Statistica Sinica Preprint No: SS-2021-0163

Title	Principal Sub-manifolds
Manuscript ID	SS-2021-0163
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0163
Complete List of Authors	Zhigang Yao, Benjamin Eltzner and Tung Pham
Corresponding Authors	Zhigang Yao
E-mails	zhigang.yao@nus.edu.sg

Statistica Sinica

Principal Sub-manifolds

Zhigang Yao[1], Benjamin Eltzner[2], Tung Pham[3]

[1] *Department of Statistics and Data Science, National University of Singapore, Singapore*

[2] *Institute for Mathematical Stochastics, University of Goettingen, Germany*

[3] *School of Mathematics and Statistics, University of Melbourne, Australia*

Abstract: We propose a novel method of finding principal components in multivariate data sets that lie on an embedded nonlinear Riemannian manifold within a higher-dimensional space. Our aim is to extend the geometric interpretation of PCA, while being able to capture non-geodesic modes of variation in the data. We introduce the concept of a principal sub-manifold, a manifold passing through a reference point, and at any point on the manifold extending in the direction of highest variation in the space spanned by the eigenvectors of the local tangent space PCA. Compared to recent work for the case where the sub-manifold is of dimension one Panaretos et al. (2014)–essentially a curve lying on the manifold attempting to capture one-dimensional variation–the current setting is much more general. The principal sub-manifold is therefore an extension of the principal flow, accommodating to capture higher dimensional variation in the data. We show the principal sub-manifold yields the ball spanned by the usual principal components in Euclidean space. By means of examples, we illustrate how to find, use and interpret a principal sub-manifold and we present an application in shape analysis.

Key words and phrases: manifold, principal component analysis, tangent space, dimension reduction, shape analysis.

1. Introduction

Many quantities of interest are best described as points in a non-Euclidean space, not as vectors in a vector space. The most well-known example are directional data represented on a circle or sphere in *directional statistics*, which has been discussed as early as Fisher (1953). Higher dimensional manifold data spaces arise in the description of shapes in terms of landmarks, e.g. by Kendall (1989). In many cases, data lie close to a low dimensional sub-manifold of the data space. Approaches to restrict consideration to such a sub-manifold broadly fall into two categories. There are approaches to represent data *explicitly* on a *known sub-manifold* embedded in the data space (Kendall et al., 1999; Patrangenaru and Ellingson, 2015). Alternative approaches represent the data on an *unknown sub-manifold* in the sense that it is not embedded in the original data space, which is determined by *manifold learning* (Roweis and Sau, 2000; Donoho and Grimes, 2003; Zhang and Zha, 2004; Guhaniyogi and Dunson, 2016; Yao et al., 2023). In this paper, we discuss a method which provides an explicit, embedded sub-manifold of a manifold data space. The setting of a manifold data space in which a data sub-manifold is embedded becomes

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0163

increasingly important, as many procedures in medical imaging (Gerber et al., 2010; Souvenir and Pless, 2007) and computer vision produce high-dimensional manifold data (Pennec, 2006; Pennec and Thirion, 1997). Such methods are so far underdeveloped because conventional statistical methodology for vector spaces cannot be easily adapted to manifold spaces. The simplest case is that the existence and uniqueness of the commonly used notion of sample mean is not guaranteed anymore on a manifold (Karcher, 1977; Kendall, 1989). To quantify statistical variation on more complex features such as curves and surface a strategy of developing statistical tools in parallel with their Euclidean counterparts is highly relevant.

Previous approaches to determine an explicit data sub-manifold in a manifold data space are typically framed as efforts to generalize principal component analysis (PCA) to manifold data spaces and broadly fall into two categories. In the *forward approach*, the sub-manifold is built up by increasing dimension stepwise. Tangent space PCA (Fletcher and Joshi, 2007) attempts to project the manifold data by simply lifting them to the relevant tangent space and approximating the data distribution locally at the lifting point on the manifold with the induced Euclidean distribution. This approach only works well if the data are fairly concentrated. An alternative line of work seeks instead to directly use geodesics, which generalize the Euclidean straight lines to manifolds.

Most notable are principal geodesics (Fletcher et al., 2004) and a sequence of improvements (Huckemann and Ziezold, 2006; Huckemann et al., 2010; Kenobi et al., 2010; Jung et al., 2012; Sommer, 2013; Pennec, 2015; Eltzner et al., 2018). In shape space, many approaches use the pre-shape space of oriented shapes and the carefully deal with the quotient space structure. Using spline functions on manifolds, Jupp and Kent (1987) and Kume et al. (2007) develop smooth curves by unrolling and unwrapping the shape space. The *backward approach* carries out the procedure in reverse order from higher to lower dimension (Jung et al., 2010), discarding the direction of lowest variation at each step. The most well known approach of this type is principal nested spheres by Jung et al. (2012), which fits a a sequence of nested sub-spheres with decreasing dimension, by minimizing the residuals of the projected data in each step.

A recent approach, which retains the classical PCA interpretation at each point of the curve, is the principal flow (Panaretos et al., 2014). The flow attempts to follow the main direction of the data cloud locally and offers a trade-off between data fidelity and curve regularity. Differing from the principal flow that starts from a given reference point, Yao et al. (2024) further develops a fixed boundary flow with fixed starting and ending point for multivariate datasets lying on an embedded non-linear Riemannian manifold. More recently, inspired by finding an optimal boundary between the two classes of data lying on manifolds, Yao and Zhang (2020) invent a novel approach – the principal boundary. From the perspective of classification, the principal boundary is defined as an optimal curve that moves in between the principal flows traced out from two classes of data, and at any point on the boundary, it maximizes the margin between the two classes. In the present paper we tackle the challenging higher dimensional generalization of principal flows. The idea is to generalize the flow to a surface or more generally a sub-manifold. To find a suitable sub-manifold, we start from any point of interest on the manifold, preferably close to a large number of data points, just like we do for the principal flow; but unlike the principal flow that moves only along the maximum direction of variation of the data, we let the sub-manifold expand in all directions along multiple dimensions simultaneously. Instead of following a given shape template in every direction, the sub-manifold expands guided by the eigenvectors of the local covariance matrix.

During the preparation of this paper, which began in 2016, a preprint was published that further elaborates on the geometric theory underpinning principal submanifolds, see Akhoj et al. (2023).

In supplement S1 we present a simulation of a data set which is close to a two-dimensional sub-manifold of $S^3 \subset \mathbb{R}^4$. Figure 1 shows the data, the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0163

superimposed principal flow and the estimated two-dimensional principal submanifold. Since the surface extends in two dimensions it can for more variance of the data points.

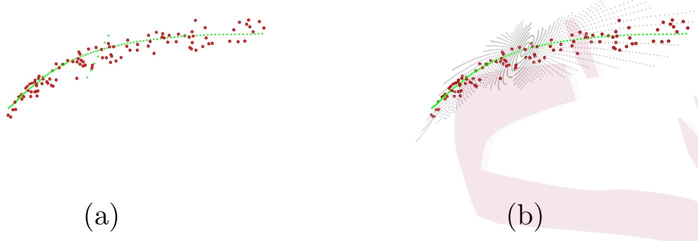


Figure 1: Visualization of the projected two-dimensional sub-manifold for data on S^3. (a) Principal flow; (b) Principal sub-manifold. The data points are labeled in red, with the first and second principal flows (in green) going through the starting point. The sub-manifold (in gray) are the estimated principal submanifold. For visualization purpose, the sub-manifold, the first and second principal direction and the data points have been projected to the first three eigenvectors of the covariance matrix at the starting point.

The optimization to determine the principal sub-manifold subject to smoothness constraints is a challenging problem. The same problem has appeared in finding the principal flow, but for higher dimensional surfaces the problem of parametrization is much more involved. We introduce two points of view on principal sub-manifolds; the first, conceptual point of view is parametrization invariant, while the second, more concrete point of view uses a specific parametrization of the surface. Since the latter point of view is more amenable to an explicit construction, our algorithm of finding the principal sub-manifold is based thereon.

We formally define the principal sub-manifold (Section 2.4) as a sub-manifold in which at any point of the sub-manifold, the tangent space of the sub-manifold attempts to be close to that of the data manifold; intuitively, this definition is analogous to the definition of the principal flow. We show that in case of a flat space, the principal sub-manifold reduces to a ball spanned by the usual principal components, in which the dimension of the sub-manifold corresponds to the number of principal components. The principal sub-manifold also provides a complementary notation to that of a principal surface by Hastie and Stuetzle (1989), as a self-consistent surface defined in Euclidean space.

2. Principal Sub-manifolds

2.1 Preliminaries

Suppose that $\{x_1, \ldots, x_n\}$ are n data points on a complete Riemannian manifold $(\mathcal{M}, g)$ of dimension m, isometrically embedded in a linear space $\mathbb{R}^d$, where $m < d$. The manifold $(\mathcal{M}, g)$ is considered known throughout all of the following. The principal sub-manifolds which are introduced here are understood to be submanifolds of this manifold $(\mathcal{M}, g)$.

Let $U \subset \mathbb{R}^d$ be an open set, which satisfies that there is an $\epsilon > 0$ such that $\{x \in \mathbb{R}^d : \exists y \in \mathcal{M} \text{ such that } |x - y| < \epsilon\} \subseteq U$. Throughout the paper, we assume that there exists a differentiable function $F : U \to \mathbb{R}^{d-m}$ such that

$$\mathcal{M} := \left\{x \in \mathbb{R}^d : F(x) = 0\right\}.$$

For each $x \in \mathcal{M}$, the tangent space at x, denoted by $T_x\mathcal{M}$ is characterized by the equation

$$T_x\mathcal{M} = \left\{y \in \mathbb{R}^d : y^T \nabla F(x) = 0\right\}.$$

Here, $\nabla F(x)$ is the $d \times (d - m)$ derivative matrix of F evaluated at $x \in \mathcal{M}$, assumed to be of full rank everywhere on $\mathcal{M}$. This full rank assumption implies that the components of F are functionally independent in a suitable sense. This tangent space $T_x\mathcal{M}$ provides a local vector space approximation of the manifold $\mathcal{M}$ analogous to the derivative of a real-valued function that provides a local approximation of the function. Let g be a smooth family of inner products associated with the manifold $\mathcal{M}$:

$$g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}.$$

Then for any $v \in T_x\mathcal{M}$, the norm of v is defined by

$$\|v\| = \sqrt{g_x(v, v)}.$$

Definition 1. An arc length parametrized curve $\gamma : [0, \delta] \to \mathcal{M}$ is a geodesic if and only if its tangent vector $\dot{\gamma}(t) \in T_{\gamma(t)}\mathcal{M} \subset \mathbb{R}^d$ satisfies

$$\frac{d(\dot{\gamma})}{dt} = 0, \quad t \in [0, \delta].$$

This means that $\frac{d(\dot{\gamma})}{dt}$ considered as a vector in $\mathbb{R}^d$ is normal to $T_{\gamma(t)}\mathcal{M}$ at any time t.

We define the manifold exponential map

$$\mathbf{exp}_x : T_x\mathcal{M} \to \mathcal{M} \tag{2.1}$$

for $\|v\| \leq \delta$ by $\mathbf{exp}_x(v) = \gamma(\|v\|)$ where γ is a geodesic starting from $\gamma(0) = x$ with initial velocity $\dot{\gamma}(0) = v/\|v\|$. For a suitable neighborhood $U_x \subset \mathcal{M}$ of x, which excludes the cut locus of x, the logarithm map

$$\mathbf{log}_x : U_x \to T_x\mathcal{M} \tag{2.2}$$

is the inverse of the exponential map.

Let $x, y \in \mathcal{M}$. Denote the set of all (piecewise) smooth curves $\gamma(t)$: $[0, 1] \to \mathcal{M}$ with endpoints such that $\gamma(0) = x$ and $\gamma(1) = y$ by $\Gamma_{x,y}$. The *geodesic distance* from x to y is defined as

$$d_{\mathcal{M}}(x, y) = \inf_{\gamma \in \Gamma_{x,y}} \ell(\gamma) \tag{2.3}$$

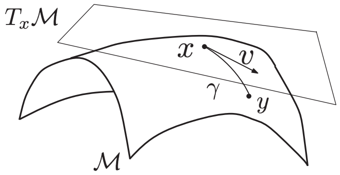


Figure 2: The vector v on the tangent subspace $T_x\mathcal{M}$ at x. The endpoint of vector v is the image of $y = \mathbf{exp}_x(v)$ under the mapping defined in (2.2).

where $\ell(\gamma) = \int_{[0,1]} \|\dot{\gamma}(t)\| \, dt = \int_{[0,1]} g_{\gamma(t)} \left(\dot{\gamma}(t), \dot{\gamma}(t)\right)^{\frac{1}{2}} dt$. Minimizing (2.3) yields geodesics as in Definition 1, providing the shortest distance between two points x and y in $\mathcal{M}$.

2.2 Principal sub-manifolds

The concept of a principal sub-manifold is strongly inspired by the principal flow, see Panaretos et al. (2014), we review this concept in Supplement S2.

In the following, we will go beyond previous work by introducing both population and sample principal sub-manifolds and discussing statistical properties. Consider a probability measure $\mathbb{P}$ on the manifold $\mathcal{M}$ as well as data $\{x_1, \cdots, x_n\} \subset \mathcal{M}$. We will give the definition of a multi-dimensional submanifold $\mathcal{N} \subset \mathcal{M}$, based on a reference point $x \in \mathcal{M}$. For any point x in $\mathcal{M}$, following Equation (S2.4) in Supplement S2, define the population and sample local tangent covariance matrices on $\mathcal{M}$

$$\Sigma_x = \frac{1}{\int_{\mathcal{M}} \kappa_h(y,x) d\mathbb{P}(y)} \int_{\mathcal{M}} \mathbf{log}_x(y) \otimes \mathbf{log}_x(y) \kappa_h(y,x) d\mathbb{P}(y), \tag{2.4}$$

$$\widehat{\Sigma}_{x,n} = \frac{1}{\sum_i \kappa_h(x_i,x)} \sum_{i=1}^{n} \mathbf{log}_x(x_i) \otimes \mathbf{log}_x(x_i) \kappa_h(x_i,x). \tag{2.5}$$

Let $\{\lambda_1(x), \ldots, \lambda_k(x)\}$ and $\{e_1(x), \ldots, e_k(x)\}$ be the first k eigenvalues and eigenvectors of Σ_x and let $\{\widehat{\lambda}_{n,1}(x), \ldots, \widehat{\lambda}_{n,k}(x)\}$ and $\{\widehat{e}_{n,1}(x), \ldots, \widehat{e}_{n,k}(x)\}$ be the first k eigenvalues and eigenvectors of $\widehat{\Sigma}_{x,n}$. Denote the linear subspace spanned by $\{e_1(x), \ldots, e_k(x)\}$ as $W(x)$ and the linear subspace spanned by $\{\widehat{e}_{n,1}(x), \ldots, \widehat{e}_{n,k}(x)\}$ as $\widehat{W}_n(x)$. For both cases, assume that each of the k principal components is smooth. Then W and $\widehat{W}_n(x)$ are smooth distributions as defined by Definition 2.

Definition 2. Let $\pi : T\mathcal{M} \to \mathcal{M}$ be the canonical projection, such that for every subset $U \subset \mathcal{M}$ the set $TU := \pi^{-1}(U) \subset T\mathcal{M}$ is well defined and contains the tangent spaces $T_p\mathcal{M}$ for all $p \in U$. Assume a subset of the tangent bundle $\mathcal{D} \subseteq T\mathcal{M}$ with the property that for every $p \in \mathcal{M}$ there is some open set $p \in U_p \subset \mathcal{M}$ such that there is a set of continuous vector fields $\mathfrak{X} = \{X_1, \ldots, X_k\}$, defined on U_p which gives rise to a homeomorphism $\mathcal{D} \cap TU_p \leftrightarrow U_p \times \mathbb{R}^k$.

(i) Then the bundle $\mathcal{D}$ is called a *distribution* of dimension k.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0163

(ii) If the vector fields used in the definition are C^r, $\mathcal{D}$ is called a C^r-*distribution* of dimension k.

(iii) If for all local vector fields $X_i, X_j \in \mathfrak{X}$ defined on the same U_p we have $[X_i, X_j] \in \mathcal{D} \cap TU_p$, the distribution is called *involutive*.

In the following, we will denote components in $\mathbb{R}^d$ by latin letters from the middle of the alphabet and components in $\mathcal{N}$ and W ranging from 1 to k by greek letter from the beginning of the alphabet.

In Supplement S6, we propose a description of principal sub-manifolds in terms of a Lagrangian problem, analogous to Panaretos et al. (2014). However, since the solution technique used there is not immediately applicable here, we instead propose an operational definition of principal sub-manifolds and provide a simpler "greedy" algorithm in Section 3.1 to approximate them.

Since principal sub-manifolds will be defined locally, we introduce the notion of a local sub-manifold.

Definition 3 (Local Sub-manifold)**.** Assume a sub-manifold, described by the image of an injective smooth function

$$N : \mathbb{R}^k \supset U \to \mathcal{N} \subset \mathcal{M} \subset \mathbb{R}^d \,. \tag{2.6}$$

In this expression, the image $\mathcal{N} := \{N(t)\}$ is the principal sub-manifold. We denote the space of local k-dimensional sub-manifolds containing some point

$A \in \mathcal{M}$, i.e. $A \in \mathcal{N}$ by this assumption, and satisfying $\forall N \in \mathcal{N} : d_{\mathcal{N}}(N, A) < L$ by $\mathrm{SubM}(A, L, k, \mathcal{M})$. Here $d_{\mathcal{N}}$ is the metric on $\mathcal{N}$ induced by the metric on $\mathcal{M}$.

Next, we define an integral sub-manifold of the distributions $W(x)$ or $\widehat{W}_n(x)$. If such integral sub-manifolds exist, the principal sub-manifolds are defined to be these integral sub-manifolds.

Definition 4 (Integral Sub-manifold)**.** A sub-manifold $\mathcal{N} \subset \mathcal{M}$ is called an *integral sub-manifold* of the distribution W, if for every point $q \in \mathcal{N}$ the tangent space is spanned by the distribution vector fields $T_q\mathcal{N} = W(q) := span\{X_1(q), \dots, X_k(q)\}$.

The following is a theorem from differential geometry on the existence of integral sub-manifolds.

Theorem 1. *For any point $p \in \mathcal{M}$ a distribution W can give rise to at most one integral sub-manifold containing p. A distribution W defines a unique C^2 integral sub-manifold for each point $p \in \mathcal{M}$ if and only if W is involutive.*

Remark 1. Involutiveness is a strong property that is not generically satisfied for the spans of eigenvectors of local covariance matrices which we consider. As an example for a simple non-involutive distribution that could arise in our setting, consider the two everywhere orthonormal vector fields

$X = \cos(y)\partial_x + \sin(y)\partial_z$ and $Y = \partial_y$ on $\mathbb{R}^3$. These define a non-involutive distribution since $[X, Y] = \sin(y)\partial_x - \cos(y)\partial_z \notin \text{span}(X, Y)$. The defining construction for principal sub-manifolds as presented in Section 3.1 always yields an interpretable sub-manifold, however it is not in general an integral sub-manifold.

In order to show the connection between population and sample principal sub-manifolds, we discuss asymptotics results in Supplement S3.

2.3 Asymptotic theory for principal sub-manifolds

For the asymptotic theory discussed here, we assume that integral sub-manifolds of $W(x)$ and $\widehat{W}_n(x)$ exist and the principal sub-manifolds are defined as these. The difference between population and sample principal sub-manifolds rests entirely on the distinction whether the distribution $W(x)$ or $\widehat{W}_n(x)$ are used, since principal sub-manifolds are defined as integral manifolds of these distributions. In Theorems 2 and 3 we also assume a fixed kernel with bandwidth h used to define both $W(x)$ and $\widehat{W}_n(x)$. We do not place specific restrictions onto the kernel or the bandwidth except for the high-level requirement that the resultant distributions $W(x)$ and $\widehat{W}_n(x)$ be involutive.

The first step towards establishing consistency and asymptotics of sample principal sub-manifolds with respect to population principal sub-manifolds requires establishing these properties for the distributions $W(x)$ and $\widehat{W}_n(x)$. Note that here and in the following we denote by $\angle(v, w)$ the angle between two vectors v and w.

Theorem 2 (Consistency of Local Covariance)**.** *If for every $x \in B_\epsilon(A)$ we have $\lambda_k(x) > \lambda_{k+1}(x)$, then we have for every $\delta > 0$ and for every sequence $a_n \to 0$*

$$\lim_{n\to\infty} \sup_{x\in B_\epsilon(A)} \mathbb{P}\left(a_n n^{1/2} \angle\left(\widehat{W}_n(x), W(x)\right) > \delta\right) = 0\,.$$

Proof. Using the CLT for principal components by Anderson (1963) the result follows immediately. □

This result does not immediately yield a consistency result for principal sub-manifolds. In fact, since $\widehat{W}_n(x)$ will in general deviate from $W(x)$ even at the reference point A, the two sub-manifolds may diverge proportionately to the distance L from the reference point A.

Theorem 3 (Consistency of Local Principal Sub-manifolds)**.** *Assume that for every $x \in B_\epsilon(A)$ we have $\lambda_k(x) > \lambda_{k+1}(x)$. Furthermore, assume a sequence $\{L_n \in \mathbb{R}^+\}_{n\in\mathbb{N}}$ which satisfies $n^{1/4}L_n \to 0$, and consider a sequence of $\{A_n \in \mathbb{R}^k\}_{n\in\mathbb{N}}$ satisfying $n^{1/2}d_{\mathcal{M}}(A_n, A) \to 0$ for some point $A \in \mathcal{N}$ on the population principal sub-manifold. Define local sample principal sub-manifolds*

$\widehat{\mathcal{N}}_n \in \mathrm{SubM}(A_n, L_n, k, \mathcal{M})$, *then we have for every* $\delta > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(n^{1/2} \max_{x\in\widehat{\mathcal{N}}_n} \min_{y\in\mathcal{N}} d_{\mathcal{M}}(x,y) > \delta \right) = 0\,.$$

Proof. The proof can be found in Supplement S3. □

The asymptotic results given above assume a fixed kernel κ and a fixed bandwidth h. This is due to the fact that the population principal sub-manifold is defined as an integral manifold of a geometric distribution defined by an optimization criterion. This might not be immediately intuitive and one might rather have the picture in mind of the population manifold being a true smooth object and the sample to be a noisy discrete representation thereof drawn via a generative model. We will therefore discuss a simple a generative model and show a consistency result for it.

Theorem 4. *Consider a sub-manifold* $\mathcal{N}_0 \subset \mathcal{M} = \mathbb{R}^d$ *and the multivariate normal probability density* $\phi(x;\mu,\Sigma)$ *on* $\mathbb{R}^d$*. Then define a one-parameter family of probability densities*

$$\phi_{n,\mathcal{N}_0}(x) = \int_{\mathcal{N}_0} \phi(x; y, \sigma_n \cdot \mathrm{Id}) dy\,.$$

Consider a sequence of bandwidths $\{h_n \in \mathbb{R}^+\}_{n\in\mathbb{N}}$ *with* $h_n \to 0$*, a sequence of noise levels* $\{\sigma_n \in \mathbb{R}^+\}_{n\in\mathbb{N}}$ *with* $n^{1/4}\sigma_n/h_n \to 0$*, and a sequence* $\{L_n \in \mathbb{R}^+\}_{n\in\mathbb{N}}$ *with* $n^{1/4}L_n \to 0$*. Then, a sequence of local population principal sub-manifolds*

$\mathcal{N}_n \in \mathrm{SubM}(A, L_n, k, \mathcal{M})$ with reference point $A \in \mathcal{N}_0$ defined by the sequence of local covariance fields

$$\Sigma_{n,x} = \frac{1}{\int_{\mathcal{M}} \kappa_{h_n}(y,x) d\mathbb{P}(y)} \int_{\mathcal{M}} \mathbf{log}_x(y) \otimes \mathbf{log}_x(y) \kappa_{h_n}(y,x) d\mathbb{P}(y),$$

leading to a sequence of distributions W_n satisfies for every $\delta > 0$

$$\lim_{n\to\infty} n^{1/2} \max_{x\in\mathcal{N}_n} \min_{y\in\mathcal{N}_0} d_{\mathcal{M}}(x,y) = 0\,.$$

Proof. The proof can be found in Supplement S3. □

3. Determination of Principal sub-manifold

3.1 An algorithm for principal sub-manifold

Recall that the principal flow is the solution of an optimization problem in equations (S2.2) and (S2.3). Finding such a solution requires an extensive search for a critical point of a Euler-Lagrange problem that involves integrating the vector field along the curve. Because it is a one dimensional curve, standard numerical methods can be applied as shown in Panaretos et al. (2014), reducing it to a problem of determining the solution of a system of ordinary differential equations (ODEs). As seen in equation (S6.6) in Supplement S6, when it comes to a sub-manifold, things turn out to be quite different. The corresponding optimization problem for sub-manifolds is much more complex and it is not clear how to approach the problem numerically. The main reason is that the

Lagrangian theory leads to a partial differential equation for which the method used in Panaretos et al. (2014) is not applicable, whereas in the case of principal flow one has a simple ordinary differential equation.

With this in mind, it is clear that the algorithm we provide should approximate curves in an integral sub-manifold, whenever the distribution is involutive.

Some complications arise when working with a sub-manifold of higher dimension than two. One problem is that computational complexity increases exponentially with dimension, which can be easily understood since the number of points in a simple rectangular grid depends exponentially on dimension. For the algorithm we present here there is an additional complication for higher dimensions, which we briefly mention below. We will discuss how to determine a two-dimensional principal sub-manifold as a special case and present an algorithm for the rest of the paper.

Principal sub-manifolds are always constructed starting from some initial point $A \in \mathcal{M}$. A possible initial point, which is used in some of the applications below, is the Fréchet mean defined below.

Definition 5. The Fréchet sample mean, $\bar{x} \in \mathcal{M}$, for a sample of data points $\{x_1, \cdots, x_n\} \in \mathcal{M}$ is a minimizer of the *Fréchet sample variance*, if the minimizer is unique:

$$\bar{x} = \operatorname*{argmin}_{x \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} d^2_{\mathcal{M}}(x, x_i).$$

Consider $\mathcal{N} \in \mathrm{SubM}(A, \epsilon, k, \mathcal{M})$ where $k = 2$. We will not perform an analytical optimization and rather present an approximating algorithm. To this end, we will work with a natural parametrization of $\mathcal{N}$ induced by the vector space structure of $\mathbf{log}_A(\mathcal{N})$. Thus, recalling the defining map from equation (2.6), we use $N := \mathbf{exp}_A$ and $U := B_\epsilon(0)$ in order to parameterize any sub-manifold $\mathcal{N} \in \mathrm{SubM}(A, \epsilon, 2, \mathcal{M})$.

We now define an equidistant pattern of directions $P := \{2j\pi/L \,|\, j = 1, \dots L\} \subset S^1 \subset \mathbb{R}^2$ and use it to define a set of points in $B_\epsilon(0) \subset \mathbb{R}^2$ as

$$\forall l \in P \subset \mathbb{R}^2 : \quad P_l := \left\{ tl \middle| t \in \{t_0 = 0, t_1, \dots, T_{N(l)}\} \subset [0, \epsilon] \right\} \subset \mathbb{R}^2 \,,$$

where $N(l)$ is the number of levels for the lth direction. These are mapped by $\mathbf{exp}_A$ into point sets, which we call *rays*,

$$\begin{aligned} \forall l \in P : \quad \mathcal{A}_l := \big\{ & A_{l,0} = A, A_{l,1} = \mathbf{exp}_A(1 \cdot l), \dots, A_{l,N(l)} = \mathbf{exp}_A(N(l) \cdot l) \big\} \\ & \subset \mathcal{M} \,, \end{aligned} \tag{3.1}$$

where we choose $L = 180$.

For dimension $k > 2$ one can also define patterns $P \subset S^{k-1}$ of finitely many directions, which are close to evenly and uniformly distributed on S^{k-1}.

However, one cannot achieve the same regularity as for an equidistant pattern in S^1. This is an additional complication which arises when trying to use the algorithm presented here in higher dimension.

In words, the key is to represent $\mathcal{N}$ discretely by a collection of ordered rays, each representing a certain amount of data variation and all of them spanning the sub-manifold of maximal variation. The rays are expected to grow and expand along all directions. While the principal flow tries to match its tangent vector to the first eigenvector at a certain point, the principal sub-manifold tries to find the best direction that belongs to the plane spanned by the first few eigenvectors, as represented by the Lagrangian $\mathscr{L}_2$. In this sense, the directions in which the sub-manifolds expands provide an extra dimension to build up the target sub-manifold of maximal variation. A set of such rays representing an approximation to the sub-manifold $\mathcal{N}$ at A—in every possible direction of variation—remain to be found.

We call all the rays for all directions a *principal sub-manifold* $\mathcal{N}$. A complete algorithm (Algorithm 1) can be found in Supplement S7. Here, we elaborate the core of the algorithm (see Figure 3): given direction l, we are at the ith level, $A_{l,i}$, there are three steps to go through to find the $(i+1)$th level, $A_{l,i+1}$

(1) *Reorientation*: identify the current tangent vector of the curve $A_{l,i}A_{l,i-1}$

and determine the direction for the next move

(2) *Projection*: expand the rays from the points $A_{l,i}$ along the direction $r_{l,i}$ by a step of ϵ' and arrive at point $A_{l,i+1}$

(3) *Updating*: project the data points x_j's$(1 \leq j \leq n)$ onto the point $A_{l,i+1}$, and re-calculate the tangent plane at $A_{l,i+1}$.

In Step (1), given the current point $A_{l,i}$ and the previous point $A_{l,i-1}$, we obtain the tangent vector of the curve $A_{l,i}A_{l,i-1}$ by backward projection

$$v_{l,i} = \mathbf{log}_{A_{l,i}}\left(A_{l,i-1}\right).$$

The best knowledge we have about the ray at $A_{l,i}$ is $v_{l,i}$. Let $u_{l,i}$ be the direction for the next move and define the projected vector

$$\tilde{v}_{l,i} := W(A_{l,i})^T W(A_{l,i}) v_{l,i} = \Big\langle v_{l,i}, e_1\big(A_{l,i}\big)\Big\rangle e_1\big(A_{l,i}\big) + \Big\langle v_{l,i}, e_2\big(A_{l,i}\big)\Big\rangle e_2\big(A_{l,i}\big)$$

where $e_1\big(A_{l,i}\big)$ and $e_2\big(A_{l,i}\big)$ are the first and second eigenvector of $\Sigma_{A_{l,i}}$. We discuss two alternatives to determine u. Let $W_{l,i}$ denote the plane spanned by $e_1\big(A_{l,i}\big)$ and $e_2\big(A_{l,i}\big)$.

(a) The straight forward choice $u_{l,i} := \tilde{v}_{l,i}$ amounts to *projection* to $W_{l,i}$.

(b) Choosing $u_{l,i} := 2\tilde{v}_{l,i} - v_{l,i}$ amounts to *reflection* at $W_{l,i}$.

While the reflection is a less obvious choice, it can achieve better data fidelity for large, but slowly varying curvature. We illustrate this point in Supplement S9.

In Step (2), we move $A_{l,i}$ on the tangent plane by a step of ϵ' along $r_{l,i}$, where

$$r_{l,i} = -\epsilon' \times \frac{u_{l,i}}{\|u_{l,i}\|},$$

and then map it back to the manifold $\mathcal{M}$

$$A_{l,i+1} = \mathbf{exp}_{A_{l,i}}(r_{l,i}).$$

Note that $u_{l,i}$ is not of unit length, and the negative sign appears as $u_{l,i}$ is obtained from $v_{l,i}$.

In Step (3), updating the covariance matrix at $A_{l,i+1}$ is necessary when local data points change significantly, where the covariance matrix is updated by replacing $\Sigma_{A_{l,i}}$ with $\Sigma_{A_{l,i+1}}$.

It is crucial to make sure that the rays always move forward and do not return to points already explored by the ray itself or another ray. Additionally, a stop condition is necessary such hat the principal sub-manifold does not extend far beyond the data domain. In accordance with the stopping rule used in Panaretos et al. (2014), we can terminate the process when the *length*

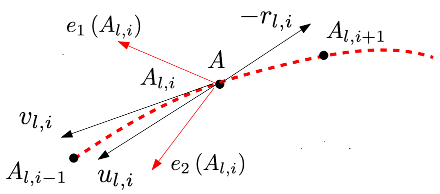


Figure 3: Illustration of the Algorithm. The goal is to determine $A_{l,i+1}$, given the current point $A_{l,i-1}$ and the previous point $A_{l,i}$: $v_{l,i}$ is the tangent vector of the curve $A_{l,i}A_{l,i-1}$, $u_{l,i}$ is the projection of $v_{l,i}$ onto the tangent plane spanned by $e_1(A_{l,i})$ and $e_2(A_{l,i})$. The point $A_{l,i+1}$ is found by mapping a small move ϵ' from $A_{l,i}$ along the direction of $r_{l,i} = -u_{l,i}/\|u_{l,i}\|$.

of the lth ray, i.e.,

$$\ell_{\mathcal{A}_l} = \sum_{i=1}^{N(l)-1} d(A_{l,i}, A_{l,i+1}),$$

exceeds 1. The length of lth ray does not necessarily have to be equal. There may exist other stopping rules that one can use. Among them, we should also consider that for all j,

$$\left\|\mathbf{log}_{A_{l,i+1}}(x_j)\right\| > \delta \text{ or } \left\langle \mathbf{log}_{A_{l,i+1}}(A_{l,i}), \mathbf{log}_{A_{l,i+1}}(x_j)\right\rangle \geq 0,$$

which implies that either there are not enough data points in the neighborhood or $A_{l,i+1}$ is already outside the convex hull of the x_j's under the logarithm map.

Remark 2. Both ϵ' and δ are pre-defined parameters. We suggest to choose ϵ' preferably with small values to ensure the stability of the local move on the tangent plane, while the choice of δ depends more on the data dispersion and configuration, which might vary from case to case.

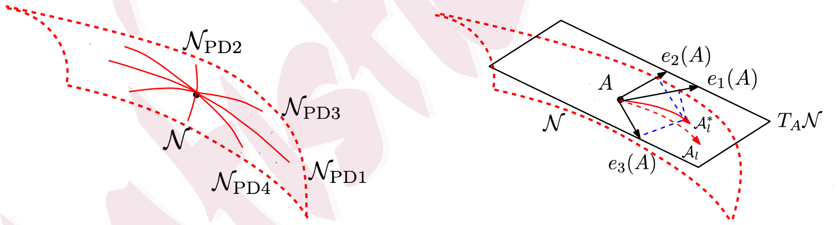
For $h \to \infty$ and a flat manifold the principal component distribution is constant over the whole space and therefore it is clear that the greedy algorithm leads to straight lines spanning the linear subspace spanned by the k largest principal components. In this sense, the limiting case of standard PCA is trivial. In Supplement S8 we investigate the convergence behavior of the greedy algorithm in the limit $\epsilon' \to 0$ if the length of all curves is fixed in advance. In general, it is very difficult to show that solution curves of the greedy algorithm approximate solution curves to either Lagrangian. Instead, we show that the curves converge to the integral sub-manifold if the distribution is involutive, as expected. This convergence result is very important, because it shows that the greedy algorithm leads to meaningful results in all cases where a unique "true" geometric solution in terms of an integral sub-manifold exists.

3.2 Visualization of the principal sub-manifold

The principal sub-manifold in general cannot be fully visualized when its dimension exceeds one. Consider a simple case where the data lies in $S^3 \subset \mathbb{R}^4$; the principal sub-manifold is then a subset of S^3; that is, it is equivalent to visualizing a two-dimensional manifold in a four-dimensional space. However, a meaningful representation of the sub-manifold is still quite relevant for understanding the shape of the manifold, at least partially. We propose two ways of visualizing the principal sub-manifold. The first one is to represent the sub-manifold in terms of principal direction rays. The second one is to visualize the sub-manifold in the projected manifold space.

- parameterize the sub-manifold in polar coordinates and represent it by the shapes of principal direction rays

- project the sub-manifold by multiplying a projection matrix in which the basis is formed by eigenvectors from the covariance matrix at the starting point


$\mathcal{N}_{\mathrm{PD2}}$
$\mathcal{N}_{\mathrm{PD3}}$
$\mathcal{N}$
$\mathcal{N}_{\mathrm{PD4}}$
$\mathcal{N}_{\mathrm{PD1}}$
$e_2(A)$
A
$e_1(A)$
$\mathcal{N}$
A_l^*
$T_A\mathcal{N}$
A_l
$e_3(A)$

(a) (b)

Figure 4: Visualization of a principal sub-manifold. (a) Visualize the sub-manifold by eight principal direction rays. (b) Visualize the sub-manifold by projecting to the three largest eigenvectors of covariance matrix at A.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0163

Visualization in principal direction rays: Choose a number of directions from the starting point of the algorithm and visualize the sub-manifold by using the corresponding rays. Recall that the entire sub-manifold can be expressed as follows

$$\mathcal{N} = \left[\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_L\right]^{\mathrm{T}}.$$

Although we denote $\mathcal{N}$ as a "matrix", the actual length of each row (i.e., $\mathcal{A}_l, 1 \le l \le L$) may vary. To visualize the sub-manifold, we select a candidate set $\mathcal{L}_s \subset \mathcal{L}$ and map the corresponding rows of $\mathcal{N}$ into the corresponding rays in shape coordinates. Thus, the principal direction rays of the sub-manifold shall be represented by

$$x_{l,i} = f^{-1}(A_{l,i}), \quad \forall l \in \mathcal{L}_s, 1 \le i \le N(l),$$

where f is the embedding function. In the case of Kendall shape space, the resultant $x_{l,1}, \cdots, x_{l,N(l)}$ is a collection of $N(l)$ k-ads.

Among all l's, the two *principal directions* of the sub-manifold are defined as follows. Recall polar coordinates on the image $\theta = 2l\pi/L$, where $L = 180$. The first principal direction, denoted as $\mathcal{N}_{\mathrm{PD1}}$, is the curve corresponding to $\theta = \pi$ and $\theta = 2\pi$ in polar coordinates on the image; this is equivalent to l equals 90 and 180 such that

$$\mathcal{N}_{\mathrm{PD1}} = \mathcal{A}_{90} \cup \mathcal{A}_{180}.$$

The second principal direction, denoted as $\mathcal{N}_{\text{PD2}}$, corresponds to the curve with $\theta = \pi/2$ and $\theta = 3\pi/2$ in polar coordinates on the same image; this is equivalent to l equals 45 and 135 such that

$$\mathcal{N}_{\text{PD2}} = \mathcal{A}_{45} \cup \mathcal{A}_{135}.$$

In addition, it is suggested to also include the curves, $\mathcal{N}_{\text{PD3}}$, corresponding to $\theta = \pi/4$ and $\theta = 5\pi/4$ as well as the ones, $\mathcal{N}_{\text{PD4}}$, corresponding to $\theta = 3\pi/4$ and $\theta = 7\pi/4$. Adding two extra directions gives additional details about the sub-manifold.

We remark here that although we have used $\mathcal{N}_{\text{PD1}} - \mathcal{N}_{\text{PD4}}$ as the principal directions, they are by no means the simple extension of the usual principal components or any variants thereof. Figure 5 gives an example of such a configuration of shapes. The entire image contains 9 by 9 small shapes. The central figure is the mean shape. Row 5 represents the shapes of $\mathcal{N}_{\text{PD1}}$. Column 5 is the shapes of $\mathcal{N}_{\text{PD2}}$. The main diagonal contains the shapes of $\mathcal{N}_{\text{PD3}}$. The other diagonal contains the shapes $\mathcal{N}_{\text{PD4}}$.

Visualization in projected space: Alternatively, one may wish to represent the sub-manifold using a projected sub-manifold rather than itself. The latter serves as a much simplified version of the original one and it is more interpretable, provided that the majority of variation of the principal sub-manifold can be explained by a reduced one. Compared to the previous representation, this visualization preserves the resolution of the sub-manifold.

To fix representation, we center the $\mathcal{N}$ row-wise by A and obtain the centered sub-manifold

$$\mathcal{N}^* = \left(\mathcal{N}^*_{l,i}\right)_{1\le l\le L, 1\le i\le N(l)}$$

where $\mathcal{N}^*_{l,i} = A_{l,i} - A$ where $1 \le l \le L, 1 \le i \le N(l)$. Clearly, for $i = 1$, $\mathcal{N}^*_{l,i} = \mathbf{0}$. Let the projection matrix for $\mathcal{N}^*$ be

$$\Psi = (\psi_{l,i})_{1\le l\le L, 1\le i\le N(l)}$$

where $\psi_{l,i}$ is the projection matrix for $A_{l,i}$. Usually, we choose $\psi_{l,i} = E_3$ where $E_3 = [e_1(A), e_2(A), e_3(A)]^{\mathrm{T}}$ of Σ_A. The process is carried out by multiplying $\mathcal{N}^*$ element-wise by the projection Ψ, so that

$$\mathcal{N}^{\text{pro}} = \Psi \odot \mathcal{N}^*$$

where $\odot$ is the element-wise product such that $\mathcal{N}^{\text{pro}}_{l,i} = \psi_{l,i}\mathcal{N}^*_{l,i}$. Figure 4(b) illustrates the main idea: the red dashed arrow starting from A denotes the lth ray (or a vector of $(A_{l,1}, \ldots, A_{l,N(l)})$) of the principal sub-manifold $\mathcal{N}$; the red solid arrow denotes the projected lth ray of the principal sub-manifold, $\mathcal{A}^*_l$. The point A is now regarded as the new origin under the new coordinate system, correspondingly. Moreover, the data points x_j's are projected in the same way by

$$x_j^* = E_3(x_j - A), \quad 1 \leq j \leq n.$$

In general, the projected points x_j^*'s are expected to lie closely to the sub-manifold $\mathcal{N}^{\text{pro}}$, provided that the projection matrix has accounted for most of the variability.

4. Applications

This section contains an illustration of principal sub-manifolds on a data set of handwritten digits. Additional simulations are presented in Supplement S10-11 and two more applications can be found in Supplement S12-13. Since we are concerned with landmark shapes, we provide a brief introduction to that topic in Supplement S14.

To illustrate the use of the principal sub-manifold in a concrete example, we consider a handwritten digit "3" data. The data, included in the GNU R package *shapes*, consists of 13 landmarks of a "3" in two dimensions, collected from 30 individuals. For visualization, we find a principal sub-manifold for the data and recover the shape variation of the "3" in four principal directions, started at two different shapes of the "3". In the first case (see Figure 5), the sub-manifold starts from the Fréchet mean of the data. In each principal

see `https://cran.r-project.org/web/packages/shapes/index.html` direction, the flow of images describes the shapes of the “3” moving from one extreme to the other extreme. The horizontal set of the images represents the various shapes of “3” recovered from the first principal direction. From there, we can see that the most varying part is the middle part of the “3”. The parts varying in the second principal direction are mainly the upper and lower parts of the “3”. Those parts of the “3” have exhibited a significant shape change along the two principal directions. Both the main diagonal and the other diagonal show certain degrees of the shape change mostly in the middle part of the “3” but in an opposite direction. By observing the fact that there are two seemingly outlying individuals of “3”s deviating from the rest in the data—the midpoint of the 3 having moved away from the center of the figure—a more sensible center of symmetry should be also considered. As in the second case (Supplement S4) serves to illustrate the slight effect of having a different choice (center of symmetry) of the starting point on the sub-manifold. However, no significant change in the representation of the sub-manifold is found.

To further understand the shape variation in configuration space, we contrast the results with that from the standard generalized Procrustes analysis (GPA). The profiling of shapes obtained from both methods along different principal directions (or principal components) in Figure 6 has suggested quite different patterns. Not only does the variation differ at various parts of the

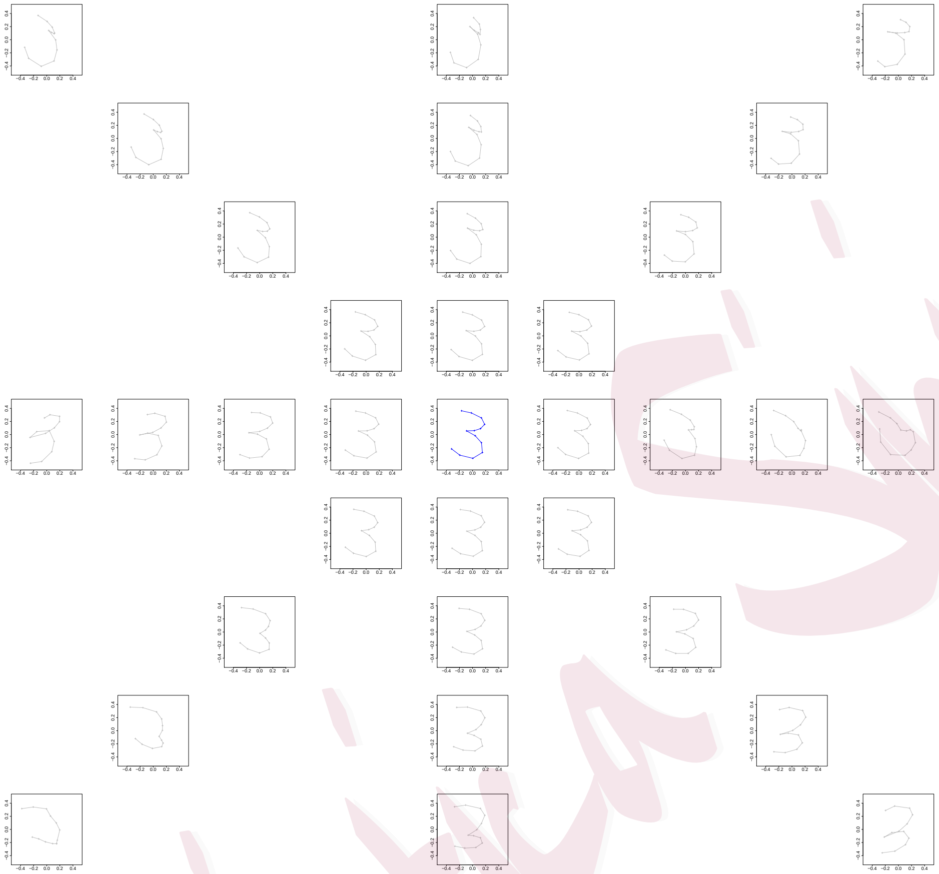
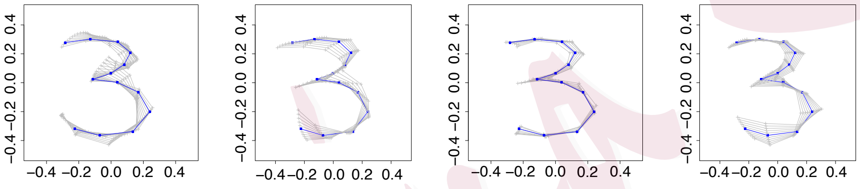


Figure 5: Principal sub-manifolds of the handwritten digits data, started from the mean. Among all the figures: the central figure (in blue) is the Fréchet mean; the horizontal row contains images recovered from the first principal direction of the sub-manifold; the vertical column is the second principal direction; the main diagonal is the third principal direction; the other diagonal is the fourth principal direction.

"3", but the images of shapes recovered from the principal directions of the sub-manifold reveals an phenomenon of asymmetrical variation around the Procrustes mean, compared to the GPA: the principal sub-manifold tries to explain most of the variation by its first principal direction, while the GPA explains the variation almost equally along its first and second principal components. This is interesting to us, as this information is not available from standard procedures that profiles the images in the configuration space where they obey a standard PCA manner.



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

Figure 6: Principal sub-manifolds and generalized Procrustes analysis on the handwritten digits data. Figures (a) and (b): the central figure (in blue) is the Procrustes mean; (a) contains images recovered from the first principal direction of the principal sub-manifold; (b) contains images recovered from the first principal component of generalized Procrustes analysis. Figures (c) and (d) give the same information for second principal direction (or principal component) of the principal sub-manifold (or generalized Procrustes analysis).

5. Discussion

The statistical analysis of data on Riemannian manifold is a very challenging topic and it plays an increasingly important role in real-world problems. Conventional approaches, such as PCA in Euclidean space, are essentially helpful in neither learning the shape of the underlying manifold nor deciding its dimensionality. The main reason for this lies in the fact that those approaches simply do not use the intrinsic Riemannian manifold structures.

With the aim of proposing a method that allows for finding a nonlinear manifold from the data, we introduced the notation of principal sub-manifold. We showed the importance of estimating a multi-dimensional sub-manifold, and its difference from finding only a one-dimensional curve. The principal sub-manifold was seen to be interpretable as a measure of non-geodesic variation of the data. Based on a polar coordinate representation, the principal sub-manifold was constructed so that it coordinated with the local data variation. We illustrated that the principal sub-manifold is an extension of the principal flow, in the sense that it depicts a multi-dimensional manifold. When the manifold is linear, the k-dimensional principal sub-manifold reduces to the ball spanned by the principal component vectors corresponding to the k largest eigenvalues.

We claim here that by definition, the implemented principal directions might or might not coincide with the principal flows that are defined in Panaretos et al. (2014), although in practice, they appear to be close to or the same as the principal flows. Under the polar coordinate representation, we observe that the principal directions (these are plotted in green in Figure 1) on the principal sub-manifold have presented the main modes of variation.

Regarding the issue of choosing the locality parameter h, or equivalently, which scale of the local covariance one should consider, we note that different sub-manifolds in this article have been fitted by choosing different parameters. Still, we refrain from making a strict statement on optimizing the h; rather, one should overview a sequence of h. Possible routes to approach this question are suggested by the criterion in Panaretos et al. (2014), which could be adapted, and the scale space perspective Chaudhuri and Marron (2000). Simultaneously, we were able to define the principal sub-manifold to any dimension $k \leq d$, and this may also be seen as the development of a heuristic understanding of a backward stepwise principle of PCA on manifolds: in backward PCA, the best approximating affine subspaces are constructed from the highest dimension to the lowest one, see Jung et al. (2010) for the case of spherical subspaces, while in the case of principal sub-manifolds, each ray of the principal sub-manifolds (i.e., the principal directions) corresponds to lower dimension sub-manifolds, compared to the entire sub-manifold.

Last but not least, the formulation of the principal sub-manifold opens the way to the generalization of many other statistical procedures. From the variance reduction perspective, one may categorize our proposed method as one of those competing methods that extend PCA on manifolds but not limited to only using lines or curves. This, potentially, can help us understand the data variation better and improve accuracy. From the classification point of view, this new method has been seen to be a useful tool to study shape changes. In the leaf growth example (Supplement S12), we studied the only two main modes of shape variation. This implies that one can extend a classification framework to manifolds. By projecting the new data points to any principal direction of the sub-manifold, one can calculate the distance and extend a classification rule based on all the distances. Surely, a successful classification also depends on 1) the data configuration; 2) how to define the local covariance matrix. If the data on the manifold is not too dense, one might consider using a kernel density estimation. The label information also needs to be considered in the local covariance matrix, in which one would account for both of the between class and within class effects. As this is one of our on-going works, we will investigate it in the future.

Supplementary Material

PDF file *Principal Sub-manifolds – Supplementary Materials*: This PDF file contains additional background theory, some additional simulation results to illustrate the greedy algorithm proposed here, and additional applications.

Acknowledgments

Z. Yao has been supported by Singapore Ministry of Education Tier 2 grant (A-0008520-00-00, A-8001562-00-00) and Tier 1 grant (A-0004809-00-00, A8000987-00-00) at the National University of Singapore. B. Eltzner gratefully acknowledges funding by the DFG CRC 803 project Z02 and DFG CRC 1456 project B02. B. Eltzner is very grateful to the National University of Singapore for funding a two-month long term visit in February and March 2018 which enabled collaboration on this project. We thank S. F. Huckemann and J. S. Marron for helpful discussions.

References

Akhoj, M., J. Benn, E. Grong, S. Sommer, and X. Pennec (2023). Principal subbundles for dimension reduction. *arXiv preprint arXiv:2307.03128*.

Anderson, T. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics 34*, 122–148.

Chaudhuri, P. and J. S. Marron (2000). Scale space view of curve estimation. *The Annals of Statistics 28*, 408–428.

Donoho, D. L. and C. Grimes (2003). Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America 100*, 5591–5596.

Eltzner, B., S. Huckemann, and K. V. Mardia (2018). Torus principal component analysis with applications to RNA structure. *The Annals of Applied Statistics 12*, 1332–1359.

Fisher, R. (1953). Dispersion on a sphere. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 217*, 295–305.

Fletcher, P. T. and S. Joshi (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing 87*, 250–262.

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

Gerber, S., T. Tasdizen, P. T. Fletcher, S. Joshi, R. Whitaker, and the Alzheimers Disease Neuroimaging Initiative (ADNI) (2010). Manifold modeling for brain population analysis. *Medical Image Analysis 14*, 643–653.

Guhaniyogi, R. and D. Dunson (2016). Compressed gaussian process for manifold regression. *Journal of Machine Learning Research 17*, 1–26.

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

Huckemann, S., T. Hotz, and A. Munk (2010). Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica 20*, 1–100.

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

Jung, S., I. L. Dryden, and J. S. Marron (2012). Analysis of principal nested spheres. *Biometrika 99*, 551–568.

Jung, S., X. Liu, J. S. Marron, and S. M. Pizer (2010). Generalized pca via the backward stepwise approach in image analysis. In *Brain, Body and Machine*, pp. 111–123. Springer: Berlin/Heidelberg.

Jupp, P. E. and J. T. Kent (1987). Fitting smooth paths to spherical data. *Journal of the Royal Statistical Soceity, Series C 36*, 34–36.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communication on Pure and Applied Mathematics 30*, 509–541.

Kendall, D. G. (1989). A survey of the statistical theory of shape. *Statistical Science 4*, 87–120.

Kendall, D. G., D. Barden, T. K. Carne, and H. Le (1999). *Shape and Shape*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0163

Theory. New York: Wiley.

Kenobi, K., I. L. Dryden, and H. Le (2010). Shape curves and geodesic modelling. *Biometrika 97*, 567–584.

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

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

Patrangenaru, V. and L. Ellingson (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press.

Pennec, X. (2006). Intrinsic statistics on riemannian manifolds: basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision 25*, 127–154.

Pennec, X. (2015). Barycentric subspaces and affine spans in manifolds. In *Geometric Science of Information (GSI) 2015*, pp. 12–21. Springer International Publishing.

Pennec, X. and J.-P. Thirion (1997). A framework for uncertainty and validation of 3d registration a framework for uncertainty and validation of 3d registration methods based on points and frames. *International Journal of Computer Vision 25*, 203–229.

Roweis, S. T. and L. K. Sau (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 2323–2326.

Sommer, S. (2013). Horizontal dimensionality reduction and iterated frame bundle development. In *Geometric Science of Information (GSI) 2013*, pp. 76–83. Springer: Berlin/Heidelberg.

Souvenir, R. and R. Pless (2007). Image distance functions for manifold learning. *Image and Vision Computing 25*, 365–373.

Yao, Z., J. Su, and S.-T. Yau (2023). Manifold fitting with cyclegan. *Proceedings of the National Academy of Sciences of the United States of America 121*, e2311436121.

Yao, Z., Y. Xia, and Z. Fan (2024). Random fixed boundary flows. *Journal of the American Statistical Association 119*(547), 2356–2368.

Yao, Z. and Z. Zhang (2020). Principal boundary on riemannian manifolds. *Journal of the American Statistical Association 115*, 1435–1448.

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

Zhigang Yao

Department of Statistics and Data Science

National University of Singapore

Singapore 117546

E-mail: zhigang.yao@nus.edu.sg

Benjamin Eltzner

Institute for Mathematical Stochastics

University of Goettingen

37077 Goettingen, Germany

E-mail: beltzne@uni-goettingen.de

Tung Pham

School of Mathematics and Statistics

University of Melbourne

Victoria 3010 Australia

E-mail: pham.t@unimelb.edu.au