscUnif}(\sigma_1^2, \ldots, \sigma_{n_2}^2),$$
$$r, t_F \sim \text{DiscUnif}\left([r_1, \ldots, r_{n_3}] \times [\phi_1, \ldots, \phi_{n_4}]\right).$$

We iterate sampling the following,

1. $\text{pr}(\sigma_m^2 \mid -) = c_1 \det(\tau^2 K + \sigma_m^2 I_n)^{-1/2} \exp(-y'(\tau^2 K + \sigma_m^2 I_n)^{-1} y)$.
2. $\text{pr}(\tau_l^2 \mid -) = c_2 \det(\tau_l^2 K + \sigma^2 I_n)^{-1/2} \exp(-y'(\tau_m^2 K + \sigma^2 I_n)^{-1} y)$.
3. $\text{pr}(r_l, \phi_m \mid -) = c_3 \det(\tau^2 K_{r_l,\phi_m} + \sigma^2 I_n)^{-1/2} \exp(-y'(\tau^2 K_{r_l,\phi_m} + \sigma^2 I_n)^{-1} y)$.

Here, c_i for $i = 1, 2, 3$ are constants such that the sum of probability on each discrete space is 1.

The histogram of sampled values of $-r$ is shown in Figure 8.1. Based on this result, $r = 2^{-5}$ was chosen as the best approximation for the evolution rate parameter in the EB model. Since $r = 2^{-5}$ is close to zero, the result obtained from our method seems to suggest a BM model for lemurs of Madagascar, and therefore indicates the hypothesis of adative radiation is not favored by the data. Although this may seem surprising at first – biology researchers have treated lemurs of Madagascar as a classic example of adaptive radiation – our results are consistent with emerging evidence that lemurs of Madagascar may not exhibit strong signals of early burst. Yet, we shall not completely reject the early burst hypothesis. In general, the power for models of data in high dimensionality is weak. A larger data sets could remedy the difficulties in assessing results for high-dimensional data. Reducing the correlation in the data sets would also help reduce the curse of high-dimensionality.

Chapter 9

Geometry for morphology at a different level: ariaDNE

This chapter looks at morphology at a different level. If before we are looking at collection of shapes that us to infer the macroevolution, we are now zooming in onto each fibire in the bundle of shapes and study features that characterize the geometry of an individual surface.

This work was done in collaboration and published previously in the following publication.

S. Shan, S. Kovalsky, J. Winchester, D. Boyer, and I. Daubechies.

"ariaDNE: A robustly implemented algorithm for Dirichlet energy of the normal."

Methods in Ecology and Evolution 10, no. 4 (2019): 541-552.

9.1 Shape characterizers

Shape characterizers that quantify aspects of the overall geometry of a surface are a promising class of metrics that enables studies on dietary preferences and evolution processes [Eva13]. They are distinguished from "shape descriptors", primarily including geometric morphometric quantifications of shape [AOC13]. Examples of shape characterizers include relief index (RFI) (e.g., [MU03]), orientation patch count (OPC) (e.g., [EWFJ07], [EJ14], [Mel17]) and Dirichlet Normal Energy (DNE) (e.g., [BBL+11], [Win16], [PWM+16]). RFI measures the relative height and sharpness of an object; OPC measures the complexity or rugosity of a surface; DNE measures the bending energy of a surface. When different teeth exhibit substantially different values of a shape characterizer, they also look different and have easily conceivable functional and ecological differences. For instance, a tooth with higher relief often has sharper blades or cusps that pierce food items more effectively than a tooth with lower relief. As another example, DNE differences in mammalian baculum shape potentially correspond to difference in genital form and function ([GBB18]).

Compared to popular shape characterizers like RFI and OPC, DNE has several advantages. Whereas DNE is landmark-free and independent of the surface's initial position, orientation and scale, RFI and OPC rely on the orientation of the tooth relative to an arbitrarily defined occlusal plane. Thus, DNE is less susceptible to observer induced error/noise. In addition, OPC relies on the orientation of the tooth with regard to rotation around the central vertical axis. Furthermore, direct comparisons show that DNE has a stronger dietary signal for teeth than RFI and OPC ([WBSC^{+}14]). This greater success in dietary separation is likely due to its more effective isolation of information on the "sharpness" of surface features. In contrast, RFI only measures the relative cusp and/or crown height which does not describe sharpness; OPC is less sensitive to changes in blade orientation due to its binning protocol ([BEJ10]).

Mathematically, DNE computes a discrete approximation to the *Dirichlet Energy of the normal*, a continuous surface attribute, coming from differential geometry. This quantity is defined as the integral, over the surface, of change in the normal direction, indicating at each point of the surface, how much the surface bends. In practical applications, a continuous surface is represented as a triangular mesh, i.e., a collection of points or nodes and triangles. (We note that the nomenclature is not standardized across all scientific fields; in computer science these would be called *vertices* and *triangular faces*, respectively; see e.g., [BKP^{+}10]). To compute DNE on such a discrete mesh, normal directions must be estimated for each point/triangle. The sum of the change of normal directions over the points/triangles is then used to approximate the Dirichlet energy of the normal for the continuous surface that the mesh represents. However, the DNE algorithm published in *MorphoTester* [Win16] and in the *R* package "molaR" [PWM$^+$16] is sensitive to varying mesh preparation protocols and requires special treatment for boundary triangles, which are triangles that have one side/node that fall on the boundary of the mesh [PSM$^+$16], [SPMK17], raising concerns regarding the comparability and reproducibility when utilizing DNE for morphological research.

Recent attempts to address this issue have developed protocols for standardizing the mesh preparation process [SPMK17]. Unlike previous work, we provide a robustly implemented algorithm for Dirichlet energy of the normal (ariaDNE), that is insensitive to a greater range of mesh preparation protocols. Fig. 9.1 shows DNE and ariaDNE values on an example tooth and various meshes representing the same surface: (from left to right) 2k triangles, 20k triangles, a different mesh representation, 0.001 noise (by adding a random number normally distributed with mean 0 and standard deviation 0.001 to the points), 0.002 noise (similar to 0.001 noise but with standard deviation set to be 0.002) and smoothing. The red surface shading indicates the value of curvature as measured by each approach; it is uniformized across each row by the row's highest local curvature value. To demonstrate this insensitivity empirically, we test the stability of our algorithm on tooth models with differing triangle count, mesh representation (i.e., a different set of points/nodes or triangles representing the same continuous surface), and simulated noise. We also test the effects of smoothing and boundary triangles as in [SPMK17]. We furthermore assess the dietary differentiation power of ariaDNE.

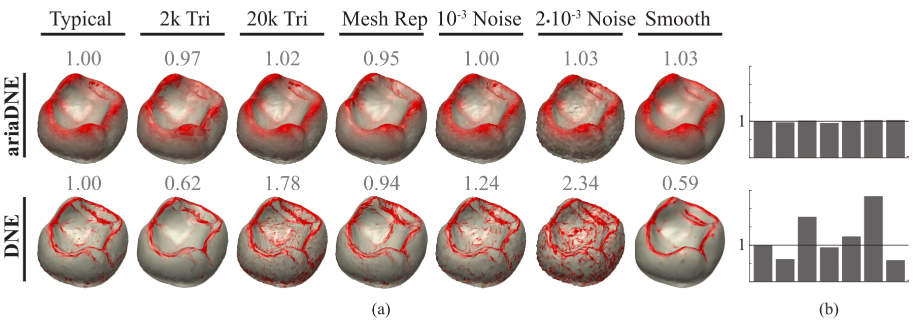


Figure 9.1: Comparing effects of triangle count, mesh representation, noise and smoothing on ariaDNE (top) and DNE (bottom). (a) shows the distribution of curvature as measured by each method overlaid in shades of red on a grey 3D rendering of the surface. Normalized ariaDNE and DNE values (by the values for the typical tooth) are shown above each surface and summarized in the bar plots (b), demonstrating the robustness of ariaDNE versus DNE.

9.2 ariaDNE: a robustly implemented algorithm for DNE

[BBL$^+$11] noted the relevance of the differential geometry concept of Dirichlet energy of the normal for morphology and provided an algorithm called DNE calculating an approximation to this quantity on discrete surface meshes by summing the local energy over all triangles. The local energy on a triangle is defined by the total change in the normals; this provides a local estimate for the curvature of the surface. However, this change in normals is sensitive to how a continuous surface is discretized. That is, a different triangle count, mesh representation, or contamination by noise or small artifacts can all lead to significantly different numerical values.

To address this sensitivity problem, we leverage the observation that the local energy can be also expressed by the curvature at the query point on the surface [Wil65]; another simple method for estimating curvature on discrete surfaces is by Principal Component Analysis (PCA). The procedure is outlined as follows. For

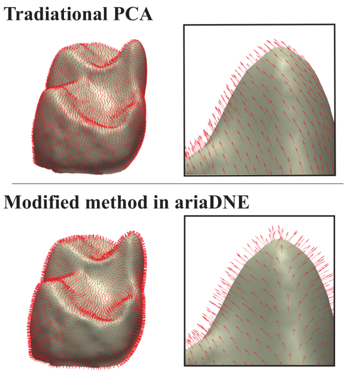


Figure 9.2: Improved normal estimation with our modified method in ariaDNE. Top: traditional PCA method gives skewed normal estimates on a pointed cusp, leading to erroneous curvature approximation. Bottom: our modification gives better normal approximation, and therefore improves curvature approximation.

each query point, find all its neighboring points within a fixed radius; the value of this radius is set as a parameter for the method [$Y^{+}06$]. Then apply PCA to the coordinates of those points; the plane spanned by the first two principal components typically approximates the tangent plane to the surface at the query point, with the third principal component approximating the normal direction. The corresponding smallest principal component score $\sigma = \lambda_0/(\lambda_0 + \lambda_1 + \lambda_2)$ where $\lambda_0 < \lambda_1 < \lambda_2$, indicates the deviation from the fitted plane, i.e. the curvature.

There are two issues with this PCA method: (1) The third principal component does not always approximate the normal direction. Therefore, the smallest principal component score may not accurately reflect curviness as we discussed above, i.e., the deviation of the surface from the tangent plane. Fig. 9.2 (top) shows an erroneous normal approximation for a cusp, where the normals should be perpendicular to the surface but using standard PCA gives skewed estimation. (2) Standard PCA becomes numerically unstable (due to ill-conditioning) when the number of nearby neighbors is low. This implies that when the triangle count is low, there may not be enough points for PCA.

To resolve the first issue, we modify the algorithm to choose at each query point the principal component closest to its normal, and set the curvature at that point to be the score of the chosen principal component. Fig. 9.2 (bottom) illustrates that this modification produces estimates more consistent with surface normals, thereby providing a better local estimate of the tangent plane, and in turn curvature. In practice, normals at a point are obtained by taking a weighted average of normals of adjacent triangles, easily computed on discrete meshes.

To resolve the second issue, we propose a modification to the traditional PCA method. Selecting neighbors within a fixed radius could result, near some point, in a small-sized neighborhood where few or even no points would be selected; instead, we apply a "weighted PCA", with weights decaying according to the distance away from the query point, retaining the rest of procedure. There are many ways to define the weight function. Indeed, using an indicator function that outputs one for points within the chosen radius and zero elsewhere recovers the tradition PCA method. For ariaDNE, we set the weight function to be the widely-used Gaussian kernel $f(x) = e^{-x^2/\epsilon^2}$.

The Gaussian kernel captures local geometric information on the surface. The parameter ϵ indicates the size of local influence. Fig. 9.3 illustrates effects of different ϵ on the weight function: the larger ϵ, the more points on the mesh have significant weight values, resulting in larger principal component scores for those points. In consequence, when ϵ increases, ariaDNE becomes larger. In practice, we suggest using ϵ ranging from 0.04 to 0.1. If ϵ is too small, ariaDNE will be highly sensitive to trivial features that are most likely to be noise (similar to traditional DNE); if ϵ is too large, the approximation will simply become non-local. Choosing an appropriate value of ϵ depends on the application in hand.

In summary, (1) we apply a weighted PCA, localized around each query point by

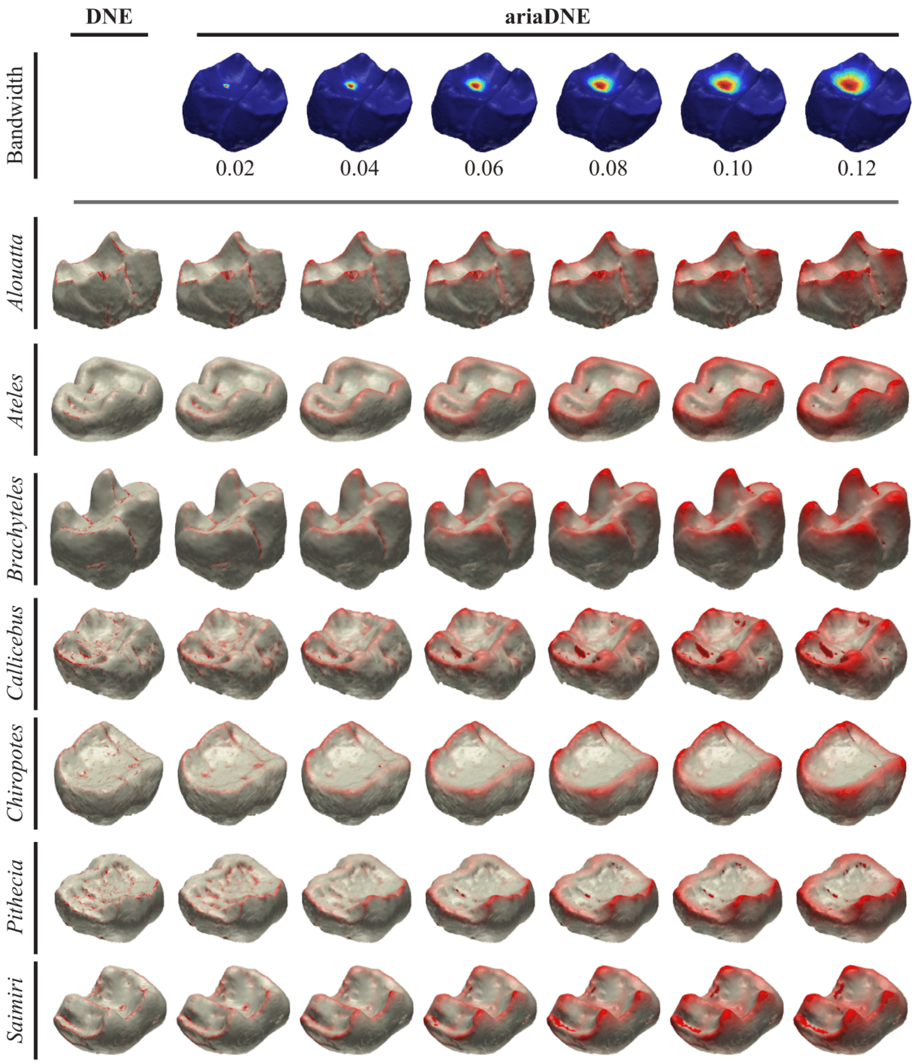


Figure 9.3: Effect of increasing the ϵ parameter (bandwidth) on the weight function (top; red indicates highest weight) and curvature computed by ariaDNE for molar teeth. Choices for ϵ are 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, with surface shading similar to Fig. 9.1. When ϵ is small, both DNE and ariaDNE capture fine-scale features on the tooth. When ϵ is larger, ariaDNE captures larger scale features.

means of the Gaussian kernel function; (2) we find the principal component that is closest to its normal and set the curvature to be its principal score; (3) we integrate this curvature estimate along the surface to obtain ariaDNE. See Appendix A for detailed procedure.

We can furthermore show that the weighted PCA approach indeed capture local curvature information on the surface.

9.3 Sensitivity tests on primate molar data

9.3.1 Study samples

Understanding the correlation between surface geometry and metrics like DNE or ariaDNE helps understand whether these metrics are relevant to questions concerning morphology, ecology and evolution. The meaningfulness and success of a metric have to be measured against relevant samples and the research questions.

Here we use a sample of new world monkey (platyrrhine) second mandibular molars downloaded from *Morphosource* [WBSC$^+$14]. The sample has substantial taxonomic breadth (7 extant platyrrhine primate genera) and depth (10 individuals per genus), consisting of meshes (117,623 - 665,001 points, 234,358 - 1,334,141 triangles) from *Alouatta*, *Ateles*, *Brachyteles*, *Callicebus*, *Chiropotes*, *Pithecia*, and *Saimiri*. Platyrrhine dentitions have been essential for questions about dental variation and dietary preference [AK93], [DUTG04], [LWSCB13], [WBSC$^+$14], [ACGK15], [PSM$^+$16]. Questions have included how dietarily diverse platyrrhines should be considered based on available behavioral data, whether and how dental morphology is reflective of diet differences, and how important tooth wear, individual variation, and scale of geometric features are when considering tooth differences between species.

In the following sections, we tested the stability of ariaDNE by perturbing attributes like triangle count and mesh representation. We also tested the effects of noise, smoothing and boundary triangles on ariaDNE. Furthermore, we assessed its power in differentiating the 7 platyrrhine primate species according to dietary habits.

9.3.2 Sensitivity tests

Triangle count

To evaluate the sensitivity of ariaDNE under varying triangle count, each tooth was downsampled to produce simplified surfaces with 20k, 16k, 12k, 8k, 4k, and 2k triangles. We computed ariaDNE values ($\epsilon = 0.04, 0.06, 0.08, 0.1$) using the MATLAB function "ariaDNE" provided in Section 9.6.2. For comparison, we also computed DNE values using the function "DNE" (Section 9.6.2), a MATLAB port of the R function "DNE" from "molaR". Default parameters were used for DNE, with outlier percentile at 0.1 and boundary triangles excluded.

Mesh representation

A continuous surface can be represented by different discrete meshes; even with the same triangle count, they can differ by altering the position of points and their adjacency relations (i.e., triangles). We would like ariaDNE to be roughly the same for all meshes that represent the same continuous surface. To evaluate the sensitivity of ariaDNE under varying mesh representation, we tested on a surface generated by a mathematical function as well as real tooth samples. First, we tested it on the surface S defined by $z = 0.3\sin(2x)\sin(2y)$ where $0 \leq x \leq 1$ and $0 \leq y \leq 1$ (Fig. 9.4). To generate a mesh, we randomly picked 2000 sets of (x, y) coordinates uniformly distributed on $0 \leq x \leq 1$ and $0 \leq y \leq 1$ and calculated their accompanying z-values using the equation above. Each set of (x, y, z) coordinates represented a node/point in the mesh, and the triangles are obtained by applying Delaunay Triangulation to these points. We generated 100 meshes by repeating these steps and computed their DNE and ariaDNE values as in 9.3.2. We remark here that meshes generated by this procedure do not necessarily have evenly distributed points; some areas of the mesh can have smaller triangles than others.

Real tooth samples are already given as meshes; we generated new mesh representations for each tooth sample by computing pairwise surface correspondences. Specifically, points and triangles from one surface were taken to the other surface in the samples by correspondence maps computed using the methods in [BLC^{+}11], between all pairs of surfaces in the sample. These correspondences resulted in 70 different mesh representations for each tooth in the sample. We computed their DNE and ariaDNE ($\epsilon = 0.04, 0.06, 0.08, 0.1$) as in 9.3.2.

Simulated noise

We tested the stability of ariaDNE when simulated noise was added to the surface defined as in 9.3.2 and real tooth samples. First, given a mesh representing the same surface S as in 9.3.2, a noisy mesh was obtained by adding a random variable uniformly distributed on $[-0.001, 0.001]$ to the x, y, z coordinates of each node/point on the mesh. We then generated 100 noisy versions of the given mesh by repeating the previous steps.

For real tooth data, we generated a noisy mesh by adding a random variable uniformly distributed on $[-0.003, 0.003]$ to the x, y, z coordinates for each node/point in the mesh. The noise level was chosen arbitrarily; we added more noise to the tooth samples to increase diversity of the test cases. We obtained 100 noisy meshes per tooth, and computed their DNE and ariaDNE ($\epsilon = 0.04, 0.06, 0.08, 0.1$) values as in

9.3.2.

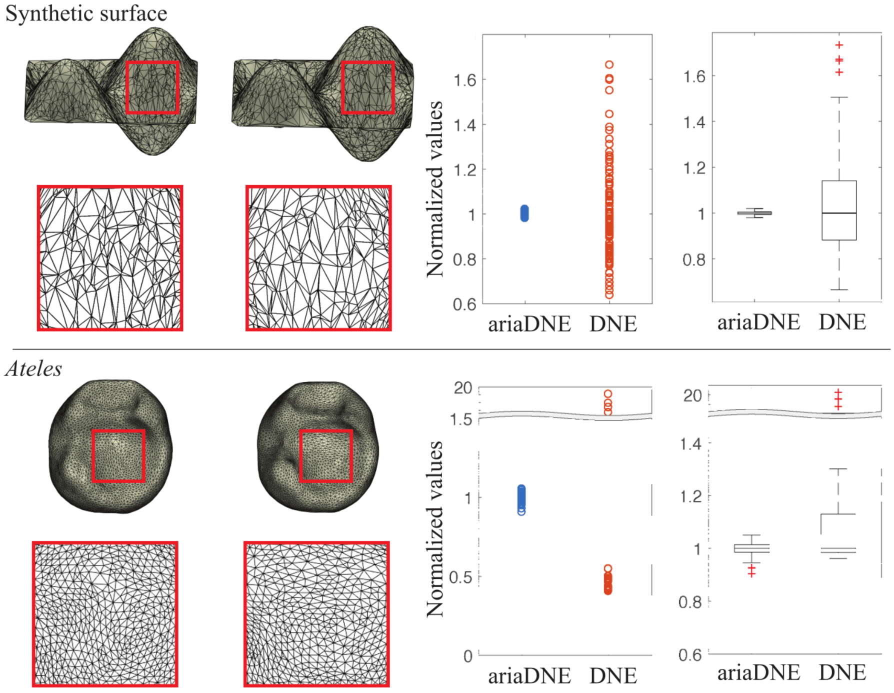


Figure 9.4: Effect of varying mesh representation on ariaDNE and DNE values computed for a synthetic surface (top) and a tooth from *Ateles* (bottom). Left panel: examples of different mesh representations. Right panel: scatter plots and box plots of ariaDNE ($\epsilon = 0.08$) and DNE values computed for N meshes representing the synthetic surface (top, $N = 100$) and the tooth surface (bottom, $N = 70$).

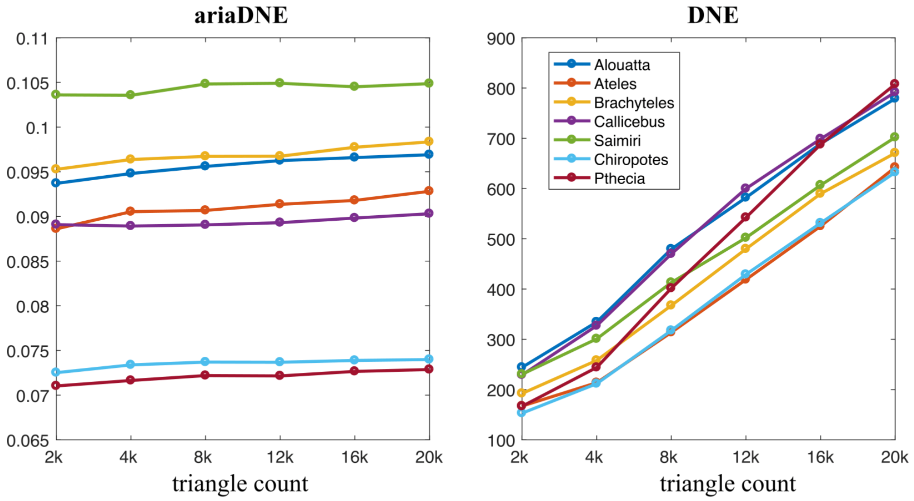


Figure 9.5: Effect of increasing triangle count on ariaDNE (left) and DNE (right) values computed for 7 teeth from *Alouatta*, *Ateles*, *Brachyteles*, *Callicebus*, *Saimiri*, *Chiropotes*, *Pithecia*. The ariaDNE values for each tooth remain relatively unchanged, compared to the DNE values, under varying triangle count.

Smoothing

Smoothing is commonly used to eliminate noise produced during mesh preparation. [SPMK17] tested the effects of various smoothing operators and smoothing amounts on DNE with surface meshes of hemispheres and primate molars. They suggested that aggressive smoothing procedures like Laplacian smoothing and implicit fairing should be avoided. To evaluate the performance of ariaDNE under different smoothing algorithms, we randomly picked 7 tooth models from our sample (one from each taxa) and generated their smooth surfaces by applying 100 iterations of the *Avizo* smoothing module, 3 iterations of the *Meshlab* function *HC Laplacian Smoothing*, or 3 iterations of the *implicit fairing method* using *MorphoTester*. Then we computed their DNE and ariaDNE ($\epsilon = 0.08$) as in 9.3.2.

We further evaluated the effects of varying amounts of *Avizo* smoothing on ariaDNE; we iteratively applied the Avizo smoothing module to a single molar tooth from *Ateles*. The smoothing function was performed in intervals of 20 on the raw surface mesh, evenly spaced from 20 to 200 to generate 10 new surface meshes. Default value for lambda was kept (lambda = 0.6). We computed their DNE and ariaDNE ($\epsilon = 0.08$) as in 9.3.2.

Boundary triangles

Triangles with one side/node that are on the boundary of the mesh have a large impact on DNE, calling for special treatment [SPMK17]. We assess how such boundary triangles affect ariaDNE on two molar teeth, one of *Ateles* where crown side walls are relatively bulged outwardly, and one from *Brachyteles* where crown side walls are relatively unbulged (Fig. 9.6). For each tooth, we found its boundary triangles and computed their local energy using ariaDNE and DNE ("BoundaryDiscard" = "none", i.e., no boundary triangles will be removed).

9.3.3 Tests on species differentiation

Previous studies have revealed systematic variation among species with different dietary habits in the values of DNE and other topographic metrics, such as RFI. To test the differentiation power for species and their dietary preferences, we compared RFI, DNE and ariaDNE on the 70 mandibular second molars in our platyrrhine sample. Under the diet-classification scheme from [WBSC$^+$14], *Alouatta* and *Brachyteles* are folivorous, *Ateles* and *Callicebus* are frugivorous, *Chiropotes* and *Pithecia* are hard-object feeding, and *Saimiri* is insectivorous. For each tooth, we computed RFI, DNE and ariaDNE ($\epsilon = 0.02, 0.04, 0.06, 0.08, 0.1, 0.12$) as in 9.3.2. We then used ANOVA and multiple comparison tests to assess the differentiation power of these different metrics for dietary preferences.

9.4 Results

9.4.1 Sensitivity tests

In numerical analysis, an algorithm is stable if perturbing inputs do not significantly affect outputs. To enable comparison, the change in the outputs can be quantified by *coefficient of variation*, which is the result of dividing the standard deviation by the mean. We computed coefficients of variation of DNE and ariaDNE values of the perturbed meshes in each collection per tooth model. For each tooth and each perturbed collection, the coefficient of variation of ariaDNE is less than that of DNE, meaning ariaDNE is relatively more stable than DNE under varying triangle count, mesh representation and noise. Table 9.1 summarizes results, indicating means of coefficients of variation from Supplementary Tables 1-4. Fig. 9.5 illustrates effects of increasing triangle count on ariaDNE ($\epsilon = 0.10$) and DNE values computed for 7 arbitrarily chosen teeth (one per genus). The ariaDNE values for each tooth (maximum percent change: 3.42 %) remain relatively unchanged compared to DNE (maximum percent change: 384%). Fig. 9.4 illustrates ariaDNE is relatively more stable when the mesh representation is changed. In the scatter plots, ariaDNE and DNE values are normalized to have a mean one in each case; in the box plots, the values are normalized to have a median one in each case. Similar result holds for adding noise to the surface.

Table 9.2 shows percent change of ariaDNE and DNE values subject to different smoothing algorithms. After 100 iterations of Avizo smoothing, ariaDNE increased 2% of its original value whereas DNE dropped to 46%. After 3 iterations of HC Laplacian smoothing and implicit fairing, ariaDNE dropped to approximately 90% of the original value whereas DNE dropped to approximately 40%. The larger drop in values using Laplacian smoothing and implicit fairing is consistent with the discussion

Table 9.1: Robustness of ariaDNE under various mesh attributes perturbation. For each tooth in the 70 platyrrhine sample, we generated three collections of perturbed meshes by varying triangle count, mesh representation or adding simulated noise. We computed the coefficient of variation of their DNE and ariaDNE values in each collection for each tooth. The numbers in the table are obtained by taking the mean across all 70 tooth samples.

Method		Triangle Count	Remeshing	Noise
ariaDNE	$\epsilon = 0.04$	0.0213	0.0824	0.0055
	$\epsilon = 0.06$	0.0114	0.0429	0.0044
	$\epsilon = 0.08$	0.0117	0.0304	0.0039
	$\epsilon = 0.10$	0.0117	0.0293	0.0038
DNE		0.420	2.3075	0.0169

Table 9.2: Effect of different smoothing algorithms on DNE and ariaDNE ($\epsilon = 0.08$). The numbers in the table are DNE and ariaDNE values divided by values of raw surfaces indicating the percent change. The table also contains coefficients of variation (COV) of DNE and ariaDNE computed on the three smooth surfaces in each taxa. The table demonstrates: (1) The effect of smoothing is limited on ariaDNE versus DNE; (2) ariaDNE is relatively more stable under varying smoothing algorithms.

	DNE					ariaDNE				
	Raw	Avizo	Laplacian	Fairing	COV	Raw	Avizo	Laplacian	Fairing	COV
Alouatta	1	0.38	0.35	0.48	0.120	1	1.00	0.90	0.90	0.049
Ateles	1	0.57	0.46	0.57	0.144	1	1.05	0.93	0.93	0.056
Brachyteles	1	0.54	0.45	0.57	0.127	1	1.06	0.96	0.95	0.053
Callicebus	1	0.46	0.39	0.52	0.122	1	0.99	0.90	0.91	0.074
Chiropotes	1	0.51	0.37	0.48	0.112	1	1.02	0.95	0.94	0.06
Pithecia	1	0.43	0.29	0.41	0.165	1	1.02	0.94	0.93	0.058
Saimiri	1	0.33	0.44	0.56	0.169	1	1.00	0.90	0.92	0.056

by [SPMK17]. However, for all smoothing algorithms, the variation in ariaDNE is significantly lower than for DNE. This suggests that ariaDNE is relatively stable under varying smoothing algorithms. For Avizo smoothing, the degree of overall change in ariaDNE from unsmoothed surfaces to smoothed surfaces is much less than the overall change in DNE. This suggests that ariaDNE is relatively more stable under varying Avizo smoothing iterations.

Fig. 9.6 shows that the local energy of the boundary triangles computed with ariaDNE are among the smallest, whereas those computed with DNE have a few larger

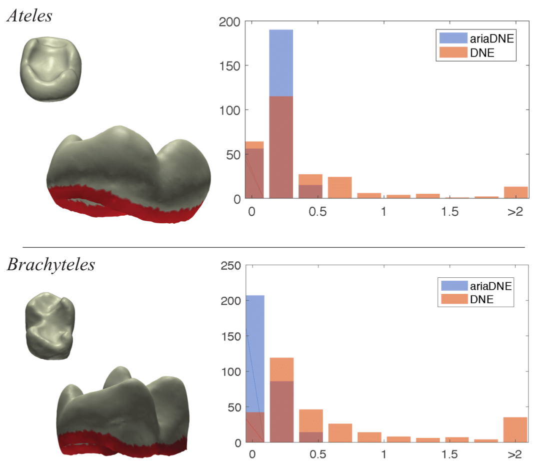


Figure 9.6: The boundary triangles have less impact on ariaDNE than DNE. The left panel shows an *Ateles* molar (top), with a curvy side wall and a *Brachyteles* molar (bottom), with a straight side wall. The right panel shows histograms of local energy values of the boundary triangles, computed by ariaDNE and DNE. To enable comparison, the values are normalized by the mean of those of all triangles.

ones, which affect the DNE value for the whole surface. This histogram suggests that the effects of boundary triangles on ariaDNE are limited, and therefore no special treatment for them is needed. This represents another improvement for ariaDNE compared to DNE.

9.4.2 Species differentiation power

For each shape characterizer (RFI, DNE and ariaDNE), ANOVA rejects the hypothesis with $P < 0.05$ that all dietary groups have the same mean, which indicates that some dietary differentiation was detected. To further determine which group means are different, we used multiple comparison tests and the results are summarized in Table 9.3. RFI separated folivore from frugivore and hard-object feeding; DNE in addition separated hard-object feeding from insectivore. As ϵ increases, ariaDNE further separated frugivore from hard-object feeding and insectivore. No metrics

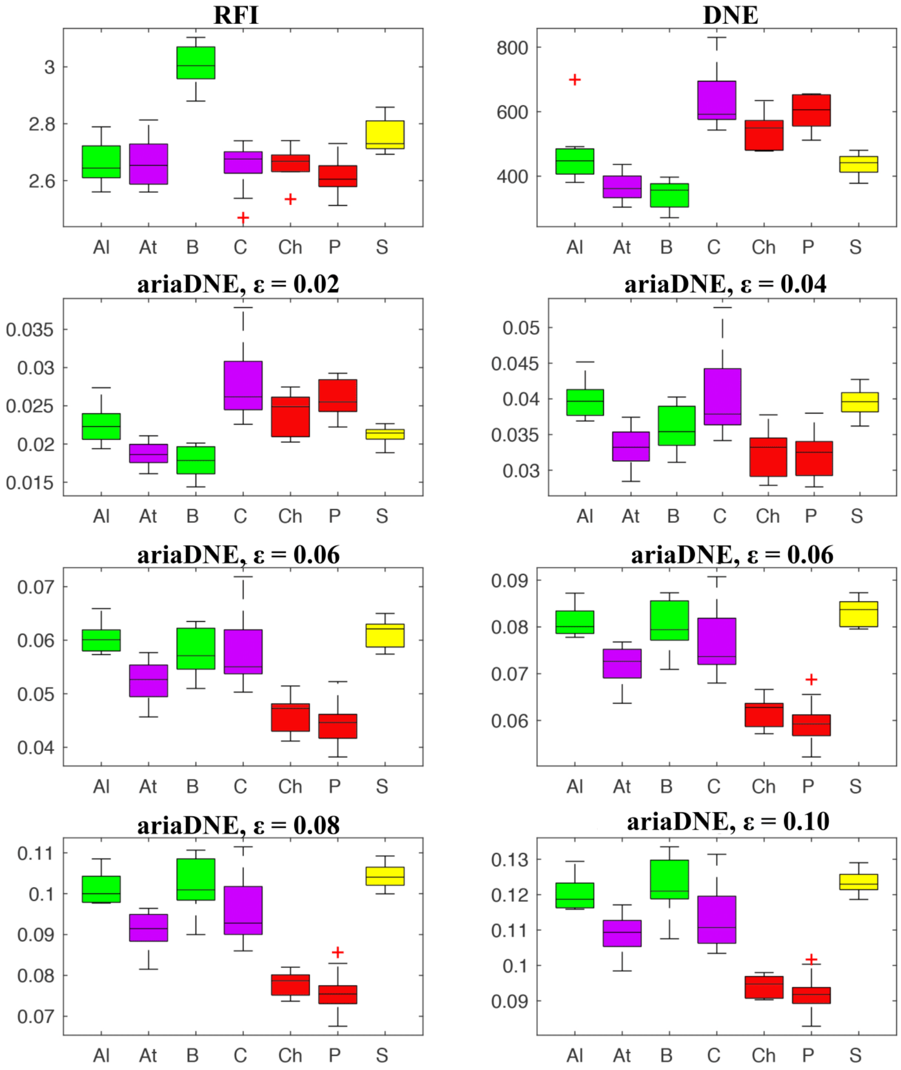


Figure 9.7: Box plots of RFI, DNE, and ariaDNE with various values of ϵ for *Alouatta*(Al), *Ateles*(At), *Brachyteles*(B), *Callicebus*(C), *Chiropotes*(Ch), *Pithecia*(P), *Saimiri*(S). Edges of the box indicate the 25th and 75th percentiles, and the outliers are plotted individually using the '+' symbol. Color indicates dietary preference: green represents folivore, purple represents frugivore, red represents hard-object feeding and yellow represents insectivore.

Table 9.3: Multiple comparison tests on RFI, DNE and ariaDNE ($\epsilon = 0.02, 0.04, 0.06, 0.08, 0.10, 0.12$) values of folivore (Fo), frugivore (Fr), hard-object feeding (H) and insectivore (I). The numbers in the table are p values for the pairwise hypothesis test that the corresponding mean difference is not equal to 0. For $\epsilon = 0.08, 0.10, 0.12$, ariaDNE differentiated folivore, frugivore and hard-object feeding. None of the metrics differentiated insectivore from folivore.

	RFI	DNE	ariaDNE					
			$\epsilon = 0.02$	0.04	0.06	0.08	0.10	0.12
Fo-Fr	0.0001	0.0224	0.0512	0.8362	0.0207	0.0002	0.0000	0.0000
Fo-H	0.0000	0.0000	0.0010	0.0003	0.0000	0.0000	0.0000	0.0000
Fo-I	0.2414	0.8463	0.8986	0.7564	0.6544	0.5584	0.5893	0.7376
Fr-H	0.9372	0.2295	0.5425	0.0049	0.0000	0.0000	0.0000	0.0000
Fr-I	0.1888	0.3910	0.4738	0.3461	0.0034	0.0000	0.0000	0.0000
H-I	0.0689	0.0125	0.0627	0.0002	0.0000	0.0000	0.0000	0.0000

separated folivore and insectivore. However, similarity in their ariaDNE values are not surprising. Insect and leaf tissues tend to be high in structural carbohydrates, which sharpened dental blades are capable of shearing, and therefore high ariaDNE values. What's more important here is the separation from teeth that have low cusps and wide basins, as these are used for crushing motions to efficiently break down soft (i.e., fruit) and hard objects. For $\epsilon = 0.08$, 0.1 and 0.12, the box plots of ariaDNE (Fig. 9.7) converge on a pattern in which folivorous *Alouatta*, *Brachyteles* and insectivorous *Saimiri* have higher values, reflecting sharper cusps, whereas frugivorous *Ateles* and *Callicebus* have lower values and hard-object feeding *Chiropotes* and *Pithecia* have the lowest values, reflecting low unsharp cusps. The separation was not as clear for RFI and DNE.

For $\epsilon = 0.02$, ariaDNE shows a pattern similar to DNE. This suggests that when ϵ is small, both methods capture fine and/or local features on tooth models, and as ϵ becomes larger, ariaDNE starts capturing larger scale features, ignoring smaller scale features. Fig. 9.3 demonstrates the feature scale of DNE and ariaDNE with various ϵ values. The pattern is particularly interesting in *Callicebus*, *Chiropotes* and *Pithecia*, which evince less pointed cusps, but which exhibit more fine details on the basin (such as enamel crenulations for the pitheciines). In these teeth, ariaDNE values are high when ϵ is small, but drop with larger ϵ. The pattern is more pronounced in the teeth of *Pithecia* because their high energy features - the enamel crenulations - are even smaller than those of *Callicebus* and so are erased more completely by using a high ϵ value.

So is it good or bad to erase small-scale features? We do not believe there is an objective answer or a universal optimum for ϵ. [BS17] emphasized the importance of small scale features in their analyses of dental topography of extant apes, which also exhibit crenulated enamel similar to Pitheciines. Additionally, erasing small scale features makes the mandibular second molars of *Pithecia* more similar to those of Aye Ayes (*Daubentonia*). Previous studies have argued that the two species are analogous from an ecological point of view [WBSC⁺14]. On the other hand, small-scale features could reflect an important functional ability of *Pithecia* not available to *Daubentonia* [LWSCB13]. Considering other research questions, these small scale features align *Pithecia* with *Callicebus*, which may be evidence of a close phylogenetic relationship between them - one that was debated prior to availability of genetic data, based on a dearth of obvious unique anatomical similarities.

9.5 Discussion

9.5.1 Bandwidth and multi-scale quantifications

Even with a less sensitive implementation, ariaDNE still requires choices on the parameter ϵ. In section 9.2, we discussed the origin and interpretation of ϵ. We showed how ϵ affects values of ariaDNE and the power to differentiate primates with differing dietary preferences in section 9.4.2. To summarize: (1) for a given ϵ, values of ariaDNE remained relatively unchanged compared to DNE, when the input mesh is perturbed (Fig. 9.1). This suggests that ϵ is independent of mesh attributes like triangle count, mesh representation, noise level, smoothness, etc. (2) The parameter ϵ indicates the size of local influence: the larger ϵ, the more points on the mesh are considered important to quantify the local energy of the query point, and therefore larger ariaDNE values. This means ϵ determines the scale of features to be included in geometric quantification. Small ϵ will make surfaces with finer features have higher ariaDNE values, and large ϵ will make surfaces with large scale features have higher ariaDNE values.

Parameter tuning was often achieved through optimization based on a priori goals, yet a single choice of parameter may not satisfy all goals. For example, the parameter that maximizes the differentiation between species in different diet groups may be different from that which minimizes the effect of wear or optimizes the differentiation between species irrespective of diet. The requirement of choosing a uniform scale applies to quantitative methods generally, and perhaps this is their biggest weakness compared to qualitative analyses of more traditional comparative morphology, where multiple scales of perception were naturally integrated into prose describing observed similarities and differences. However, the freedom to check patterns under different parameters also presents potential for more informative comparisons, as seen in (Fig. 9.7). Future work should aim to characterize samples using values computed across a range of ϵ values.

9.5.2 Wider applicability of ariaDNE

Many other applications of ariaDNE beyond functional questions of teeth are possible (Fig. 9.8). For instance, in bivalves, burrowing benthic forms should benefit from shells with greater rugosity (higher ariaDNE) to help them stay embedded in the sea floor, whereas more planktonic forms should benefit from smoother, more hydrodynamic shells (lower ariaDNE). It might also be helpful to use ariaDNE on distal phalanges (bones supporting the nail/claw) as claws suited for climbing are narrower and sharper (higher ariaDNE) while those suited for burrowing (or grasping) will be broader and blunter (lower ariaDNE). In addition to studying shape complexity across species, ariaDNE might also be used for complexity in a shape over developmental time. [SMSC17], for example, used DNE to study embryo shape development and compared it with disparity in gene expression.

Comparing the distribution of ariaDNE values over surfaces will likely provide even more insight into ecologically meaningful shape variation. For example, two surfaces with the same total ariaDNE may have very different distributions: one may have greater spatial variance in ariaDNE, with high ariaDNE features more clustered in one case than another. In all, ariaDNE opens doors to defining other interesting shapes metrics that could potentially assists our understanding in morphology, evolution and ecology.

9.5.3 ariaDNE for previously published DNE analysis

The insensitivity of ariaDNE under varying mesh preparation protocols makes it more widely usable than DNE for comparing and combining results from studies with varying samples or mesh preparation protocols. The computed ariaDNE values for previously published DNE studies ([BBL+11], [WBSC+14], [PBS16], [PWM+16] [PSM+18], [BS17], [BDK18], [LTSP+18]) are included in the supplementary materials and available to download as csv files from `https://sshanshans.github.io/articles/ariadne.html`. We will continue to update our website as we obtain access to more data samples.

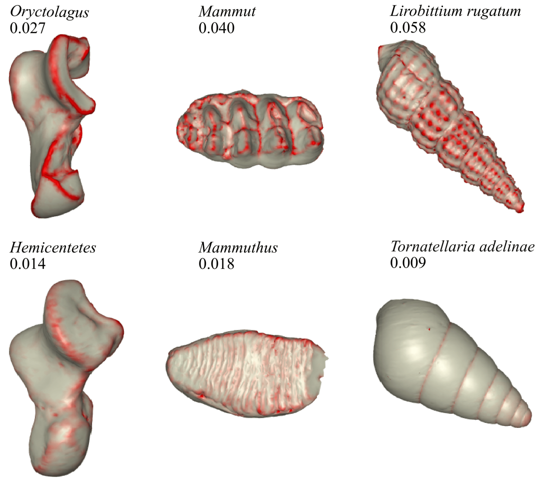


Figure 9.8: ariaDNE values for surfaces representing astragulus (ankle bone) of *Oryctolagus* (saltatorial) and *Hemicentetes* (ambulatory), molars of *Mammut* (folivorous) and *Mammuthus* (grazing) and shells of *Lirobittium rugatum* and *Tornatellaria adelinae*. Surface shading indicates curvature computed by our algorithm; ariaDNE values are above each surface.

9.5.4 Conclusion

We provided a robust algorithm for computing the Dirichlet energy of the normal by utilizing weighted PCA. Compared to DNE, ariaDNE is stable under a greater range of mesh preparation protocols. Specifically, analyses indicated that the effects of differing triangle count, mesh representation, noise, smoothing, and boundary triangles are much more limited on ariaDNE than DNE. Furthermore, ariaDNE retains the potential of DNE for biological studies, illustrated by it effectively differentiating platyrrhine primate species according to dietary preferences. While increasing the ϵ parameter of the method can erase small scale features and significantly affect how ariaDNE characterizes structures with small scale features compared to those with larger features (as it did with *Chiropotes* and *Pithecia* primates in our sample), we think this property can be leveraged to provide more informative comparisons. Future work should aim to characterize samples using values computed across a range of

ϵ values. In this type of analysis, parameters could be optimized according to model selection criteria. Finally, as with other topographic metrics, ariaDNE is likely most informative when deployed in combination with other shape metrics to achieve the goal of more accurately inferring morphological shape attributes.

9.6 Data Accessibility

9.6.1 Sample locations

The platyrrhine sample was published by [WBSC+14], and is available on *Morphosource*, a project-based data archive for 3D morphological data.

9.6.2 Matlab scripts

Matlab scripts are available from the GitHub repository: `https://github.com/sshanshans/ariaDNE_code` and are archived with Zenodo DOI: `https://doi.org/10.5281/zenodo.1465949`.

Chapter 10

Conclusions

The final chapter consists of a recap of the main results in this thesis and further generalizations, connections and applications of the probabilistic models on fibre bundles.

10.1 Diffusion kernel as the primary object

We have looked at shapes from two different levels.

On each individual shape, our main concern is features. The mathematical advancement we have made here is a robust algorithm for computing curvature, ariaDNE. Requiring no strict mesh preparation protocols, biologists can use ariaDNE on a diverse data and use the computed features to study dietary preferences. Through diet, we understand better about the enviorment that the animals live in.

On a macroscopic level, we think of the collection of shapes form one geometric object. Our focus is dimension reduction. We came to realize that diffusion kernel is indeed the primary object in dimension reduction. The shift of view has brought us several results: (1) in Chapter 3, we have used the semi-group properties of the diffusion kernel for parameter tuning in diffusion maps, which can often be difficult in practice; (2) in Chapter 4, we have proposed a projected kernel method that builds a robust low-dimensional representation of data under noise; (3) Chapter 6 shows that the diffusion kernel is a natural choice for the covariance function in Gaussian processes, and this leads us to define regression on fibre bundles in Chapter 7. Finally, in Chapter 8, we then use Gaussian processes on fibre bundles to study the evolution of lemurs on Madagascar.

10.2 Visions for the future

As my advisor playfully observed, the phylogenetic tree looks a lot like a cauliflower, plentiful in the crown and hierarchical in the stem. We have so far only looked at primates, more precisely molars of lemurs and new world monkeys. This perhaps constitutes only a floret of all the different life forms on earth. It is then natural to ask why not look at teeth from all animals? The marvelous yet unfortunate[1] fact is that not all teeth are alike. Take molars of crab-eater seal and mammoth for example. How do we compare two shapes that look so different?

There are several obstructions involved. First, most registration algorithms optimize the proximity between two shapes through an iterative search of the best possible rotation and the best possible correspondence. When there is little likeliness between the two given shapes, the algorithm may not converge. Second, even when the convergence happens, the optimal correspondence map might not indicate homologous structures at all. In biology, homology refers to a shared ancestry between two different taxa. As the result of descent from a common ancestor, homologous structures are similar parts on the shapes, and variation in the homologous structures suggests adaptation to ecological changes. It is therefore crucial to identify the homologous structures in order to understand how shapes evolve in adaptation to environment over time. Third, the "small hops" trick fails when we need to leap. As discussed in Chapter 4, the "small hops" trick involves composing together a series

---

[1] for our task

of maps between similar shapes. However, when there are not enough samples – imagine leaping across the space to go between the target shapes – we cannot get an informative correspondence map.

Schematically, the problem we now face is the following. Given a tree $\mathcal{T}$ as in Figure. Species A_i, $i = 1, 2, 3, 4$ sharing the same common ancestor form a group A and species B_i, $i = 1, 2, 3, 4$ sharing the same ancestor form a group B. Within each group, there is a global registration among the shapes, i.e., the space of A or B is in fact a product space with flat connection. Between the two groups, no known methods can compute the correspondence maps that indicate homologous structures. How can we find biologically meaningful mappings between A and B?

Between widely separated groups, the difficulty in identifying homologous structures was due to their large deviation from the common ancestor. In the long time of independent evolution, old features faded away while new features emerged. We believe these changes should be gradual enough so that we can trace them, remove lately developed features and recover the homologous structures on the phylogenetic tree. Towards this, we shall build a hierarchical representation for the shapes, with increasing level of simplification. On the simplified shapes, correspondence maps are enabled.

To state it more precisely, we return to the tree $\mathcal{T}$. We shall first find a simplified shape A_i° for A_i, $i = 1, 2, 3, 4$. (The procedure is exactly the same for B_i°.) Since the connection on A is flat, we can obtain a shape/fibre template via horizontal diffusion maps. For each point on the fibre template, we compute its variance in A. In other words, we register and align A_i, $i = 1, 2, 3, 4$ to the template shape $\mathcal{F}$, and denote the processed shapes as A_i^*. For each point $l \in \mathcal{F}$, we compute the variance of $A_i^*(l)$, the l-th point in A_i^*. To obtain A_i° we agglomerate points that have a larger variance. We adjust the agglomeration until A° and B° together give a product space, which we can check by the semi-group test! To compute the correspondence maps between A and B, we can now take composition of the maps along the path $A \to A^\circ \to B^\circ \to B$.

Once we have A° and B°, we can move down another level on the tree of life. The simplification-and-register steps would allow us to compare even more distinctly different shapes that we cannot do before. We hope this first step on this simplified model $\mathcal{T}$ can eventually provide us tools to expand the scope of comparative evolutionary studies and add to our study the breadth and diversity of the tree of life.

Bibliography

[ACGK15] Kari L Allen, Siobhán B Cooke, Lauren A Gonzales, and Richard F Kay. Dietary inference from upper and lower molar morphology in platyrrhine primates. *PloS one*, 10(3):e0118732, 2015.

[AK93] Mark RL Anthony and Richard F Kay. Tooth form and diet in ateline and alouattine primates: reflections on the comparative method. *American Journal of Science*, 293(A):356, 1993.

[All14] Kari Leigh Allen. *Endocranial volume and shape variation in early anthropoid evolution*. PhD thesis, Duke University, 2014.

[AOC13] Dean C Adams and Erik Otárola-Castillo. geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, 4(4):393–399, 2013.

[BBL$^+$11] Jonathan M Bunn, Doug M Boyer, Yaron Lipman, Elizabeth M St Clair, Jukka Jernvall, and Ingrid Daubechies. Comparing dirichlet normal surface energy of tooth crowns, a new technique of molar shape quantification for dietary inference, with previous methods in isolation and in combination. *American Journal of Physical Anthropology*, 145(2):247–261, 2011.

[BDK18] Michael A Berthaume, Lucas K Delezene, and Kornelius Kupczik. Dental topography and the diet of homo naledi. *Journal of human evolution*, 118:14–26, 2018.

[BEJ10] Doug M Boyer, Alistair R Evans, and Jukka Jernvall. Evidence of dietary differentiation among late paleocene–early eocene plesiadapids (mammalia, primates). *American Journal of Physical Anthropology*, 142(2):194–210, 2010.

[BGKM16] Doug M Boyer, Gregg F Gunnell, Seth Kaufman, and Timothy M McGeary. Morphosource: Archiving and sharing 3-d digital specimen data. *The Paleontological Society Papers*, 22:157–181, 2016.

[BKP$^+$10] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. *Polygon mesh processing*. AK Peters/CRC Press, 2010.

[BLC$^+$11] Doug M Boyer, Yaron Lipman, Elizabeth St Clair, Jesus Puente, Biren A Patel, Thomas Funkhouser, Jukka Jernvall, and Ingrid Daubechies. Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proceedings of the National Academy of Sciences*, 108(45):18221–18226, 2011.

[BN08] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

[Boy08] Doug M Boyer. Relief index of second mandibular molars is a correlate of diet among prosimian primates and other euarchontan mammals. *Journal of Human Evolution*, 55(6):1118–1137, 2008.

[BS17] Michael A Berthaume and Kes Schroer. Extant ape dental topography and its implications for reconstructing the emergence of early homo. *Journal of human evolution*, 112:15–29, 2017.

[BU09] Jonathan M Bunn and Peter S Ungar. Dental topography and diets of four old world monkey species. *American Journal of Primatology*, 71(6):466–477, 2009.

[BWGP15] Doug M Boyer, Julia M Winchester, Chris Glynn, and Jesus Puente. Detailed anatomical orientations for certain types of morphometric measurements can be determined automatically with geometric algorithms. *The Anatomical Record*, 298(11):1816–1823, 2015.

[Can70] Peter B Canham. The minimum energy of bending as a possible explanation of the biconcave shape of the human red blood cell. *Journal of theoretical biology*, 26(1):61–81, 1970.

[CBC$^+$01] Jonathan C Carr, Richard K Beatson, Jon B Cherrie, Tim J Mitchell, W Richard Fright, Bruce C McCallum, and Tim R Evans. Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 67–76. ACM, 2001.

[CG97] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[CLT$^+$16] Lynn E Copes, Lynn M Lucas, James O Thostenson, Hopi E Hoekstra, and Doug M Boyer. A collection of non-human primate computed tomography scans housed in morphosource, a repository for 3d data. *Scientific data*, 3:160001, 2016.

[CPFA17] JL Cantalapiedra, JL Prado, M Hernández Fernández, and MT Alberdi. Decoupled ecomorphological evolution and diversification in neogene-quaternary horses. *Science*, 355(6325):627–630, 2017.

[DG03] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[DUTG04] John C Dennis, Peter S Ungar, Mark F Teaford, and Kenneth E Glander. Dental topography and molar wear in alouatta palliata from costa rica. *American Journal of Physical Anthropology*, 125(2):152–161, 2004.

[EAJ$^+$11] Jonathan M Eastman, Michael E Alfaro, Paul Joyce, Andrew L Hipp, and Luke J Harmon. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, 65(12):3578–3589, 2011.

[EEFJ10] Jussi T Eronen, Alistair R Evans, Mikael Fortelius, and Jukka Jernvall. The impact of regional climate on the evolution of mammals: a case study using fossil horses. *Evolution: International Journal of Organic Evolution*, 64(2):398–408, 2010.

[EJ09] Alistair R Evans and Jukka Jernvall. Patterns and constraints in carnivoran and rodent dental complexity and tooth size. *J Vert Paleo*, 29:24A, 2009.

[EJ14] Alistair R Evans and Christine M Janis. The evolution of high dental complexity in the horse lineage. In *Annales Zoologici Fennici*, volume 51, pages 73–79. BioOne, 2014.

[EPF$^+$10] Jussi T Eronen, P David Polly, Marianne Fred, John Damuth, David C Frank, Volker Mosbrugger, Christoph Scheidegger, Nils Chr Stenseth, and Mikael Fortelius. Ecometrics: the traits that bind the past and present together. *Integrative Zoology*, 5(2):88–101, 2010.

[Eva74] Evan A Evans. Bending resistance and chemically induced moments in membrane bilayers. *Biophysical journal*, 14(12):923–931, 1974.

[Eva13] Alistair Robert Evans. Shape descriptors as ecometrics in dental ecology. *Hystrix, the Italian Journal of Mammalogy*, 24(1):133–140, 2013.

[EWFJ07] Alistair R Evans, Gregory P Wilson, Mikael Fortelius, and Jukka Jernvall. High-level similarity of dentitions in carnivorans and rodents. *Nature*, 445(7123):78, 2007.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[Gao16] Tingran Gao. The diffusion geometry of fibre bundles: Horizontal diffusion maps. *arXiv preprint arXiv:1602.02330*, 2016.

[GBB18] James D Gardiner, Julia Behnsen, and Charlotte A Brassey. Alpha shapes: Determining 3d shape complexity across morphologically diverse structures. *BMC Evolutionary Biology*, 18(1):184, 2018.

[GKBD19] Tingran Gao, Shahar Z Kovalsky, Doug M Boyer, and Ingrid Daubechies. Gaussian process landmarking for three-dimensional geometric morphometrics. *SIAM Journal on Mathematics of Data Science*, 1(1):237–267, 2019.

[GKD19] Tingran Gao, Shahar Z Kovalsky, and Ingrid Daubechies. Gaussian process landmarking on manifolds. *SIAM Journal on Mathematics of Data Science*, 1(1):208–236, 2019.

[Gre] WK Gregory. The origin and evolution of the human dentition, baltimore, 1922.

[GWK$^+$12] Laurie R Godfrey, Julia M Winchester, Stephen J King, Doug M Boyer, and Jukka Jernvall. Dental topography indicates ecological contraction of lemur communities. *American Journal of Physical Anthropology*, 148(2):215–227, 2012.

[Her17] James P Herrera. Testing the adaptive radiation hypothesis for the lemurs of madagascar. *Royal Society open science*, 4(1):161014, 2017.

[HLJD$^+$10] Luke J Harmon, Jonathan B Losos, T Jonathan Davies, Rosemary G Gillespie, John L Gittleman, W Bryan Jennings, Kenneth H Kozak, Mark A McPeek, Franck Moreno-Roark, Thomas J Near, et al. Early bursts of body size and shape evolution are rare in comparative data. *Evolution: International Journal of Organic Evolution*, 64(8):2385–2396, 2010.

[Hot33] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[IM13] Travis Ingram and D Luke Mahler. Surface: detecting convergent evolution from comparative data by fitting ornstein-uhlenbeck models with stepwise akaike information criterion. *Methods in Ecology and Evolution*, 4(5):416–425, 2013.

[Jol11] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[KANP$^+$05] Stephen J King, Summer J Arrigo-Nelson, Sharon T Pochron, Gina M Semprebon, Laurie R Godfrey, Patricia C Wright, and Jukka Jernvall. Dental senescence in a long-lived primate links infant survival to rainfall. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16579–16583, 2005.

[Kay78] Richard F Kay. Molar structure and diet in extant cercopithecidae. *Development, function and evolution of teeth. Academic Press, London*, pages 309–339, 1978.

[Kay84] R Kay. On the use of anatomical features to infer foraging behavior in extinct primates. *Adaptations for foraging in nonhuman primates*, pages 21–53, 1984.

[KC84] Richard F Kay and Herbert H Covert. Anatomy and behaviour of extinct primates. In *Food acquisition and processing in primates*, pages 467–508. Springer, 1984.

[KH78] Richard F Kay and WL. Hylander. The dental structure of mammalian folivores with special reference to primates and phalangeroidea (marsupialia). *The ecology of arboreal folivores*, pages 173–191, 1978.

[LMR17] Anna V Little, Mauro Maggioni, and Lorenzo Rosasco. Multiscale geometric methods for data sets i: Multiscale svd, noise and curvature. *Applied and Computational Harmonic Analysis*, 43(3):504–567, 2017.

[LTSP$^+$18] Sergi López-Torres, Keegan R Selig, Kristen A Prufrock, Derrick Lin, and Mary T Silcox. Dental topographic analysis of paromomyid (plesiadapiformes, primates) cheek teeth: more than 15 million years of changing surfaces and shifting ecologies. *Historical Biology*, 30(1-2):76–88, 2018.

[Luc04] Peter W Lucas. *Dental functional morphology: how teeth work*. Cambridge University Press, 2004.

[LWSCB13] Justin A Ledogar, Julia M Winchester, Elizabeth M St Clair, and Doug M Boyer. Diet and dental topography in pitheciine seed predators. *American Journal of Physical Anthropology*, 150(1):107–121, 2013.

[Mel17] Keegan M Melstrom. The relationship between diet and tooth complexity in living dentigerous saurians. *Journal of morphology*, 278(4):500–522, 2017.

[MK97] D Jeffery Meldrum and Richard F Kay. Nuciruptor rubricae, a new pitheciin seed predator from the miocene of colombia. *American Journal of Physical Anthropology*, 102(3):407–427, 1997.

[MR10] Carlo Meloro and Pasquale Raia. Cats and dogs down the tree: the tempo and mode of evolution in the lower carnassial of fossil and living carnivora. *Evolutionary Biology*, 37(4):177–186, 2010.

[MTS+06] Gildas Merceron, Sarah Taylor, Robert Scott, Yaowalak Chaimanee, and Jean-Jacques Jaeger. Dietary characterization of the hominoid khoratpithecus (miocene of thailand): evidence from dental topographic and microwear texture analyses. *Naturwissenschaften*, 93(7):329–333, 2006.

[MU03] Francis M'kirera and Peter S Ungar. Occlusal relief changes with molar wear in pan troglodytes troglodytes and gorilla gorilla gorilla. *American Journal of Primatology*, 60(2):31–41, 2003.

[OBA+03] Yutaka Ohtake, Alexander Belyaev, Marc Alexa, Greg Turk, and Hans-Peter Seidel. Multi-level partition of unity implicits. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 463–470. ACM, 2003.

[Osb02] Henry Fairfield Osborn. The law of adaptive radiation. *The American Naturalist*, 36(425):353–363, 1902.

[PBS16] Kristen A Prufrock, Doug M Boyer, and Mary T Silcox. The first major primate extinction: an evaluation of paleoecological dynamics of north american stem primates using a homology free measure of tooth shape. *American journal of physical anthropology*, 159(4):683–697, 2016.

[PFC+08] Rudolph Pienaar, Bruce Fischl, V Caviness, Nikos Makris, and P Ellen Grant. A methodology for analyzing curvature in the developing brain from preterm to adult. *International journal of imaging systems and technology*, 18(1):42–68, 2008.

[PH15] P DAVID Polly and Jason J Head. Measuring earth-life transitions: Ecometric analysis of functional traits from late cenozoic vertebrates. *The Paleontological Society Papers*, 21:21–46, 2015.

[PM02] J Podani and I Miklos. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*, 83(12):3331–3343, 2002.

[PSM+16] James D Pampush, Jackson P Spradley, Paul E Morse, Arianna R Harrington, Kari L Allen, Doug M Boyer, and Richard F Kay. Wear and its effects on dental topography measures in howling monkeys (alouatta palliata). *American journal of physical anthropology*, 161(4):705–721, 2016.

[PSM+18] James D Pampush, Jackson P Spradley, Paul E Morse, Darbi Griffith, Justin T Gladman, Lauren A Gonzales, and Richard F Kay. Adaptive wear-based changes in dental topography associated with atelid (mammalia: Primates) diets. *Biological Journal of the Linnean Society*, 124(4):584–606, 2018.

[PWM$^+$16] James D Pampush, Julia M Winchester, Paul E Morse, Alexander Q Vining, Doug M Boyer, and Richard F Kay. Introducing molar: A new r package for quantitative topographic analysis of teeth (and other topographic surfaces). *Journal of Mammalian Evolution*, 23(4):397–412, 2016.

[Ras03] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[Rev12] Liam J Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[RL17] Sabrina Renaud and Ronan Ledevin. Impact of wear and diet on molar row geometry and topography in the house mouse. *Archives of oral biology*, 81:31–40, 2017.

[Rob07] Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media, 2007.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[SF18] Graham J Slater and Anthony R Friscia. Hierarchy, morphology, and adaptive radiation: a test of osborn's law in the carnivora. *bioRxiv*, page 285700, 2018.

[SKW$^+$19] Shan Shan, Shahar Z Kovalsky, Julie M Winchester, Doug M Boyer, and Ingrid Daubechies. ariadne: A robustly implemented algorithm for dirichlet energy of the normal. *Methods in Ecology and Evolution*, 2019.

[SMSC17] Irepan Salvador-Martínez and Isaac Salazar-Ciudad. How complexity increases in development: An analysis of the spatial-temporal dynamics of gene expression in ciona intestinalis. *Mechanisms of development*, 144:113–124, 2017.

[SPMK17] Jackson P Spradley, James D Pampush, Paul E Morse, and Richard F Kay. Smooth operator: The effects of different 3d mesh retriangulation protocols on the computation of dirichlet normal energy. *American journal of physical anthropology*, 163(1):94–109, 2017.

[SUB$^+$05] Robert S Scott, Peter S Ungar, Torbjorn S Bergstrom, Christopher A Brown, Frederick E Grine, Mark F Teaford, and Alan Walker. Dental microwear texture analysis shows within-species diet variability in fossil hominins. *Nature*, 436(7051):693, 2005.

[TA13] Magdalena Toda and Bhagya Athukoralage. Geometry of biological membranes and willmore energy. In *AIP Conference Proceedings*, volume 1558, pages 883–886. AIP, 2013.

[TB99] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[Tor52] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[TP91] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[UB08] Peter S Ungar and Jonathan M Bunn. 11 primate dental topographic analysis and functional morphology. *Technique and application in dental anthropology*, 53:253, 2008.

[UKSH04] Lilian Ulhaas, Ottmar Kullmer, Friedemann Schrenk, and Winfried Henke. A new 3-d approach to determine functional morphology of cercopithecoid molars. *Annals of Anatomy-Anatomischer Anzeiger*, 186(5-6):487–493, 2004.

[Ung98] Peter Ungar. Dental allometry, morphology, and wear as evidence for diet in fossil primates. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 6(6):205–217, 1998.

[Ung04] Peter Ungar. Dental topography and diets of australopithecus afarensis and early homo. *Journal of Human Evolution*, 46(5):605–622, 2004.

[Ung07] Peter S Ungar. Dental topography and human evolution with comments on the diets of australopithecus africanus and paranthropus. In *Dental Perspectives on Human Evolution: State of the Art Research in Dental Paleoanthropology*, pages 321–343. Springer, 2007.

[UW00] PS Ungar and Malcolm Williamson. Exploring the effects of tooth wear on functional morphology: a preliminary study using dental topographic analysis. *Palaeontologia electronica*, 3(1):1–18, 2000.

[WBSC$^+$14] Julia M Winchester, Doug M Boyer, Elizabeth M St Clair, Ashley D Gosselin-Ildari, Siobhán B Cooke, and Justin A Ledogar. Dental topography of platyrrhines and prosimians: convergence and contrasts. *American Journal of Physical Anthropology*, 153(1):29–44, 2014.

[Wil65] Thomas J Willmore. Note on embedded surfaces. *An. Sti. Univ. "Al. I. Cuza" Iasi Sect. I a Mat.(NS) B*, 11:493–496, 1965.

[Win16] Julia M Winchester. Morphotester: an open source application for morphological topographic analysis. *PloS one*, 11(2):e0147649, 2016.

[WWS$^+$17] Ian J Wallace, Julia M Winchester, Anne Su, Doug M Boyer, and Nicolai Konow. Physical activity alters limb bone structure but not entheseal morphology. *Journal of human evolution*, 107:14–18, 2017.

[Y$^+$06] Yong-Liang Yang et al. Robust principal curvatures on multiple scales. *Symposium on Geometry Processing*, 2006.

[ZZ04] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.

Biography

Shan Shan graduated summa cum laude in mathematics from Agnes Scott College in 2014. She is completing the PhD in mathematics at Duke in September 2019 under the supervision of Professor Ingrid Daubechies. Her research interest includes dimension reduction techniques for high-dimensional data, geometry processing, probabilistic models on shapes. In 2019, her paper "ariaDNE: A robustly implemented algorithm for Dirichlet energy of the normal" was published in *Methods and Ecology*.