Principal Sub-manifolds – Supplementary Materials

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

[2]*Max-Planck-Institute for Biophysical Chemistry, Goettingen, Germany*

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

Supplementary Material

S1 Illustration referenced in the Introduction

The geometry plays an essential role for the shape of such a sub-manifold. To illustrate the concept of a principal sub-manifold, consider a set of data points $(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$ on $S^3 \subset \mathbb{R}^4$, where the coordinates $(x_{i,1}, x_{i,2}, x_{i,3}) \subset \mathbb{R}^3$ of each data point form a rough sinusoid and the fourth coordinate $x_{i,4}$ is added so that every point lies on a sphere. The data is originally constructed by sampling the triplets $(x_{i,1}, x_{i,2}, x_{i,3}) \subset \mathbb{R}^3$ from the distribution

$$\begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{pmatrix} = \begin{pmatrix} (i - n/2)/n \\ \sin(2x_{i,1})/6 + 32U \\ 1 + 1/100V \end{pmatrix}, \quad 1 \leq i \leq n$$

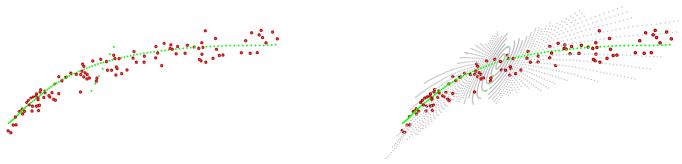
where U and V are independent normal variable $N(0, 1/10)$ and $N(0, 1/100)$. To lift each point to $\mathbb{R}^4$, we add the fourth coordinate so that every point lies on a sphere satisfying

$$x_{i,1}^2 + x_{i,2}^2 + x_{i,3}^2 + x_{i,4}^2 = C,$$

where a shifting parameter C (e.g., 0.45) is chosen such that

$$\sqrt{C - x_{i,1}^2 - x_{i,2}^2 - x_{i,3}^2} \geq 0.$$

We still need to normalize the data to guarantee that the data are in S^3.



(a) (b)

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 sub-manifold. 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 ambient manifold determines some of the inherited geometry of the projected sub-manifold. In this case, the sub-manifold carries an S-pattern

that mainly stems from the rough sinusoid of the first three coordinates. Figure 1 shows the data, the superimposed principal flow and the estimated principal sub-manifold. The starting point (in black) is the center of the data points (in red). The two green curves (Figure 1(a)) are the first and second principal flows, while the estimated principal sub-manifold (Figure 1(b)) is highlighted in gray, contrasting the two green principal flows lying on the surface of the estimated principal sub-manifold. All of them are projected to the first three eigenvectors of the covariance matrix at the starting point. The surface is able to bend wherever the curvature of the manifold changes rapidly. The first principal direction is towards the direction of maximum variance of the data points, while the surface extends in more directions that automatically account for more variance of the data points.

S2 Principal flows

Since the concept of a principal sub-manifold is strongly inspired by the principal flow, see Panaretos et al. (2014), we will review this concept here. The principal flow yields a one-dimensional, not necessarily geodesic, approximation of a data set $\{x_1, \cdots, x_n\} \subset \mathcal{M}$. We parameterize the one

dimensional sub-manifolds as a set of unit speed curves

$$\mathrm{SubM}(A,1,v,\mathcal{M}) = \Big\{\gamma : [0,r] \to \mathcal{M}, \gamma \in C^2(\mathcal{M}), \gamma(s) \neq \gamma(s') \text{ for } s \neq s', \\ \gamma(0) = A, \dot\gamma(0) = v, \ell(\gamma[0,t]) = t \text{ for all } 0 \leq t \leq r \leq 1\Big\}, \quad \text{(S2.1)}$$

where $\gamma(0) = A$ and $\dot\gamma(0) = v$ are initial conditions for γ and $\ell(\gamma)$ is the length of γ. The starting point A can be chosen as the Fréchet sample mean $\bar{x}$ or any other point of interest. Then $\mathrm{SubM}(A,1,v,\mathcal{M})$ contains all smooth curves of length less than 1 with given initial speed and starting point.

The principal flow is defined by two curves

$$\gamma^+ = \arg \sup_{\gamma \in \mathrm{SubM}(A,1,v_1,\mathcal{M})} \int_0^{\ell(\gamma)} \langle \dot\gamma(t), e_1(\gamma(t)) \rangle \, dt \quad \text{(S2.2)}$$

$$\gamma^- = \arg \inf_{\gamma \in \mathrm{SubM}(A,1,v_2,\mathcal{M})} \int_0^{\ell(\gamma)} \langle \dot\gamma(t), e_1(\gamma(t)) \rangle \, dt \quad \text{(S2.3)}$$

where $v_1 = e_1(\gamma(t))$, $v_2 = -v_1$, $e_1(\gamma(t))$ is the first eigenvector of the covariance matrix $\Sigma_{\gamma(t)}$ at $\gamma(t)$. The integral for γ^- is negative, therefore the infimum appears in its definition. At each point of γ, $\dot\gamma(t)$ is maximally compatible to the eigenvector to the largest eigenvalue of the local covariance matrix at scale h

$$\Sigma_{h,\gamma(t)} = \frac{1}{\sum_i \kappa_h(x_i, \gamma(t))} \sum_{i=1}^n \mathbf{log}_{\gamma(t)}(x_i) \otimes \mathbf{log}_{\gamma(t)}(x_i) \kappa_h(x_i, \gamma(t)) \,, \quad \text{(S2.4)}$$

where $y \otimes y := yy^{\mathrm{T}}$ and $\kappa_h(x, \gamma(t)) = K(h^{-1} d_{\mathcal{M}}(x, \gamma(t))$ with a smooth

non-increasing kernel K on $[0, \infty]$. All the above definitions are under the assumption that the first and second eigenvalues of $\Sigma_{h,\gamma(t)}$ are distinct.

Principal flows achieve higher data fidelity than geodesics and are more flexible than other curve-fitting approaches in trading off between data fidelity and avoiding too high curvature of the curve. The question whether non-linear variation can be captured in higher dimension has been discussed (for other assumptions on the embedding data space) under the names of *principal curves* or *principal surfaces* by Hastie and Stuetzle (1989). Note that principal surfaces are the extension of principal curves to higher dimensions in Euclidean space, restricted to a two-dimensional scenario. The present work is connected to both principal flow and principal surfaces, using a more general setting than Hastie and Stuetzle (1989).

S3 Proofs of Theorems 3 and 4

Proof of Theorem 3. First, since $\widehat{\mathcal{N}}_n$ is C^2, we can use a Taylor expansion around A_n to see that for any $x \in \widehat{\mathcal{N}}_n$ we have some $w_x \in \widehat{W}_n(A_n)$ with $|w_x| = 1$ such that

$$x = A_n + |A_n - x| w_x + \mathcal{O}(L_n^2)\,.$$

Next, we note that since W is C^2, we have a constant $C > 0$, such that

$$\angle\left(\widehat{W}_n(A_n), W(A)\right) \leq \angle\left(\widehat{W}_n(A_n), W(A_n)\right) + Cd_{\mathcal{M}}(A_n, A) + \mathcal{O}(d_{\mathcal{M}}(A_n, A)^2)\,.$$

Now, define the vector $w_y \in W(A)$ by

$$w_y := \underset{w_y \in W(A), |w_y|=1}{\operatorname{argmin}} \angle(w_x, w_y)$$

and let γ_{A,w_y} be the integral curve of W starting at A with tangent vector w_y which stays closest to the straight line $c(t) := A + w_y t$. Then define $y := \gamma_{A,w_y}(|A_n - x|) \in \mathcal{N}$. By construction, we now have

$$y = A + |A_n - x| w_y + \mathcal{O}(L_n^2)\,.$$

Thus we get the following bound

$$\begin{aligned}
d_{\mathcal{M}}(x, y) &\leq d_{\mathcal{M}}(A_n, A) + |A_n - x| \angle(w_x, w_y) + \mathcal{O}(L_n^2) \\
&\leq d_{\mathcal{M}}(A_n, A) + |A_n - x| \angle\left(\widehat{W}_n(A_n), W(A)\right) + \mathcal{O}(L_n^2) \\
&\leq d_{\mathcal{M}}(A_n, A) + |A_n - x| \angle\left(\widehat{W}_n(A_n), W(A_n)\right) \\
&\quad + C|A_n - x| d_{\mathcal{M}}(A_n, A) + |A_n - x| \mathcal{O}(d_{\mathcal{M}}(A_n, A)^2) + \mathcal{O}(L_n^2) \\
&\leq (1 + CL_n) d_{\mathcal{M}}(A_n, A) + L_n \angle\left(\widehat{W}_n(A_n), W(A_n)\right) \\
&\quad + \mathcal{O}(L_n d_{\mathcal{M}}(A_n, A)^2) + \mathcal{O}(L_n^2)\,.
\end{aligned}$$

Using this bound, we see that

$$n^{1/2} d_{\mathcal{M}}(x,y) \leq (1+CL_n) n^{1/2} d_{\mathcal{M}}(A_n, A) + L_n n^{1/2} \angle\left(\widehat{W}_n(A_n), W(A_n)\right) + \mathcal{O}(n^{1/2} L_n d_{\mathcal{M}}(A_n, A)^2) + \mathcal{O}(n^{1/2} L_n^2)\,.$$

Now, using the assumptions $n^{1/2} d_{\mathcal{M}}(A_n, A) \to 0$ and $n^{1/4} L_n \to 0$, we get

$$n^{1/2} d_{\mathcal{M}}(x,y) \to L_n n^{1/2} \angle\left(\widehat{W}_n(A_n), W(A_n)\right)\,,$$

which goes to 0 in probability due to Theorem 2 as desired. □

Proof of Theorem 4. There are vectors $w_x \in W_n(A)$ with $|w_x| = 1$ and $w_y \in W(A)$ with $|w_y| = 1$ such that

$$x = A + |A-x| w_x + \mathcal{O}(L_n^2) \qquad y = A + |A-y| w_y + \mathcal{O}(L_n^2)\,.$$

In consequence,

$$\begin{aligned} d_{\mathcal{M}}(x,y) &\leq \max(|A-x|, |A-y|) \angle(w_x, w_y) + \mathcal{O}(L_n^2) \\ &\leq L_n \angle(W_n(A), W(A)) + \mathcal{O}(L_n^2)\,, \end{aligned}$$

so it remains to show that

$$n^{1/4} \angle(W_n(A), W(A)) \xrightarrow{\mathbb{P}} 0\,.$$

Because of $\sigma_n / h_n \to 0$ and $h_n \to 0$, there is a constant C_1 such that for $v \in T_A \mathcal{N}_0$ with $|v| = 1$ one gets $v^T \Sigma_{n,A} v \to C_1 h_n^2$ and a constant C_2 such that for $v \in N_A \mathcal{N}_0$ normal to $\mathcal{N}_0$ with $|v| = 1$ Because of $\sigma_n / h_n \to 0$ and

$h_n \to 0$, one gets $v^T \Sigma_{n,A} v \to h_n$ for $v \in T_x \mathcal{N}_0$ and $|v| = 1$ while $v^T \Sigma_{n,A} v \to \sigma_n$ for $v \in N_x \mathcal{N}_0$ normal to $\mathcal{N}_0$ and $|v| = 1$ one gets $v^T \Sigma_{n,A} v \to C_2 \sigma_n^2$.

As a result, the eigenvectors of $\Sigma_{n,A}$ corresponding to the k largest eigenvalues approach the tangent space $T_x \mathcal{N}_0$ with the maximal angle converging in probability

$$\begin{aligned} \angle\left(W_n(A), W(A)\right) = & \arctan\left(\frac{C_2 \sigma_n}{C_1 h_n}\right) \mathcal{O}_p(1) \\ \Rightarrow \angle\left(W_n(A), W(A)\right) = & \left(\frac{C_2 \sigma_n}{C_1 h_n} + \mathcal{O}\left(\frac{\sigma_n^2}{h_n^2}\right)\right) \mathcal{O}_p(1)\,. \end{aligned}$$

The claim then follows from $n^{1/4}\sigma_n / h_n \to 0$. □

S4 Principal sub-manifolds of the handwritten digits data, started from the center of symmetry

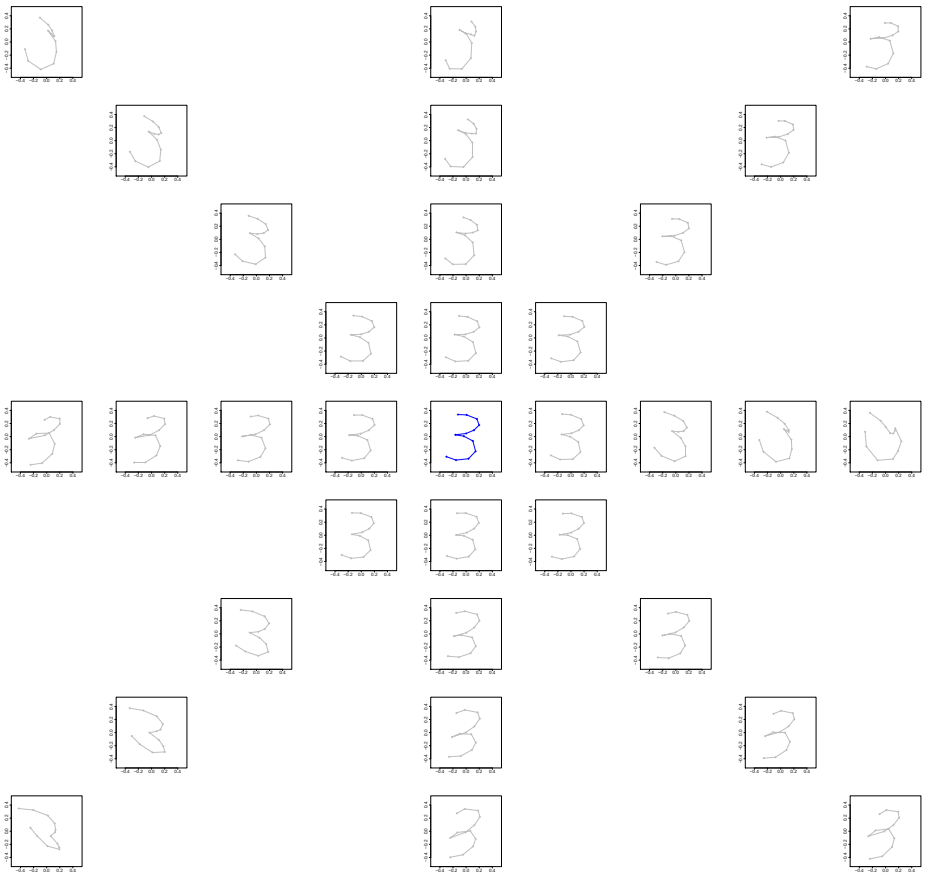


Figure 2: Principal sub-manifolds of the handwritten digits data, started from the center of symmetry. Among all the figures: the central figure (in blue) is the center of symmetry; the horizontal row contains images recovered from the first principal direction of the sub-manifold; the vertical column is the second principal direction; the main diagonal is the third principal direction; the other diagonal is the fourth principal direction.

S5 Angle Between Linear Subspaces

Assume two k-dimensional linear subspaces $A, B \subset \mathbb{R}^d$. Let $v_j := \sum_\alpha A_{j\alpha} w_\alpha$ any normal vector in A, which means that $\|w\| = 1$. Then, the projection of this vector to B is given by $\tilde{v}_j = \sum_{\alpha,\beta,l} B_{j\beta} B_{l\beta} A_{l\alpha} w_\alpha$. Let $\phi(v, \tilde{v})$ denote the angle between the two vectors, then $\|\tilde{v}\| = \cos\phi(v, \tilde{v})$ and therefore

$$\cos^2\phi(A,B) = \min_v \cos^2\phi(v,\tilde{v}) = \min_v \sum_j v_j \cdot \tilde{v}_j = \min_w \sum_{\alpha,\beta,\gamma,j,l} w_\gamma A_{j\gamma} B_{j\beta} B_{l\beta} A_{l\alpha} w_\alpha \,,$$

where minimization runs over unit vectors. Since the matrix $\left(\sum_{\beta,j,l} A_{j\gamma} B_{j\beta} B_{l\beta} A_{l\alpha}\right)_{\gamma\alpha}$ is symmetric and positive semidefinite, the expression is minimized for w being the eigenvector to the minimal eigenvalue $\lambda_{\min}$ of this matrix. We therefore note

$$\cos^2\phi(A,B) = \lambda_{\min}\left(\sum_{\beta,j,l} A_{j\gamma} B_{j\beta} B_{l\beta} A_{l\alpha}\right)_{\gamma\alpha} \quad \Rightarrow \quad \cos\phi(A,B) = \lambda_{\min}\left(\sum_l B_{l\beta} A_{l\alpha}\right)_{\beta\alpha} .$$

In the latter expression, we assume all eigenvalues of $\left(\sum_l B_{l\beta} A_{l\alpha}\right)_{\beta\alpha}$ to be positive, which can be achieved by suitably choosing the signs of spanning vectors of A and B.

Analogously, note that the angle between a 1-dimensional linear subspace A spanned by the unit vector v and a k-dimensional linear subspace

B is given by

$$\cos^2\phi(A,B) = \cos^2\phi(v,\tilde{v}) = \sum_j v_j \cdot \tilde{v}_j = \sum_{\alpha,j,l} v_j B_{j\alpha} B_{l\alpha} v_l \,.$$

S6 A Lagrangian Formulation

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.

We are aiming to describe a principal sub-manifold in terms of a Lagrangian problem, analogous to Panaretos et al. (2014). To make the idea precise, we propose the following definition.

Definition 1. Assume a sub-manifold, described by the image of an injective smooth function

$$N : \mathbb{R}^k \supset U \to \mathcal{N} \subset \mathcal{M} \subset \mathbb{R}^d \,. \tag{S6.5}$$

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

We denote a generic point in the principal sub-manifold by N, suppressing the argument t, and write $\dot{N}_{j\alpha} := \nabla_\alpha N_j$ in a slight abuse of notation.

The main term of the Lagrangian characterizes the compatibility of the principal sub-manifold with the tensor field W. A simple measure for goodness of fit at some point $N \in \mathcal{N}$ is given by the angle between $T_N\mathcal{N}$ and $W(N)$, which we will denote by $\phi(T_N\mathcal{N}, W(N))$. Furthermore, the Lagrangian will contain two sets of constraints. One set will enforce that the tangent vectors of $T\mathcal{N}$ will always be orthonormal and the second set consists of the algebraic equations $F(N) = 0$, restricting to $\mathcal{M}$.

Using the result of Supplement S3, we can formulate the Lagrangian for the principal sub-manifold as

$$\mathscr{L}_1(N, \dot{N}) := \lambda_{\min}\left(\sum_l W_{l\beta}(N)\dot{N}_{l\alpha}\right)_{\beta\alpha} + \sum_{\alpha\beta}\kappa_{\alpha\beta}\left(\sum_l \dot{N}_{l\alpha}\dot{N}_{l\beta} - \delta_{\alpha\beta}\right) + \sum_\nu z_\nu F_\nu(N)\,. \tag{S6.6}$$

For the solution of the dynamic system resulting from this Lagrangian we have the following result.

Theorem 1. *Assume that $\mathcal{M} = \mathbb{R}^d$, which means that $F \equiv 0$, and let $h = \infty$. Assume that the first $k+1$ eigenvalues of $\Sigma_{x,\mathcal{M}}$ are distinct for any point $x \in \mathcal{M}$. Then the solution of the dynamic system corresponding to $\mathscr{L}_1$ with initial conditions $N(0) = A$ and $\forall\alpha : \dot{N}_\alpha(0) = e_\alpha(A, \mathbb{R}^d)$ is the affine space containing A and spanned by $\left\{e_1(A, \mathbb{R}^d), e_2(A, \mathbb{R}^d), \ldots, e_k(A, \mathbb{R}^d)\right\}$, the first k eigenvectors of $\Sigma_{A,\mathcal{M}}$.*

Proof. Since $h = \infty$, we have for any point $x \in \mathbb{R}^d$ that $W_{j\beta}(x) = \left(e_\beta(A, \mathbb{R}^d)\right)_j$. Assuming the constraint $\forall \alpha, \beta \,:\, \sum_l \dot{N}_{l\alpha}\dot{N}_{l\beta} = \delta_{\alpha\beta}$, it is clear that

$$\lambda_{\min}\left(\sum_l W_{l\beta}(N)\dot{N}_{l\alpha}\right)_{\beta\alpha} \leq 1$$

where equality holds if and only if $\forall \alpha, j \,:\, W_{j\alpha}(N) = \dot{N}_{j\alpha}$. Integrating the equation $\dot{N}_\alpha = e_\alpha(A, \mathbb{R}^d)$ with the initial conditions yields the desired affine subspace. □

However, many complications arise when trying to solve the dynamic system resulting from this Lagrangian. In particular, the solution strategy outlined in Panaretos et al. (2014) cannot be applied in this generalized setting. We will therefore turn to a simpler approach, which aims at constructing a principal sub-manifold in terms of a spherical mesh, centered at the reference point A. This means, that we are looking for curves γ starting at A, whose tangent vectors $\dot{\gamma}$ stay as close as possible to $W(\gamma)$.

Consider the following set of curves, which encodes the constraints

$$\Gamma_{1,\mathcal{M}} := \left\{ \gamma : [a, b] \to \mathbb{R}^d \,\middle|\, \sum_j \dot{\gamma}_j\dot{\gamma}_j = 1 \text{ and } F(\gamma) = 0 \right\}. \tag{S6.7}$$

We then maximize

$$\underset{\gamma\in\Gamma_{1,\mathcal{M}}}{\operatorname{argmax}} \sum_{\alpha,j,l} \dot{\gamma}_j W_{j\alpha}(\gamma)W_{l\alpha}(\gamma)\dot{\gamma}_l\,. \tag{S6.8}$$

We can write this in terms of a Lagrangian, where we square the main term of the Lagrangian to achieve a simpler formulation

$$\mathscr{L}_2(\gamma, \dot{\gamma}) := \sum_{\alpha,j,l} \dot{\gamma}_j W_{j\alpha}(\gamma) W_{l\alpha}(\gamma) \dot{\gamma}_l + \kappa \left(\sum_j \dot{\gamma}_j \dot{\gamma}_j - 1 \right) + \sum_\nu z_\nu F_\nu(\gamma) \,. \tag{S6.9}$$

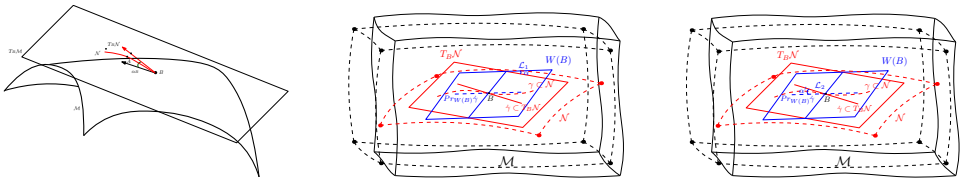
We get the following Theorem for this Lagrangian

Theorem 2. *Assume that $\mathcal{M} = \mathbb{R}^d$, which means that $F \equiv 0$, and let $h = \infty$. Assume that the first $k+1$ eigenvalues of $\Sigma_{x,\mathcal{M}}$ are distinct for any point $x \in \mathcal{M}$. Then the solutions γ_α of the dynamic system corresponding to $\mathscr{L}_2$ with initial conditions $\gamma(0) = A$ and $\dot{\gamma}(0) = e_\alpha(A, \mathbb{R}^d)$ span the affine space containing A and spanned by $\left\{e_1(A, \mathbb{R}^d), e_2(A, \mathbb{R}^d), \ldots, e_k(A, \mathbb{R}^d)\right\}$, the first k eigenvectors of $\Sigma_{A,\mathcal{M}}$.*

The Lagrangian $\mathscr{L}_2$ is much simpler than $\mathscr{L}_1$ and is in fact very similar to the principal flow Lagrangian. However, the solution technique used in Panaretos et al. (2014) is not suitable here, therefore we provide a simpler "greedy" algorithm to approximate principal submanifolds in Section 3.1 of the article.

For geometric intuition on the Lagrangians, denote the angle between the k-dimensional tangent space $\dot{N}$ from $\mathscr{L}_1$ and the tensor field $W(N)$ at point B as α_B and denote the angle between the 1-dimensional tangent

vector $\dot{\gamma}$ and $W(\gamma)$ at point B as α'_B and note that $\alpha'_B \leq \alpha_B$. Figure 3 illustrates the two angles and compares them to the situation of the principal flow.



(a) (b) (c)

Figure 3: Principal sub-manifolds. (a) Principal flow, α_B is the angle between $T_B\mathcal{N}$ $(k = 1)$ and $W(B)$ $(k = 1)$; (b) Principal sub-manifold (according to $\mathscr{L}_1$), α_B is the angle between $T_B\mathcal{N}$ $(k \geq 2)$ and $W(B)$ $(k \geq 2)$; (c) Principal sub-manifold (according to $\mathscr{L}_2$), α'_B is the angle between $\dot{\gamma}|_B \in T_B\mathcal{N}$ $(k = 1)$ and $W(B)$ $(k \geq 2)$.

We conclude this session with a remark on the interpretation of Theorem 1. Theorem 1 shows that, in a flat space, the principal sub-manifold reduces to the k-dimensional space spanned by the k eigenvectors of $\Sigma_{A,\mathcal{M}}$ when $h = \infty$. In this sense, the approach reduces to linear PCA. In connection with the principal flow, recall the similar result (Proposition 5.1, Panaretos et al. (2014)) where the first order of the principal flow on a flat space has been shown to coincide with the first principal direction if the locality parameter h of the tangent covariance matrix is chosen to be infinity.

S7 A Greedy Algorithm

Algorithm 1. *two-dimensional principal sub-manifold*

1. *At a point A (mean or other point), use the logarithm map:* $\mathbf{log}_A(x_i) = y_i$.

2. *Find the covariance matrix $\Sigma_{A,\mathcal{M}}$ from $y_1, \ldots, y_n$ by (2.9).*

3. *Let $e_1(A)$ and $e_2(A)$ be the first and second eigenvector of $\Sigma_{A,\mathcal{M}}$. Define*

$$Z_l = \epsilon' \times \Big[\cos\big(2l\pi/L\big)e_1(A) + \sin\big(2l\pi/L\big)e_2(A)\Big],$$

with $l = 1, \ldots, L$.

4. *Use exponential map to map Z_l onto $\mathcal{M}$ so we get a set of new points $\mathbf{exp}_A(Z_l) = A_l$.*

5. *Assume that we stay at point $A_{l,i}$, we are going to find $A_{l,i+1}$ ($A_{l,0} = A$ and $A_{l,1} = A_l$) via steps (a)-(g)*

   (a) find $\Sigma_{A_{l,i},\mathcal{M}}$.

   (b) find $e_1\big(A_{l,i}\big)$ and $e_2\big(A_{l,i}\big)$.

   (c) find $v_{l,i} = \mathbf{log}_{A_{l,i}}\big(A_{l,i-1}\big)$.

   (d) find

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

where $\langle a, b\rangle = \sum_{i=1}^{n} a_i b_i$ *with* $a = (a_1, \ldots, a_n)$ *and* $b = (b_1, \ldots, b_n)$.

(e) let $u_{l,i} = \tilde{v}_{l,i}$ *(or* $u_{l,i} = 2\tilde{v}_{l,i} - v_{l,i}$*).*

(f) calculate

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

(g) update

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

(h) stop at $A_{l,i+1}$ *when*

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

for all $j = 1, \ldots, n$.

6. *For every* $l = 1, \ldots, L$*, connect* $A_{l,i}$ *with* $A_{l,i+1}$ *by* i *we get* $\mathcal{A}_l$*, a ray of principal sub-manifold.*

7. **Output***: all* $\mathcal{A}_l$*'s as in (3.15), where* $1 \leq l \leq L$.

Remark 1. In Step 3, there is no difference in either forming a circle or an ellipse for small ϵ'. In case of an ellipse, the axes of ellipse would be proportional to the first and second eigenvalue of $\Sigma_{A,\mathcal{M}}$. In case of a k-dimensional sub-manifold, Step 3 and Step 5(b), (d) will need to be updated with the first k eigenvectors.

S8 Convergence of the Greedy Algorithm

First we simplify the algorithm from Appendix S7 to only represent abstract calculation steps. We call curves defined by this algorithm *principal spokes*.

Algorithm 2. *Algorithm for principal spokes of fixed length L.*

1. *Start with a point $p^{(0)} \in \mathcal{M}$ and a tangent vector $v^{(0)} \in W(p^{(0)})$,*

2. *for $i \geq 0$ and writing an orthonormal basis of $W(p^{(i+1)})$ by vectors $W_\alpha(p^{(i+1)})$ for $1 \leq \alpha \leq k$ let*

$$
\begin{aligned}
p^{(i+1)} &= \mathbf{exp}_{p^{(i)}}\left(\epsilon' v^{(i)}\right) \\
\widetilde{u}^{(i+1)} &= -\mathbf{log}_{p^{(i+1)}}(p^{(i)}) \\
u^{(i+1)} &= \sum_{\alpha=1}^{k}\sum_{j=1}^{m} \widetilde{u}_j^{(i+1)} W_{\alpha j}(p^{(i+1)}) W_\alpha(p^{(i+1)}) \\
v^{(i+1)} &= \frac{u^{(i+1)}}{\left\|u^{(i+1)}\right\|}
\end{aligned}
$$

3. *stop when $(j+1)\epsilon' \geq L$.*

Assume a sequence of points $(p_j)_j \in \mathcal{M}$ which converges to a point $p \in \mathcal{M}$. Then smoothness of $\mathcal{M}$ yields that $d_g(p_j, p) \to \|p_j - p\|$, where d_g is the geodesic distance. Thus proving any convergence statement in terms of euclidean distance immediately yields the same convergence statement in terns of geodesic distance in $\mathcal{M}$.

Lemma 1. *Assume W to be C^1 and assume a sequence of points $(p_j)_j \in \mathcal{M}$ which converges to a point $p \in \mathcal{M}$. For any vector $v \in W(p)$ let $v_j \in W(p_j)$ define the sequence of its projections. Then the angle between v_j and v decreases as $\mathcal{O}(\|p - p_j\|)$.*

Proof. The claim follows immediately by applying the Taylor expansion in $\|p - p_j\|$ to the local spanning vector fields X_j of W. The linear order dominates for $p_j \to p$. □

Lemma 2. *Assume W to be C^1 and assume a sequence of points $p^{(j)}$ constructed by algorithm 2 with fixed ϵ'. Then there is a constant K_0 such that*

$$\|v^{(j+1)} - v^{(j)}\| \leq K_0\epsilon' + o(\epsilon'). \tag{S8.10}$$

Proof. The points $p^{(j+1)}$ and $p^{(j)}$ are connected by an arc length parametrized geodesic γ, which is C^2, since $\mathcal{M}$ is C^2. Let $\gamma(0) = p^{(j)}$ and $\gamma(\epsilon') = p^{(j+1)}$ then $v^{(j)} = \dot\gamma(0)$ and $\widetilde{v}^{(j+1)} := \dot\gamma(\epsilon') = \frac{1}{\epsilon'}\widetilde{u}^{(j+1)}$ by construction. Since $\dot\gamma$ is C^1, we can use the Taylor expansion to note that there is a constant A_1 such that we have for the angle

$$\angle\left(\widetilde{v}^{(j+1)}, v^{(j)}\right) \leq A_1\epsilon' + o(\epsilon')$$

$$\angle\left(v^{(j+1)}, P_{W(p^{(j+1)})}v^{(j)}\right) = \angle\left(P_{W(p^{(j+1)})}\widetilde{v}^{(j+1)}, P_{W(p^{(j+1)})}v^{(j)}\right) \leq A_1\epsilon' + o(\epsilon')\,.$$

From Lemma 1 we conclude that there is a constant A_2 such that

$$\angle\left(P_{W(p^{(j+1)})}v^{(j)}, v^{(j)}\right) \leq A_2\epsilon' + o(\epsilon')$$

$$\angle\left(v^{(j+1)}, v^{(j)}\right) \leq \left(v^{(j+1)}, P_{W(p^{(j+1)})}v^{(j)}\right) + \left(P_{W(p^{(j+1)})}v^{(j)}, v^{(j)}\right) \leq (A_1 + A_2)\epsilon' + o(\epsilon')\,.$$

The claim follows immediately since $v^{(j)}$, $\widetilde{v}^{(j+1)}$ and $v^{(j+1)}$ are unit vectors.

□

Theorem 3. *Assume W to be C^2 and assume there is a C^2 integral submanifold $\mathcal{N}$ of W through $p^{(0)}$. From the fact that $\mathcal{M}$ and $\mathcal{N}$ are C^2 and from Lemma 1 we can conclude the following uniform bounds in a ball of radius L around $p^{(0)}$, where P_X denotes orthogonal projection onto X and v is always normalized:*

$$\forall\, p \in \mathcal{M},\, v \in T_p\mathcal{M} :\, \|\mathbf{exp}_p\big(\epsilon' v\big) - p - \epsilon' v\| \leq \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2), \tag{S8.11}$$

$$\forall\, p \in \mathcal{N},\, v \in T_p\mathcal{N} :\, \|P_{\mathcal{N}}(p + \epsilon' v) - p - \epsilon' v\| \leq \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2), \tag{S8.12}$$

$$\forall\, p, q \in \mathcal{M},\, v \in T_p\mathcal{N} :\, \|P_{\mathcal{W}_k(q)}v - v\| \leq K_1\|p - q\| + o(\|p - q\|). \tag{S8.13}$$

Then the curves of length L starting at $\gamma(0) = p^{(0)}$ constructed by algorithm 2 converge for step size $\epsilon' \to 0$ to curves that lie within the integral submanifold of W through $p^{(0)}$.

Proof. For every point $q \in \mathcal{N}$ we have $T_q\mathcal{N} = W(q)$. We show that the distance of $p^{(s)}$ from $\mathcal{N}$ is of order $\mathcal{O}\left((\epsilon')^2 \sum_{j=0}^{s-1}(1 + K_1\epsilon')^j\right)$. The proof is done by induction. Note that we will simply bound the distance between the points $p^{(j)}$ and some corresponding points $q^{(j)} \in \mathcal{N}$. Since these $q^{(j)}$ need not be the closest points in $\mathcal{N}$ to the $p^{(j)}$, we derive, strictly speaking, upper bounds for the distances of the $p^{(s)}$ from $\mathcal{N}$.

As above let $v^{(0)} \in W(p^{(0)}) = T_{p^{(0)}}\mathcal{N}$ the initial direction of the curve. Then, using that $v^{(0)}$ is tangent to $\mathcal{N}$ and $p^{(1)} = \mathbf{exp}_{p^{(0)}}\left(\epsilon' v^{(0)}\right)$, we get from equation (S8.11)

$$\|p^{(1)} - p^{(0)} - \epsilon' v^{(0)}\| = \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2).$$

Let $q^{(1)}$ be the orthogonal projection of $p^{(0)} + \epsilon' v^{(0)}$ to $\mathcal{N}$, which is unique for small enough ϵ'. Then by equation (S8.12) we get $\|q^{(1)} - p^{(0)} - \epsilon' v^{(0)}\| = \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2)$ and thus $\|q^{(1)} - p^{(1)}\| \leq K_2(\epsilon')^2 + o((\epsilon')^2)$. This concludes the beginning of the induction.

Now assume $\|q^{(s)} - p^{(s)}\| \leq K_2(\epsilon')^2 \sum_{j=0}^{s-1}(1+K_1\epsilon')^j + o\left((\epsilon')^2 \sum_{j=0}^{s-1}(1 + K_1\epsilon')^j\right)$ where $q^{(s)}$ is some point on $\mathcal{N}$. As before we have

$$\|p^{(s+1)} - p^{(s)} - \epsilon' v^{(s)}\| = \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2).$$

Let $w^{(s)}$ be the projection of $v^{(s)}$ to $T_{q^{(s)}}\mathcal{N}$, then equation (S8.13) yields

$$\epsilon'\|w^{(s)} - v^{(s)}\| \leq K_1\epsilon'\|q^{(s)} - p^{(s)}\| + o(\epsilon'\|q^{(s)} - p^{(s)}\|).$$

Let $q^{(s+1)}$ be the orthogonal projection of $q^{(s)} + \epsilon' w^{(s)}$ to $\mathcal{N}$. As above

$$\|q^{(s+1)} - q^{(s)} - \epsilon' w^{(s)}\| = \frac{K_2}{2}(\epsilon')^2 + o((\epsilon')^2).$$

From the above considerations we can thus conclude

$$\begin{aligned}
\|p^{(s+1)} - q^{(s+1)}\| &\leq \|p^{(s)} + \epsilon' v^{(s)} - q^{(s)} - \epsilon' w^{(s)}\| + K_2(\epsilon')^2 + o((\epsilon')^2) \\
&\leq \|q^{(s)} - p^{(s)}\| + K_1\epsilon'\|q^{(s)} - p^{(s)}\| + K_2(\epsilon')^2 + o(\epsilon'\|q^{(s)} - p^{(s)}\|) + o((\epsilon')^2) \\
&\leq K_2(\epsilon')^2 \sum_{j=1}^{s}(1 + K_1\epsilon')^j + K_2(\epsilon')^2 + o\left((\epsilon')^2 \sum_{j=1}^{s}(1 + K_1\epsilon')^j\right) + o((\epsilon')^2 \\
&= K_2(\epsilon')^2 \sum_{j=0}^{s}(1 + K_1\epsilon')^j + o\left((\epsilon')^2 \sum_{j=0}^{s}(1 + K_1\epsilon')^j\right).
\end{aligned}$$

This concludes the induction step.

Now, note that

$$\begin{aligned}
\sum_{j=0}^{s-1}(1 + K_1\epsilon')^j &= \sum_{j=0}^{s-1}\sum_{l=0}^{j}\binom{j}{l}(K_1\epsilon')^l = \sum_{l=0}^{s-1}(K_1\epsilon')^l \sum_{j=l}^{s-1}\binom{j}{l} = \sum_{l=0}^{s-1}\binom{s}{l+1}(K_1\epsilon')^l \\
&\leq \sum_{l=0}^{s-1}\frac{s^{l+1}}{(l+1)!}(K_1\epsilon')^l = \frac{1}{K_1\epsilon'}\sum_{l=1}^{s}\frac{(K_1 s\epsilon')^l}{l!} \leq \frac{\exp(K_1 s\epsilon') - 1}{K_1\epsilon'},
\end{aligned}$$

where we use the hockey-stick identity in the first line.

To achieve a curve of length at least L, we need $s = \lceil \frac{L}{\epsilon'} \rceil$ steps of length ϵ'. The maximum distance of the curve from $\mathcal{N}$ is thus bounded by $d_{\max} = \frac{K_2(\exp(K_1 L)-1)}{K_1}\epsilon' + o(\epsilon')$ which goes to 0 as $\epsilon' \to 0$. □

Remark 2. Note that, by definition of an integral manifold, $\dot{N}_{j\alpha} = W_{j\alpha}$ at every point in $\mathcal{N}$ and therefore $\mathcal{N}$ can be understood as a principal submanifold in our setting.

Corollary 1. *Assume W to be C^2 and involutive, then there is a C^2 integral submanifold $\mathcal{N}$ of W through $p^{(0)}$. Then the curves of length L starting at $\gamma(0) = p^{(0)}$ constructed by algorithm 2 converge for step size $\epsilon' \to 0$ to curves that lie within the integral submanifold of W through $p^{(0)}$.*

S9 Reflection versus Projection

To illustrate the properties of the projection and reflection algorithms we consider the simpler case of the principal flow to make illustration easier. However, the qualitative results can be generalized to the principal submanifold setting. In general, the projection algorithm will incur an error proportional to the local curvature of true integral curves of the vector field W at each step. If the vector field changes strongly, this leads to an increasingly bad fit. The reflection algorithm does not incur an error when following a field whose integral curves have constant curvature, which means they are circles. In the general case, it will incur an error at each step which is proportional to the change of curvature. In this sense, the error is of "higher order" and can be expected to be smaller in general.

Figure 4 illustrates three example cases: spherical, elliptical and sinus integral curves. To highlight the qualitative behaviour, the step sizes of the algorithms are strongly exaggerated. One can clearly see that the projec-

tion algorithm determines a curve which successively departs from the true integral curve. The reflection algorithm, in contrast, stays close to the true curve. It is thus also much less susceptible to perturbations in the vector field W.

Figure 4 also shows that the integral curves for the reflection algorithm are much smoother, if only every second step is taken into account. This is due to the fact that the curve starts out tangential to the true integral curve, changes direction after the first step and is then again close to tangential after the second step, if curvature is only slowly varying. Therefore, the change of direction after the second step and indeed every even numbered step is much smaller than after an odd numbered step, if curvature changes slowly.

S10 Simulated data

To further investigate the behavior of the principal sub-manifold as dependent on the configuration of the data points and the choice of scale parameter, we considered three sets of examples on S^3. We chose this manifold as a "test manifold" since it represents one of the most natural spaces from which the projected sub-manifold can be well understood, and since it provides a manifold for which we can compare the principal sub-manifolds

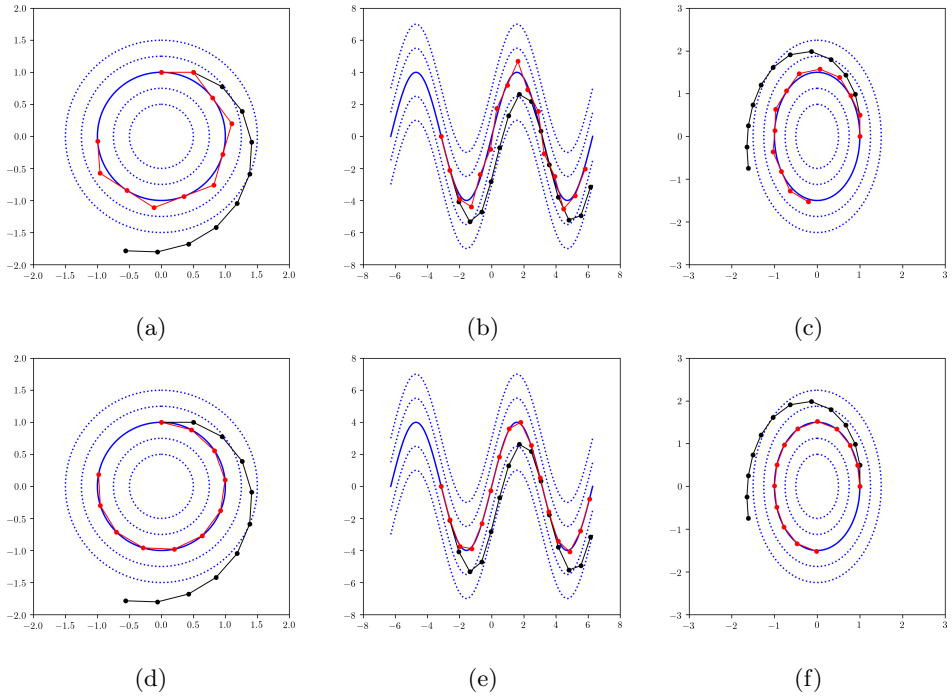


Figure 4: Comparing the results of the projection and reflection algorithms for a simple principal flow. The blue lines indicate true integral curves of the principal flow, thus the local PCA vector field W is always tangential to these lines. The solid blue line is the curve on which the algorithm starts and which it should ideally follow. The black points and lines represent steps of the algorithm with projection and the red points and lines represent steps of the algorithm with reflection. In Figures (a)-(c) all steps of the reflection algorithm are shown, whereas in Figures (d)-(f) the reflection algorithm uses halved step size and only every second point is used.

with the principal geodesic. We observe here that the full manifold variation of the sub-manifold from the data can be very complicated; hence, we do not look at them on a quantitative basis, but rather investigate them

qualitatively.

The first set of examples involves five data clouds in S^3, each presenting a different curvature. As the curvature is non-constant, the Fréchet mean is no longer a good starting point for the principal sub-manifold. Instead, we choose the center of symmetry for each data set as a starting point. The first and second data cloud are constructed in a way that the first three coordinates of each point are concentrated around a one dimensional curve; the configuration of the third and fourth are such that the points are on a two-dimensional surface/plane; the fifth one is much more diffuse: the points lie on a sea-wave-like surface.

For each one of them, a two-dimensional principal sub-manifold was fitted using three different bandwidths h and the results are presented. The results indicate that the corresponding sub-manifolds perform well in capturing the local and global variation. We note that the sub-manifold fits well for data Cloud 1 no matter what scale of h is used (Figure 6(a)-(c)); the sub-manifold seems to capture a finer structure with a reduced value for h for data Cloud 2 (Figure 6(d)-(f)): this can be also seen as the first principal direction evolves with the scale of h. When the surface becomes two-dimensional for data Clouds 3 and 4, the principal sub-manifolds also excel: the fitted sub-manifold remains unchanged for different h as the

surface is flat (Figure 6(g)-(i)), while it picks up the appropriate structure with a reduced h for the bent surface (Figure 6(j)-(l)); for data Cloud 5 (Figure 6(m)-(o)), it is more obvious that using a sub-manifold is more appealing than using only a curve or its equivalent; the sub-manifold fits the data points surprisingly well even with a surface of high curvature.

To probe how a sub-manifold performs with a noisy surface, we created four sets of data by blurring the sea-wave-like surface aforementioned with increasing levels of noise. Although the data reside in S^3, most of the variation originates around but not exactly on the surface. By knowing how the data points lie around the surface, we can get a sense of such variability. As we should no longer look at the local scale when the points tend to have large variability, we found a two-dimensional sub-manifold by choosing an appropriate scale parameter h, potentially a larger one, for each data set. In Figure 6(a), when there is no noise, it is expected that the sub-manifold would capture total variation of the data in the projected space. When the noise increases (Figure 6(b), (c), and (d)), where all points are more diffused away from the underlying projected surface, the fitted sub-manifold is, although not a perfect sub-manifold, still well explaining for total data variability.

The last sets of examples are from a "lifted" ellipsoid in S^3. Intuitively,

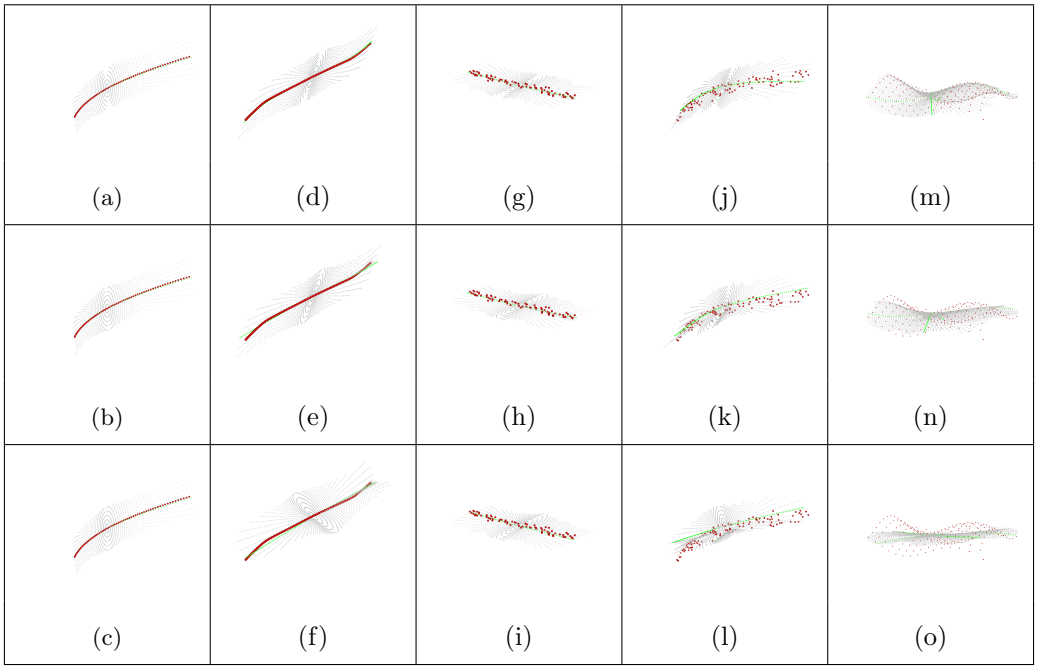
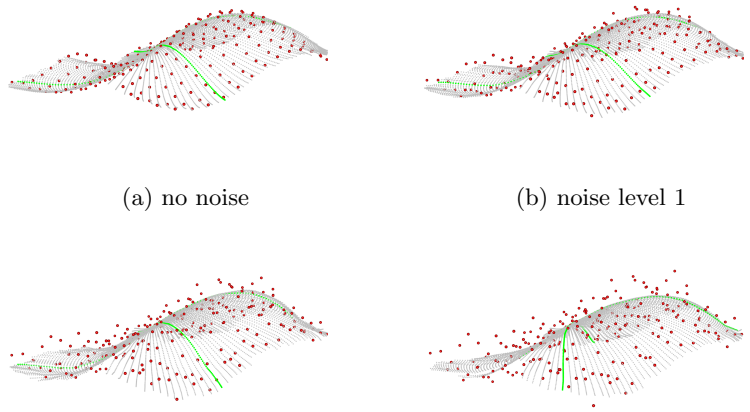


Figure 5: Principal sub-manifolds (with superimposed principal directions) for five data clouds in S^3, with different scale parameters. (a)-(c) Principal sub-manifolds (in gray) and principal directions (in green) for data Cloud 1 (in red) for different values of h (small, middle, large). (d)-(f), (g)-(i), (j)-(l) and (m)-(o) provide the same information for data Clouds 2, 3, 4 and 5.



(a) no noise (b) noise level 1

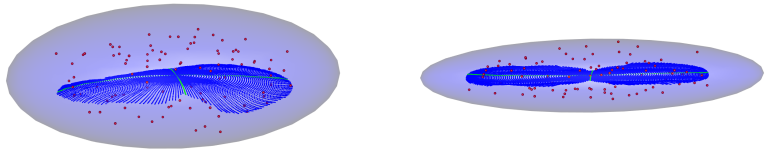
(c) noise level 2 (d) noise level 3

Figure 6: Principal sub-manifolds (with superimposed principal directions) for four sea wave sets of data with noise on S^3. (a) Principal sub-manifolds with no noise added. (b), (c) and (d) provide the same information for three different levels of noise.

the four data sets we generated represent different but inter-connected types of situation: (1) the triplets are well spread out inside the ellipsoid; (2)-(3) the triplets are mostly being concentrated in the middle of a more flatter ellipsoid; (4) the triplets are chosen nearly on the diameter of the ellipsoid (potentially around an ellipse). For case (1) (Figure 7(a)), where most points are inside the ellipsoid, neither one-dimensional nor two-dimensional sub-manifold would be a perfect sub-manifold. As the diffusion decreases, such as in case (2) (Figure 7(b)) and (3) (Figure 7(c)), the sub-manifold of dimension two appears to be more and more appropriate. In case (4) (Figure

7(d)), the sub-manifold provides the best fit such that all the projected data points lie on the sub-manifold. As one has already observed, the benefit of using a two-dimensional sub-manifold in this example is only marginal. Arguably though, one can go further, for instance, having a higher dimensional sub-manifold in case (1) or case (2). Such an extension of the algorithm would be very natural, but the details of implementing the algorithm are quite subtle and we choose not to pursue this further.

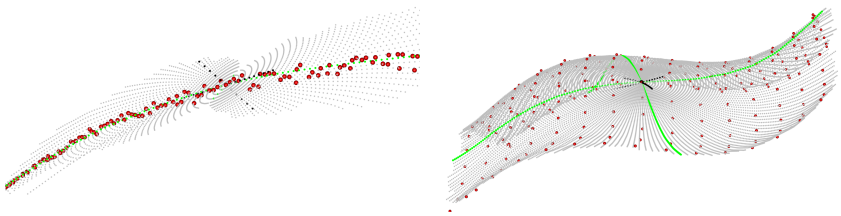
To contrast the principal sub-manifold with the standard principal geodesic, we include the results of principal geodesics adjusted to its 2d version, for the case of Figure 6(j) and Figure 6(m). Specifically, the best h has been chosen for either method to perform appropriately. It is expected that the principal geodesic, essentially a principal great circle along its first and second principal component, is not capable of capturing the curvature of the manifold; that is, the two principal geodesics (in black) for both cases (Figure 8(a)) and (Figure 8(b)) tend to deviate from the principal directions (in green) shortly after the starting point, thus not lying on the surface. In contrast, the principal sub-manifold handles the curvature well in both cases.



(a) $\frac{x_1^2}{2.5^2} + \frac{x_2^2}{2} + \frac{x_3^2}{1} \leq 1$ (b) $\frac{x_1^2}{5^2} + \frac{x_2^2}{2} + \frac{x_3^2}{1} \leq 1$

(c) $\frac{x_1^2}{10^2} + \frac{x_2^2}{2} + \frac{x_3^2}{1} \leq 1$ (d) $\frac{x_1^2}{20^2} + \frac{x_2^2}{2} + \frac{x_3^2}{1} \leq 1$

Figure 7: Principal sub-manifolds (with superimposed principal directions) for four ellipsoid sets of data on S^3. (a) Principal sub-manifolds (in blue) and principal directions (in green) for data set (in red) of case (1). (b), (c) and (d) provide the same information for case (2), (3) and (4).



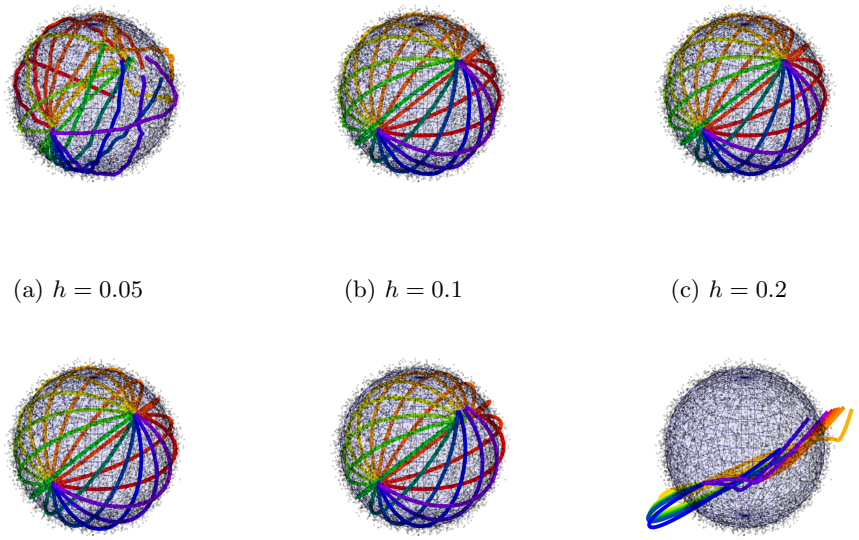
(a) (b)

Figure 8: Comparison of principal sub-manifolds and principal geodesics. The principal geodesics (in black) are superimposed to the principal directions (in green) in the projected space. Only segments of the principal geodesics are highlighted for visualization purpose.

S11 Simulations for Different Kernel Bandwidths

In this subsection, we present some simulation results highlighting the influence of the kernel bandwidth on the results of the principal submanifold estimation. We sample n points uniformly for the surface of a sphere of unit radius $S^2 \subset \mathbb{R}^3$ and add Gaussian white noise with a standard deviation of σ. Then we apply the principal submanifold algorithm with 16 initial directions with a Gaussian kernel of bandwidth h.

From Figures 9, 10, and 11, one can clearly see that the principal submanifold algorithm requires a bandwidth which is more than a factor of 2 larger than the noise level of the data. For lower bandwidth, the eigenvectors of local PCA become too variable and the two largest eigenvalues may

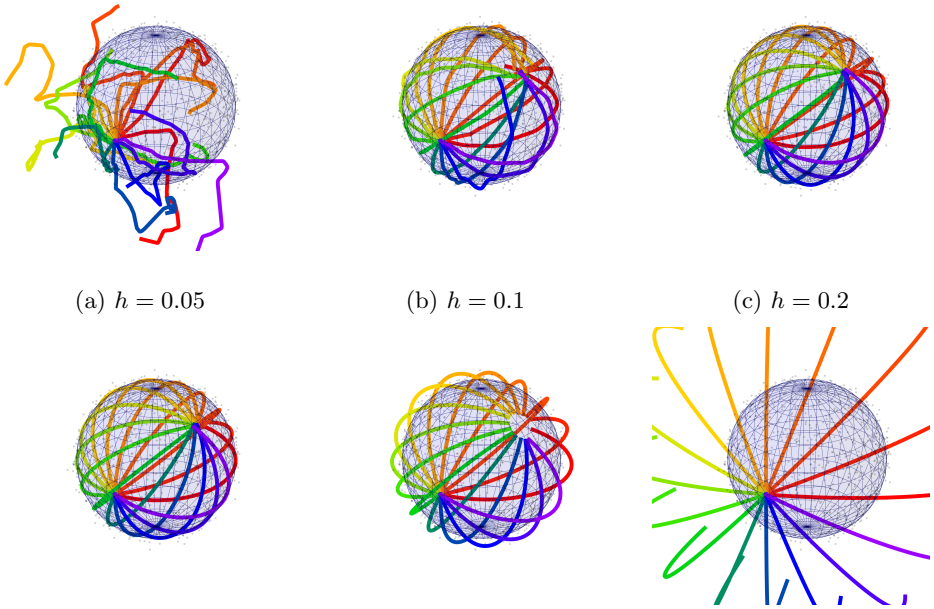


(a) $h = 0.05$ (b) $h = 0.1$ (c) $h = 0.2$

(d) $h = 0.5$ (e) $h = 1$ (f) $h = 2$

Figure 9: Principal submanifold results for noise level $\sigma = 0.05$ and sample size $n = 10\,000$. One can clearly see that a bandwidth much larger than the noise level but smaller than the diameter of the sphere, i.e. $\sigma \ll h \ll 2$ one gets a good fit to the sphere. For small bandwidths the eigenvectors to the two largest eigenvalues are in some cases not tangential to the sphere, leading to erratic direction changes of the rays. For large bandwidths, the opposite side of the sphere starts to contribute substantially to the local PCA, leading to curves that do not follow the curvature of the sphere, but are instead too straight.

not correspond to vectors whose span is close to tangential to the underlying sphere. For large bandwidths, which are less than a factor of 2 below the diameter of the sphere, the whole data set contributes to the local PCA,



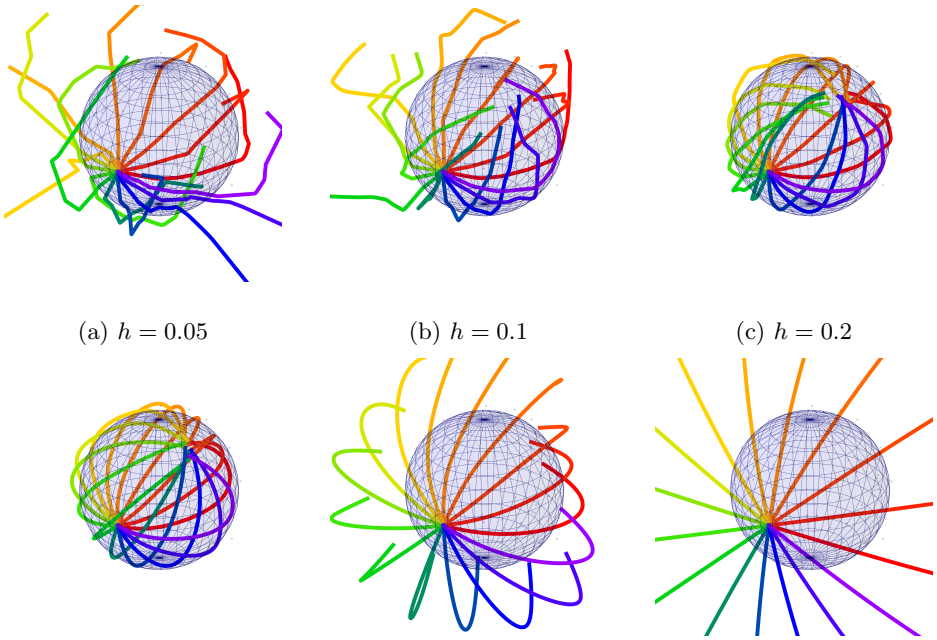
(a) $h = 0.05$ (b) $h = 0.1$ (c) $h = 0.2$

(d) $h = 0.5$ (e) $h = 1$ (f) $h = 2$

Figure 10: Principal submanifold results for noise level $\sigma = 0.05$ and sample size $n = 1\,000$. In comparison to the simulation with $n = 10\,000$, the results are more bandwidth dependent but intermediate bandwidths $h = 0.2$ and $h = 0.5$ still yield excellent results.

leading to too weak dependence of eigenvectors on the neighborhood of a point and thus to too slow variation of the eigenspaces. As a result, the lines curve much less than the sphere does and therefore progressively move away from the data. These two effects are to be expected and constitute fundamental limitations of any local neighborhood approach.

By comparing Figures 9, 10, and 11, one can see that the bandwidth dependence is exacerbated by reducing the sample size. However, even for



(a) $h = 0.05$ (b) $h = 0.1$ (c) $h = 0.2$

(d) $h = 0.5$ (e) $h = 1$ (f) $h = 2$

Figure 11: Principal submanifold results for noise level $\sigma = 0.05$ and sample size $n = 100$. Even in these extremely sparse data sets, intermediate bandwidths $h = 0.2$ and $h = 0.5$ still yield quite good results and recover the spherical shape.

$n = 100$ the intermediate bandwidths $h = 0.2$ and $h = 0.5$ still achieve remarkably good results. This underscores the potential strength of principal submanifolds as a means of identifying low dimensional structure in data sets even of moderate size.

S12 Principal variation on the MNIST data set

This section presents a detailed exploration of the principal sub-manifolds within the MNIST handwritten digit dataset, a comprehensive repository of handwritten digits widely utilized for the development and testing of various image processing algorithms. The dataset, accessible at `https://yann.lecun.com/exdb/mnist/`, encompasses a collection of 70,000 grayscale images, each of 28×28 pixel resolution, depicting digits from "0" to "9".

In this analysis, we specifically focus on the digit "3" to demonstrate the variability inherent in handwriting styles. This choice is motivated by the digit's capacity to exhibit a wide range of variations, such as differences in inclination angles, stroke thickness, opening sizes, and the curvature of junctions. Through the analysis of 7,141 instances of the digit "3" from the MNIST dataset, we construct the principal sub-manifold centered around the digit's mean and examine it across four principal directions, see Figure 12. This approach allows us to systematically describe the morphological changes of the digit "3" across distinct dimensions of variation.

The first principal direction reveals a significant variation in the inclination of the digit "3", transitioning from a leftward lean at the top to a rightward orientation. The second principal direction highlights a transformation in the corners of the "3", ranging from smooth to pronouncedly

Figure 12: Principal sub-manifolds of the digit "3" extracted from the MNIST dataset, originating from the dataset's mean. Each row illustrates the variation of the principal sub-manifolds across distinct principal directions, sequentially ordered from the first to the fourth principal direction, from the top row to the bottom. The digit displayed within a red frame at the center of each row denotes the mean.

sharp angles. In the third dimension, we observe the two semi-circular arcs of the "3" evolving from fuller to more slender forms, indicating a variation in the digit's overall robustness. Finally, the fourth direction underscores changes in stroke thickness, illustrating a spectrum from markedly thick to fine lines.

This multifaceted analysis not only underscores the diversity of handwriting styles captured within the MNIST dataset but also demonstrates

the utility of principal sub-manifold analysis in uncovering the underlying patterns of variation in digital representations of handwritten digits. Such insights are invaluable for enhancing the accuracy and robustness of machine learning models in tasks related to handwriting recognition and image processing.

S13 Principal variation of leaf growth

We also considered a landmark data set consisting of leaf growth, collected from three Clones and a reference tree of young black Canadian poplars at an experimental site at the University of Göttingen (`http://stochastik.math.uni-goettingen.de/~huckeman/ishapes_1.0.1.tar.gz`). The landmark configurations of the leaves were collected from three Clones ('C1', 'C2', 'C3') and a reference tree ('r') collected at two different levels: breast height (Level 1) and the crown (Level 2). They consist of the shapes of 27 leaves (nine from Level 1 and eighteen from Level 2) from Clone 1; of 22 leaves (six from Level 1 and sixteen from Level 2) from Clone 2; and of 24 leaves (eighteen from Level 1 and seventeen from Level 2) from Clone 3 as well as of the shapes of 21 leaves (thirteen from Level 1 and eighteen from Level 2) from the reference tree, all of which have been recorded non-destructively over several days during a major portion of their growing

period of approximately one month. There are four landmarks corresponding to quadrangular configuration at petiole, tip, and largest extensions orthogonal to the connecting line. Figure 13 represents the four landmarks extracted from the contour image of each leaf on a flat plane, the four landmarks contain, in particular, the information of length, width, vertical and horizontal asymmetry.

Although it is known that the leaf growth of the genetically identical trees along a period of time reveals a non-Euclidean pattern Huckemann (2011), the study only focused on the mean geodesic difference (therefore essentially a one-dimensional variation), which is used for the discriminant analysis across the trees. However, the shape change along different directions—especially the principal directions in shape space—has not been fully explored. We will investigate the shape variation using principal sub-manifold among three Clones and the reference tree. As can be expected (see in Section 2.2 in the paper), each landmark configuration, represented by a polygon in Figure 13, corresponds to a point in Kendall shape space. We focus on the non-geodesic shape variation primarily in vertical and horizontal direction of the leaf growth, the analysis of which requires a multi-dimensional scale treatment.

As all the leaves are very young, we first combine the leaves from the

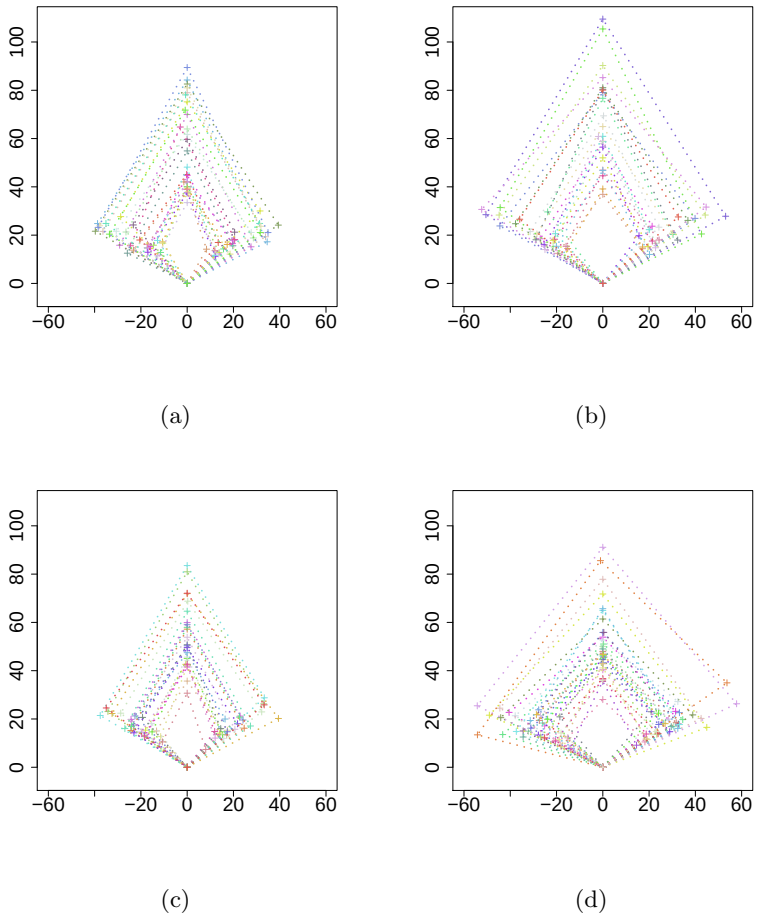
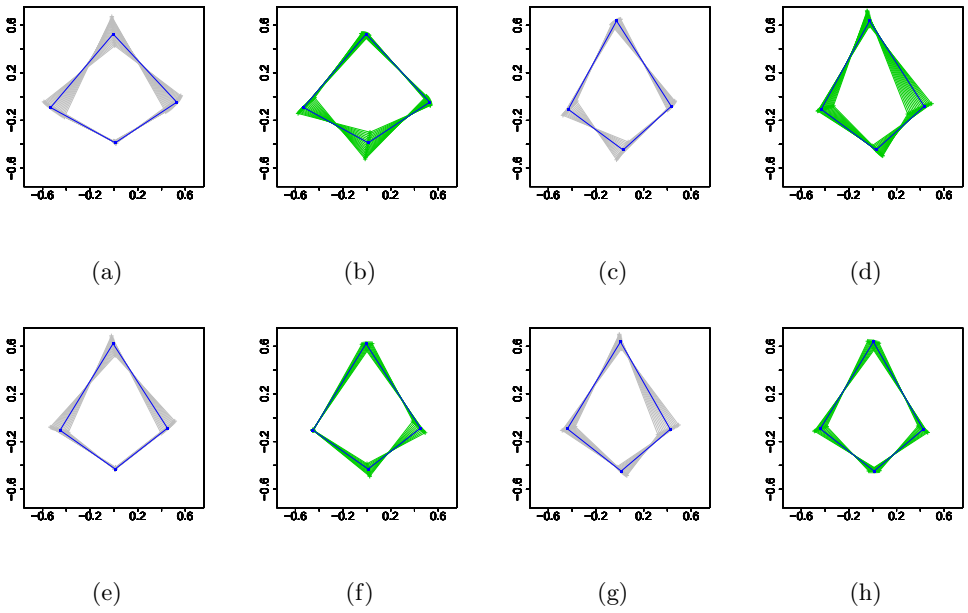


Figure 13: Leaf growth over a growing period of Clone 1 (a), Clone 2 (b), Clone 3 (c), and a reference tree (d). (a) Four landmarks on the leaf of Clone 1 have been connected and represented by a polygon at each growing period (27 polygons totally); (b)-(d) provide the same information for Clone 2 (22 polygons), Clone 3 (24 polygons) and the reference tree (31 polygons).

breast height and crown for each tree. For each tree, a principal submanifold is found, where two principal directions are extracted from the



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

(e) (f) (g) (h)

Figure 14: Principal sub-manifolds of the leaf growth data. (a) First principal direction obtained from the combined leaves at breast height and the crown of the reference tree; (b) Second principal direction obtained from the combined leaves at breast height and the crown of the reference tree. (c)-(h) provide the same information for Clone 1, 2 and 3.

fitted sub-manifold. The two principal directions are then transformed to the preshape space and all the landmarks recovered are superimposed. Results for all the three Clones and the reference tree are displayed in Figure 14. The leaves of the reference tree exhibit two main kinds of variation: the first one tends to follow the horizontal direction with some effects along the vertical direction at tip. This can be well seen by the first principal direction

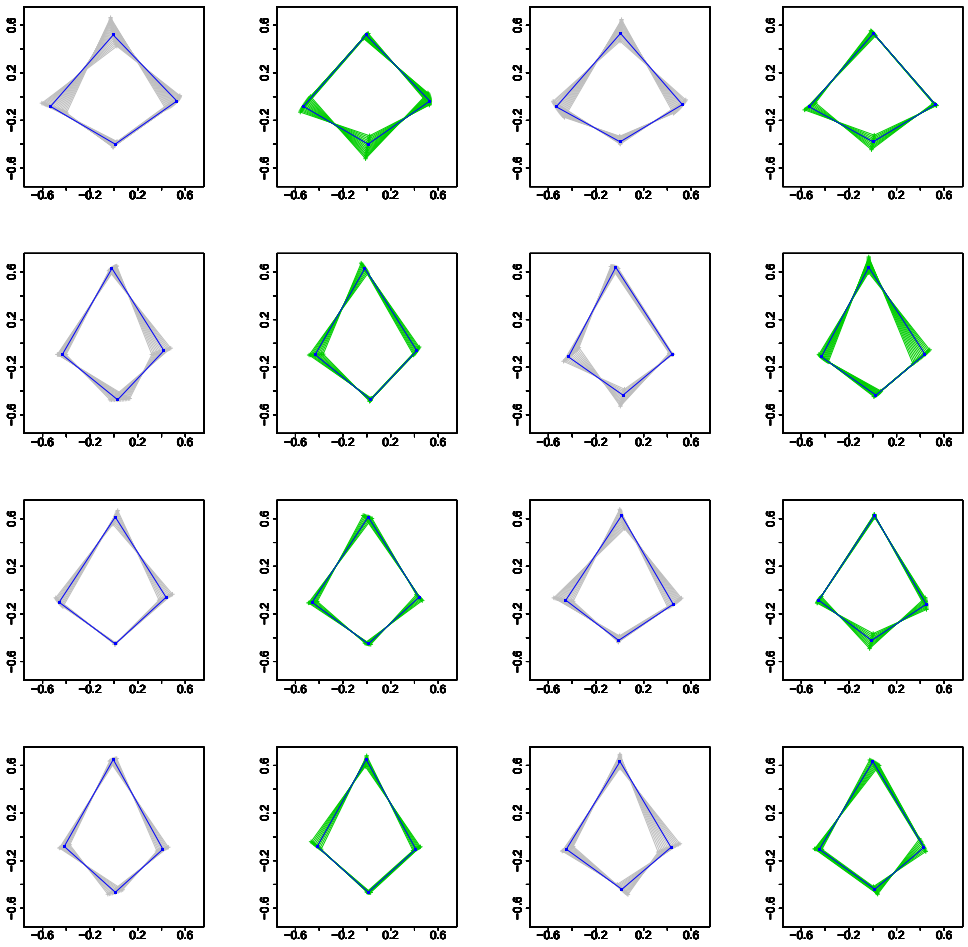


Figure 15: Principal sub-manifolds of the leaf growth data. Row 1 (reference tree): first principal direction at breast height; second principal direction at breast height; first principal direction at the crown; second principal direction at the crown. Row 2 - Row 4 provide the same information for Clone 1, 2 and 3.

in Figure 14(a); the second one concentrates on petiole, which is displayed by the second principal direction in Figure 14(b). The three Clones reveal

different patterns of variation from the reference tree; between them, each differs from the other. Clone 1 shows more variation at the petiole and the left extension in the first principal direction, while the second principal direction shows more variation at the right extension; the two principal directions of Clone 2 behave more similarly as that of the reference tree, with some other variation appearing in the second principal direction of Clone 2 at the right extension and the tip; unlike Clone 1 and 2, variation in both vertical and horizontal directions appear evenly in either the first or the second principal direction for Clone 3. The same analysis for the leaves at breast height and crown alone has also been performed separately with a similar outcome, as shown in Figure 15, the result suggesting no different conclusion.

S14 Introduction to landmark shape spaces

Here we introduce the notion of landmark shape spaces, which are used in one of the applications below. From the shape analysis point of view, landmark coordinates retain the geometry of a certain point configuration. The landmarks are observations, which are usually positions or correspondences on an object in an appropriate coordinate system. See, e.g., Dryden and Mardia (1998) for an accessible overview for a rapid introduction. Consider

a suitable ordered set of k landmarks of an object, namely a k-ad (where $k \geq 2$), with each landmark lying in $\mathbb{R}^{d'}$. That is,

$$z = \left\{ z^j \in \mathbb{R}^{d'} : 1 \leq j \leq k \right\},$$

To compare the shapes of objects described by k landmarks z_j, one can define the Kendall shape space $\Sigma_{d'}^k$ of configurations, which are invariant under translation, scaling, and rotation. This is achieved by transforming k-ads, $z = (z_1^T, \ldots, z_k^T)^T$, to points on the unit sphere:

Translation invariance: $z^* = ((z_1 - \bar{z})^T, \ldots, (z_k - \bar{z})^T)^T$, where $\bar{z} = \frac{1}{k} \sum_{j=1}^k z_j$

Scale invariance: $z_{\text{pre}} = \frac{z^*}{\|z^*\|}$

Rotation invariance: $[z] = Rz_{\text{pre}}$ for $R := \text{id}_k \otimes \widetilde{R}$ with the Kronecker product $\otimes$ and $\widetilde{R} \in SO(d')$. For $d' = 2$ this reduces to $[z] = R(\theta) z_{\text{pre}}$ or $[z] = e^{i\theta} z_{\text{pre}}$ if $\mathbb{R}^2$ is identified with $\mathbb{C}$, where $-\pi < \theta \leq \pi$.

Remark 3. Kendall shape space only leads to a manifold if $d' = 2$, therefore we restrict to $d' = 2$ here: the translation and scale invariant $z_{\text{pre}} \in S^{2k-3} \subset \mathbb{R}^{2k-2}$ is called the *preshape* of z. Centering the data to achieve translation invariance reduces dimension by 2 and projecting to the unit sphere $S^{2k-3} = \left\{ v \in \mathbb{R}^{2k-2} : \|v\| = 1 \right\}$ to achieve scale invariance reduces dimension by 1. Then, $[z]$ is the *shape* of z given by the *orbit* of the preshape z_{pre} under

rotation. Σ_2^k is a quotient space of S^{2k-3} with dimension $2k-4$ of equivalence classes of k-ads.

Bibliography

Dryden, I. L. and K. V. Mardia (1998). *Statistical Shape Analysis*. New York: Wiley.

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

Huckemann, S. (2011). Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics 39*, 1098–1124.

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