Adv. Appl. Prob. (SGSA) **38,** 299–319 (2006)
Printed in Northern Ireland

PRINCIPAL COMPONENT ANALYSIS FOR RIEMANNIAN MANIFOLDS, WITH AN APPLICATION TO TRIANGULAR SHAPE SPACES

STEPHAN HUCKEMANN * AND
HERBERT ZIEZOLD,* ** *University of Kassel*

Abstract

Classical principal component analysis on manifolds, for example on Kendall's shape spaces, is carried out in the tangent space of a Euclidean mean equipped with a Euclidean metric. We propose a method of principal component analysis for Riemannian manifolds based on geodesics of the intrinsic metric, and provide a numerical implementation in the case of spheres. This method allows us, for example, to compare principal component geodesics of different data samples. In order to determine principal component geodesics, we show that in general, owing to curvature, the principal component geodesics do not pass through the intrinsic mean. As a consequence, means other than the intrinsic mean are considered, allowing for several choices of definition of geodesic variance. In conclusion we apply our method to the space of planar triangular shapes and compare our findings with those of standard Euclidean principal component analysis.

Keywords: Shape analysis; principal component analysis; Riemannian manifold; geodesic

2000 Mathematics Subject Classification: Primary 60D05
Secondary 62H11; 53C22

1. Introduction

Means and principal component analysis (PCA) play an important role in statistics. In shape analysis means and principal components (PCs) are sought on a shape space which can be viewed as a *preshape* sphere modulo a compact group action, i.e. a Riemannian manifold (apart from possible singularities) with nonzero curvature; see e.g. [10, pp. 149ff.], [17], [20, pp. 69ff.], [21], and [22]. Presently, in order to perform PCA, a *Fréchet mean* on the preshape sphere is computed with respect to the Euclidean metric when embedding the sphere in Euclidean space. In the tangent space of that mean, standard PCA is employed, again based on the Euclidean metric; see e.g. [3] and [7]. We will give a brief review of this in Section 6.2.

Based on older works treating intrinsic means on arbitrary Riemannian manifolds in the sense of *centres of mass* (see, e.g. [9] and [13, pp. 108ff.]), more recently an algorithm and convergence bounds for computing such means have been established on general Riemannian manifolds with applications to shape spaces [15]; also, see [6] for intrinsic means of Lie groups.

In this paper we propose a method of PCA based on the intrinsic metric. This will be done in the following section. One might think (following, e.g. [6]) that an intrinsic mean would qualify as an offset of a geodesic best approximating a given data set, in the sense of least squared distances, as is the case in a Euclidean setting. Intuitively, however, owing to

Received 27 October 2004; revision received 15 December 2005.
* Postal address: Department of Mathematics, University of Kassel, D-34109 Kassel, Germany.
** Email address: ziezold@mathematik.uni-kassel.de

curvature, principal component geodesics might meet at a point different from the intrinsic (Fréchet) mean. Actually, proving by example that they do so turns out to be nontrivial. The set-up and the proof of this in the following sections are one main result of this paper. As the intrinsic mean does not come to lie on the minimizing geodesic, various possible definitions of total geodesic variance come to mind.

In the fifth section, based on the general method presented in Section 2.3 we develop an algorithm finding the intrinsic mean and the principal component geodesics on spheres of any dimension. As the shape space of planar triangles is a two-dimensional sphere in three-dimensional Euclidean space, we can illustrate our results with planar triangular shape data. Our method allows us to compare principal component geodesics of different data samples graphically.

In fact, by using arbitrary geodesics instead of projections of straight lines in the tangent space of a mean (the projections are usually also geodesics) we obtain a better fit and find an increase in the amount of variance explained by the first principal component geodesic. These findings encourage our effort to apply the method to more general shape spaces in future work.

We note that in the case of a hyperbolic model for simplex shape spaces, rather than Kendall's shape space model, which we treat here, much work has been done in finding algorithms converging to the intrinsic Fréchet mean; see [14], [16], and [18]. In [12] planar circular shapes were modelled in an infinite-dimensional Riemannian shape space and shape variation along geodesics was studied.

The idea to use curves as principal components is not new. In [6] principal component geodesics in special Lie groups were used and in [4] principal curves of low frequency were sought for planar distributions.

2. PCA based on geodesics

Let M be an m-dimensional Riemannian manifold with induced metric $d(\cdot, \cdot)$. When speaking of a geodesic we will mean a geodesic of *maximal length*. Therefore, we define

$$G(M) := \{\gamma : \gamma \text{ is a geodesic on } M \text{ of maximal length}\}.$$

For $p \in M$ and $\gamma \in G(M)$ we define

$$d(p, \gamma) := \inf_{q \in \gamma} d(p, q).$$

By X we denote an M-valued random variable, e.g. one given by N data points, $p_1, \ldots, p_N \in M$, with equal probabilities.

2.1. Means and principal component geodesics

A point $\bar{p} \in M$ is called an *intrinsic mean* (or an *intrinsic Fréchet mean*) of X if it minimizes the function

$$p \mapsto \mathrm{E}(d(X, p)^2) \tag{2.1}$$

on M.

A geodesic $\gamma_1 \in G(M)$ is called a *first principal component geodesic* to X if it minimizes

$$\gamma \mapsto \mathrm{E}(d(X, \gamma)^2) \tag{2.2}$$

on $G(M)$.

In Euclidean space the intrinsic mean and the first principal component geodesic are uniquely determined (except for obviously generic cases), and the first principal component geodesic passes through the intrinsic mean. We shall see that in general the latter is no longer the case for spaces with non-Euclidean geometry. *We assume in this paper that $\bar{p}$, γ_1, and all other means and principal component geodesics defined below exist and are uniquely determined. In most experimental situations this is the case.*

We call a geodesic, $\gamma_2 \in G(M)$, that minimizes (2.2) over all geodesics $\gamma \in G(M)$ that have at least one point in common with γ_1 and are orthogonal to γ_1 at all common points a *second principal component geodesic* to X. Every point $\hat{p}$ that minimizes (2.1) over all common points of γ_1 and γ_2 will be called a *principal component geodesic mean* (e.g. on spheres any γ_1 and γ_2 have at least two common points).

Given the first and second principal component geodesics γ_1 and γ_2 with principal component geodesic mean $\hat{p}$, we say that a geodesic γ_3 is a *third principal component geodesic* if it minimizes (2.2) over all geodesics that meet γ_1 and γ_2 orthogonally at $\hat{p}$. Principal component geodesics of higher order are defined analogously.

Given the principal component geodesics to X, $\gamma_1, \gamma_2, \ldots, \gamma_m$, we denote by $X^{(j)}$ the orthogonal projection of X onto γ_j, $1 \leq j \leq m$. In most practical situations these projections will also be uniquely determined.

A minimizer $\bar{p}^{(j)} \in \gamma_j$, $1 \leq j \leq m$, of the function $\mathrm{p} \mapsto \mathrm{E}(d(X^{(j)}, p)^2)$ on the geodesic γ_j will be called an *intrinsic mean of X on the geodesic γ_j*.

2.2. Geodesic variance

Suppose that we have specified the principal component geodesics $\gamma_1, \ldots, \gamma_m$, the intrinsic mean, $\bar{p}$, the principal component geodesic mean, $\hat{p}$, and the intrinsic mean $\bar{p}^{(1)}$ (on γ_1) of an M-valued random variable X.

In Euclidean space we have $\hat{p} = \bar{p} = \bar{p}^{(1)}$ and, for the total variance,

$$V_{\mathrm{Eucl.}}(X) := \mathrm{E}(d(X, \bar{p})^2) = \sum_{s=1}^{m} V_{\mathrm{Eucl.}}^{(s)}(X),$$

with the variances explained by the sth principal component ($1 \leq s \leq m$) given by

$$V_{\mathrm{Eucl.}}^{(s)}(X) = \mathrm{E}(d(X^{(s)}, \bar{p})^2) = \mathrm{E}\left(\frac{1}{m-1} \sum_{j=1}^{m} d(X, \gamma_j)^2 - d(X, \gamma_s)^2\right). \tag{2.3}$$

The generalization of (2.3) is inspired by two facts. First, we shall see that in general $\hat{p} \neq \bar{p}$ on arbitrary manifolds. Second, it is clear that the Pythagoras theorem does not extend to arbitrary manifolds, i.e. for a geodesic parallelogram with ordered vertices q_1, q_2, q_3, and q_4 we in general have $d(q_1, q_2) \neq d(q_3, q_4)$ and $d(q_1, q_4) \neq d(q_2, q_3)$.

We call the generalization of the second term in (2.3) the *geodesic variance* explained by the sth principal component geodesic ($1 \leq s \leq m$) as *obtained by projection*,

$$V_{\mathrm{gp}}^{(s)}(X) := \mathrm{E}(d(X^{(s)}, \hat{p})^2), \tag{2.4}$$

and set

$$V_{\mathrm{gp}}(X) = \sum_{s=1}^{m} V_{\mathrm{gp}}^{(s)}(X).$$

These quantities, however, can be unduly distorted as a result of curvature (see the third column of Table 1, below).

The generalization of the third term in (2.3),

$$V_{\mathrm{gr}}^{(s)}(X) := \mathrm{E}\left(\frac{1}{m-1}\sum_{j=1}^{m} d(X,\gamma_j)^2 - d(X,\gamma_s)^2\right), \tag{2.5}$$

will be called the *geodesic variance* explained by the sth principal component geodesic (with $1 \le s \le m$) as *obtained by residuals*. Furthermore, we set

$$V_{\mathrm{gr}}(X) = \sum_{s=1}^{m} V_{\mathrm{gr}}^{(s)}(X).$$

For dimension $m = 2$, or for $m > 2$ when comparing only total variance with variation explained by the first principal component geodesic, mixing both definitions we let

$$V_{\mathrm{gm}}(X) := V_{\mathrm{gp}}^{(1)}(X) + \mathrm{E}(d(X,\gamma_1)^2) = \mathrm{E}(d(X^{(1)},\hat{p})^2) + \mathrm{E}(d(X,X^{(1)})^2).$$

Replacing $\hat{p}$ by $\bar{p}^{(1)}$, we propose the following definition of *mixed geodesic variance*:

$$\begin{aligned} V_{\mathrm{gm}}(X) &:= \mathrm{E}(d(X^{(1)},\bar{p}^{(1)})^2) + \mathrm{E}(d(X,\gamma_1)^2) \\ &= \mathrm{E}(d(X^{(1)},\bar{p}^{(1)})^2) + \mathrm{E}(d(X,X^{(1)})^2). \end{aligned} \tag{2.6}$$

The definition (2.4) is very similar to the definition of geodesic variance for Riemannian Lie groups in [6]. There, however, $\bar{p}$ is used instead of $\hat{p}$. Also, in defining higher-order principal component geodesics the authors did not require orthogonality at the point of intersection but made use of the multiplicative structure of the group in a natural way. The results were applied to three-dimensional medical imaging; see also [5].

2.3. A method of principle component analysis based on geodesics

In this section we propose a method of principle component analysis based on geodesics for a special class of Riemannian manifolds. The method will produce a fixed-point equation

$$y = f(y),$$

which naturally defines a numerical algorithm $y_{n+1} = f(y_n)$. We apply this method in Section 5 without proving convergence of the numerical iterations in general. In fact, it turns out that the convergence is rather good when applying the algorithms to the data in Section 6.

As is well known, every Riemannian manifold M can be embedded isometrically in a Euclidean space of sufficiently high dimension. This is a famous theorem of Nash [19]. Sometimes a suitable non-isometrical embedding $M \hookrightarrow \mathbb{R}^n$ is preferred. In any case, for a sufficiently large $n > m$ we assume here that an m-dimensional Riemannian manifold, M, and its tangent spaces are implicitly defined by

$$\begin{aligned} M &= \{x \in \mathbb{R}^n : \phi(x) = 0\}, \\ T_x M &= \{v \in \mathbb{R}^n : \mathrm{d}\phi(x)\, v = 0\}, \qquad x \in M, \end{aligned} \tag{2.7}$$

for a suitable smooth function $\phi : \mathbb{R}^n \to \mathbb{R}^{n-m}$ with $\mathrm{d}\phi(x) : \mathbb{R}^n \to \mathbb{R}^{n-m}$ having full rank for all $x \in M$. The manifold M is closed and, thus, complete, implying that maximal geodesics

$t \mapsto \gamma(t)$ are defined for all $t \in \mathbb{R}$. We denote by $\langle \cdot, \cdot \rangle$ the Riemannian metric on $T_x M$. For an isometrical embedding this is the standard Euclidean inner product.

Before continuing we remark that in general the representation (2.7) will be possible in local charts only. Our method, explained below, might become much more complicated if we incorporate transitions between several charts, and might be even more complicated for topological spaces for which only subspaces can be treated as Riemannian manifolds, as is the case for some shape spaces (see e.g. [10, p. 62, pp. 110ff.]). In particular, if the manifold is noncomplete, shortest geodesics might not exist. Obviously, the applicability of our method has to be checked in the respective examples and cannot be assumed in full generality.

We return to the representation (2.7). Every geodesic on M is uniquely determined by an offset, $x \in M$, and an initial direction, $v \in T_x M$, of unit length, i.e. $\langle v, v \rangle = 1$. Geodesics are denoted by $t \mapsto \gamma_{x,v}(t)$, with offset $\gamma_{x,v}(0) = x$ and initial direction $\dot{\gamma}_{x,v}(0) = v$. Given a data sample $p_1, \ldots, p_N \in M$ and a geodesic $\gamma_{x,v}$ define (cf. (2.2))

$$F(x, v) := \sum_{i=1}^{N} d(p_i, \gamma_{x,v})^2.$$

2.3.1. *First principal component geodesic.* Letting

$$\Phi_1 \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{2n-2m+1}, \qquad (x, v) \mapsto \begin{pmatrix} \phi(x) \\ \mathrm{d}\phi(x)\, v \\ \langle v, v \rangle - 1 \end{pmatrix},$$

finding a first principal component geodesic is equivalent to solving the extremal problem

$$\begin{aligned} &\text{find } (x^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^n \text{ such that} \\ &\quad F(x^*, v^*) = \inf\{F(x, v) \colon x, v \in \mathbb{R}^n \text{ with } \Phi_1(x, v) = 0\}. \end{aligned} \tag{2.8}$$

A standard method of solving this nonlinear extremal problem under constraining conditions involves employing Lagrange multipliers. Every solution, $(x, v) \in \mathbb{R}^n \times \mathbb{R}^n$, of (2.8) also solves

$$\mathrm{d}F + \lambda^\top \mathrm{d}\Phi_1 = 0 \tag{2.9}$$

for a suitable vector $\lambda \in \mathbb{R}^{2n-2m+1}$. Of course, '$\top$' means transposition of matrices and vectors. From (2.9) two fixed-point equations can be derived by separately considering the partial derivatives with respect to the coordinates of x and v. These equations naturally yield an algorithm to determine the solution (x^*, v^*), as explained above.

Making sure that a thus-obtained sequence (x_n, v_n) leads to decreasing values of F will entail that we approach not a local maximum but a local minimum. Convergence to the same local minimum from several starting values will give further evidence that we have found a global minimum.

We will exemplify the method in Section 5.

2.3.2. *Second principal component geodesic and principal component mean.* Given a first principal component geodesic $t \mapsto \gamma_{x,v}(t)$, a second principal component geodesic must pass through a point $y = \gamma_{x,v}(\tau)$ with an initial direction, $w \in T_y M$, orthogonal to $\dot{\gamma}_{x,v}(\tau)$. Hence, with

$$\Phi_2 \colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n-m+2}, \qquad (\tau, w) \mapsto \begin{pmatrix} \mathrm{d}\phi(\gamma_{x,v}(\tau))\, w \\ \langle \dot{\gamma}_{x,v}(\tau), w \rangle \\ \langle w, w \rangle - 1 \end{pmatrix},$$

and $F_2(\tau, w) := F(\gamma_{x,v}(\tau), w)$, finding a second principal component geodesic is equivalent to solving the extremal problem

$$\text{find } (\hat{\tau}, \hat{w}) \in \mathbb{R} \times \mathbb{R}^n \text{ such that}$$
$$F_2(\hat{\tau}, \hat{w}) = \inf\{F_2(\tau, w) \colon \tau \in \mathbb{R} \text{ and } w \in \mathbb{R}^n \text{ with } \Phi_2(\tau, w) = 0\}.$$

This will again be solved by the method of Lagrange multipliers, by solving

$$\mathrm{d}F_2 + \lambda^\top \mathrm{d}\Phi_2 = 0 \tag{2.10}$$

for $\tau \in \mathbb{R}$, $w \in \mathbb{R}^n$, and $\lambda \in \mathbb{R}^{n-m+2}$. For convenience, having found $\hat{\tau}$ and $\hat{w}$, let $v_2 := \hat{w}$ and write $\gamma_{x,v}$ as $\gamma_{\hat{x},v_1}$, where $\hat{x} := \gamma_{x,v}(\hat{\tau})$ and

$$v_1 := \frac{\dot{\gamma}_{x,v}(\hat{\tau})}{|\dot{\gamma}_{x,v}(\hat{\tau})|}.$$

Note that $\hat{x}$ is a principal component geodesic mean.

2.3.3. *Higher-order principal component geodesics.* All principal component geodesics of order j, $3 \le j \le m$, pass through the principal component geodesic mean $\hat{x} \in M$, i.e. each is determined only by an initial direction $v_j \in \mathbb{R}^n$ at offset $\hat{x}$. In particular, v_j is perpendicular to all lower-order principal component geodesics at $\hat{x}$.

Suppose that we have already found $j - 1 \ge 2$ principal component geodesics, namely $\gamma_{\hat{x},v_1}, \ldots, \gamma_{\hat{x},v_{j-1}}$. Then, defining

$$\Phi_j \colon \mathbb{R}^n \to \mathbb{R}^{n-m+j}, \qquad v \mapsto \begin{pmatrix} \mathrm{d}\phi(\hat{x})\, v \\ \langle v, v_1 \rangle \\ \vdots \\ \langle v, v_{j-1} \rangle \\ \langle v, v \rangle - 1 \end{pmatrix},$$

and $F_3(v) := F(\hat{x}, v)$, finding a jth principal component geodesic is equivalent to solving the extremal problem

$$\text{find } v_j \in \mathbb{R}^n \text{ such that } F_3(v_j) = \inf\{F_3(v) \colon v \in \mathbb{R}^n \text{ with } \Phi_j(v) = 0\}.$$

As before, this leads to the task of solving the equation

$$\mathrm{d}F_3 + \lambda^\top \mathrm{d}\Phi_j = 0 \tag{2.11}$$

for $v \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^{n-m+j}$.

2.3.4. *Intrinsic mean.* An intrinsic mean $\bar{x}$ can be found in a similar fashion. For this purpose consider

$$G(x) := \sum_{i=1}^{N} d(x, p_i)^2.$$

Finding an intrinsic mean is equivalent to solving the extremal problem

$$\text{find } \bar{x} \in \mathbb{R}^n \text{ such that } G(\bar{x}) = \inf\{G(x) \colon x \in \mathbb{R}^n \text{ with } \phi(x) = 0\}.$$

The method of Lagrange multipliers yields

$$\mathrm{d}G + \lambda^\top \mathrm{d}\phi = 0 \tag{2.12}$$

for $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^{n-m}$.

2.3.5. *Intrinsic mean on a geodesic.* Given a geodesic $t \mapsto \gamma(t) := \gamma_{x,v}(t)$, we want to find the point $\bar{x}^\gamma = \gamma(\bar{t})$ best approximating the orthogonal projections onto $\gamma_{x,v}$, q_i, of the data points p_i, $i = 1, \ldots, N$. This is an unconstrained extremal problem for

$$G_1(t) := \sum_{i=1}^{N} d(q_i, \gamma_{x,v}(t))^2 \tag{2.13}$$

in one variable, $t \in \mathbb{R}$.

In Section 5 we will determine the functions F, ϕ, ψ, G, and G_1 explicitly for spheres, and obtain the corresponding fixed-point equations and the respective algorithms. In Section 6 we will apply the method to shape data on a two-sphere.

3. Distance to geodesics on spheres

Denote by $\langle \cdot, \cdot \rangle$ the inner product of the standard Euclidean space $\mathbb{R}^{m+1}$ and let

$$S = \{p \in \mathbb{R}^{m+1} \colon \langle p, p \rangle = 1\}$$

be the m-dimensional unit sphere. The immersion $S \hookrightarrow \mathbb{R}^{m+1}$ induces a Riemannian metric, i.e. the spherical metric on S. For any two points $a, b \in S$, the spherical distance is given by

$$d(a, b) = 2 \arcsin\left(\frac{\sqrt{\langle a - b, a - b \rangle}}{2} \right) = \arccos\langle a, b \rangle. \tag{3.1}$$

Geodesics on spheres are precisely the great circles given by

$$\gamma \colon t \mapsto a \cos t + b \sin t$$

for any $a, b \in S$, $\langle a, b \rangle = 0$, and $t \in \mathbb{R}$.

Proposition 3.1. *Suppose that $t \mapsto \gamma(t) = a \cos t + b \sin t$ with $a, b \in S$, $\langle a, b \rangle = 0$, is a geodesic on S. The Riemannian distance between any point $z \in S$ and γ is then given by*

$$d(z, \gamma) = \arccos \sqrt{\langle a, z \rangle^2 + \langle b, z \rangle^2}.$$

Proof. The assertion is obvious for a point z on γ. For any other point, let $e_1 := a$ and $e_2 := b$ and find a third unit vector e_3, orthogonal to e_1 and e_2, such that $z = z_1 e_1 + z_2 e_2 + z_3 e_3$ with $z_3 > 0$. If $z_1 = 0 = z_2$, the assertion is again obvious. Otherwise, choose an $\alpha \in (0, \pi/2)$ such that $\cos\alpha = (z_1^2 + z_2^2)^{1/2}$ and $\sin\alpha = z_3$, introduce new orthonormal coordinates

$$e_1' := \frac{z_1 e_1 + z_2 e_2}{\cos\alpha}, \qquad e_2' := \frac{z_2 e_1 - z_1 e_2}{\cos\alpha}, \qquad e_3' := e_3$$

and reparametrize the geodesic $t \mapsto \gamma(t)$ as

$$s \mapsto \gamma(s) = e_1' \cos(s) + e_2' \sin(s) = (\cos s, \sin s, 0).$$

Since $z = (\cos\alpha, 0, \sin\alpha)$ in the new coordinates, using (3.1) we have

$$d(z, \gamma(s)) = \arccos(\cos s \cos\alpha),$$

which attains its minimum α for $\cos s = 1$. The assertion then follows from $z_1 = \langle a, z \rangle$, $z_2 = \langle b, z \rangle$, and $(z_1^2 + z_2^2)^{1/2} = \cos\alpha$.

Corollary 3.1. *For $0 \le d(z, \gamma) < \pi/2$ the geodesic projection of z onto γ is the point*

$$e_1' = \frac{\langle a, z \rangle a + \langle b, z \rangle b}{\sqrt{\langle a, z \rangle^2 + \langle b, z \rangle^2}}.$$

4. The principal component geodesics omit the intrinsic mean: an example on $S^2(1)$

In this section we shall show that principal component geodesics do not always pass through the intrinsic Fréchet mean. This will be a rather tedious task involving spherical trigonometry. We first fix $\alpha,\ 0 < \alpha < \pi/2$, arbitrarily and consider a one-parameter ($\delta,\ 0 \le \delta \le \pi/2$) family of triples of data points given by vertices of isosceles triangles on the two-dimensional unit sphere. For each δ, define the corresponding triple (p_1, p_2, p_3) by the following three points:

$$\begin{aligned} p_1 &:= (\cos\alpha, \sin\alpha, 0), \\ p_2 &:= (\cos\alpha, -\sin\alpha, 0), \\ p_3 \equiv p_3(\delta) &:= (\cos\delta, 0, \sin\delta). \end{aligned}$$

Any first principal component geodesic to these three points, i.e. to a random variable X taking each point with probability $\frac{1}{3}$, will pass through a point of the form

$$p = (\cos\varepsilon, 0, \sin\varepsilon). \tag{4.1}$$

Once we know that the geodesic and ε are uniquely determined (this will be established for small $\delta > 0$ in Theorem 4.1), we can write

$$\varepsilon = \nu(\delta) \tag{4.2}$$

for a function ν. In fact, we show in Lemma 4.2 that, by (4.1), ν uniquely determines the principal component geodesic mean for small $\delta > 0$.

The intrinsic mean of the three points p_1, p_2, and p_3 is uniquely determined (see, e.g. [9]) and, by symmetry, is of the form

$$\bar{p} = (\cos\epsilon, 0, \sin\epsilon)$$

for a unique ϵ. We analogously write

$$\epsilon = \mu(\delta) \tag{4.3}$$

for a function μ. Note that $0 \le \mu(\delta) \le \delta$. The main result of this section is the following theorem, which implies that $\varepsilon > \epsilon$ for small $\delta > 0$.

Theorem 4.1. *Let $p_1, p_2, p_3 \in S^2(1)$ ($p_3 \equiv p_3(\delta)$) be the above-defined data points. Then, for sufficiently small $\delta > 0$ and $\alpha,\ 0 < \alpha < \pi/2$,*

(i) *there is a unique first principal component geodesic to these points;*

(ii) *this first principal component geodesic does not pass through the intrinsic Fréchet mean of the data points;*

(iii) $$\frac{\delta}{3} < \mu(\delta) = \frac{\delta}{1 + 2\alpha\cot\alpha} + O(\delta^3) < \nu(\delta) = \frac{\delta}{1 + 2\cos^2\alpha} + O(\delta^3) < \delta. \tag{4.4}$$

Let us prepare for a first lemma. With p as in (4.1), any first principal component geodesic with an initial direction

$$v = (\sin\eta\sin\varepsilon, \cos\eta, -\sin\eta\cos\varepsilon)$$

at p, orthogonal to p, with $\eta \in (-\pi, \pi]$, will be of the form

$$\gamma_1(t) = p\cos t + v\sin t. \tag{4.5}$$

Lemma 4.1. *For small $\delta > 0$ there is a unique first principal component geodesic to the points p_1, p_2, and $p_3(\delta)$. This principal component geodesic is parallel to the equator at p, i.e.* $\cos\eta = 1$. *As a consequence, for small $\delta > 0$ there is a unique point p of the form (4.1) on the first principal component geodesic.*

Proof. Let $a := \cos\alpha$ and

$$A := \cos^2\varepsilon\cos^2\alpha + \sin^2\alpha = 1 - a^2\sin^2\varepsilon,$$
$$B := \sin\varepsilon\sin\alpha\cos\alpha,$$
$$C := \cos^2(\delta - \varepsilon).$$

First, we claim that $\cos\eta = 1$ for any first principal geodesic with $\delta > 0$ sufficiently small. To see this we use Proposition 3.1, write $\sum_{i=1}^3 d(p_i, \gamma_1)^2$ as a function of η and ε, and verify that

$$\begin{aligned}\left(\sum_{i=1}^3 d(p_i,\gamma_1)^2\right)(\eta,\varepsilon) = {}& \arccos^2\sqrt{A + 2B\sin\eta\cos\eta - (A - a^2)\sin^2\eta} \\ &+ \arccos^2\sqrt{A - 2B\sin\eta\cos\eta - (A - a^2)\sin^2\eta} \\ &+ \arccos^2\sqrt{C + (1 - C)\sin^2\eta}.\end{aligned}$$

Series expansion yields

$$\frac{\partial}{\partial\eta}\left(\sum_{i=1}^3 d(p_i,\gamma_1)^2\right)(0,\varepsilon) = 0,$$
$$\frac{\partial^2}{\partial\eta^2}\left(\sum_{i=1}^3 d(p_i,\gamma_1)^2\right)(0,\varepsilon) > 2(1 - a^2),$$

for ε sufficiently small, which implies, as claimed, that for small $\delta > 0$ any first principal geodesic will be parallel to the equator at p, i.e. $v = (0, 1, 0)$.

Second, consider the derivative with respect to ε:

$$\frac{\partial}{\partial\varepsilon}\left(\sum_{i=1}^3 d(p_i,\gamma_1)^2\right)(0,\varepsilon) = -2(\delta - \varepsilon) + 4a\cos\varepsilon\frac{\arccos\sqrt{A}}{\sqrt{A}}.$$

Thus, γ_1 is minimizing only if

$$\delta = \varepsilon + 2\cos\varepsilon\cos\alpha\frac{\arccos\sqrt{1 - \cos^2\alpha\sin^2\varepsilon}}{\sqrt{1 - \cos^2\alpha\ \sin^2\varepsilon}}. \tag{4.6}$$

Note that $(\mathrm{d}/\mathrm{d}\varepsilon)(\cos\varepsilon\arccos\sqrt{A}/\sqrt{A})$ is strictly decreasing in ε, $0 < \varepsilon < \pi/2$ (recall that $0 < a \leq 1$). Hence, the right-hand side of (4.6) is a strictly concave function in ε, $0 < \varepsilon < \pi/2$, taking values from 0 to $\pi/2$. Thus, given δ, $0 < \delta \leq \pi/2$, there is a unique ε, $0 < \varepsilon \equiv \varepsilon(\delta) < \delta$, solving (4.6).

Finally, we see that this solution of (4.6) yields a local minimum, since

$$\frac{\partial^2}{\partial\varepsilon^2}\left(\sum_{i=1}^3 d(p_i,\gamma_1)^2\right)(0,\varepsilon) > 2\left(1 + \frac{2a^2\cos^2\varepsilon}{A} - \left(\frac{\pi}{2} - \varepsilon\right)\tan\varepsilon\,\frac{1 - a^2}{A}\right) > 0$$

because $(1 - a^2)/A < 1$ and $(\pi/2 - \varepsilon)\tan\varepsilon < 1$ for $0 < \varepsilon < \pi/2$.

Next we establish that (4.2) uniquely determines the principal component mean for sufficiently small $\delta > 0$.

Lemma 4.2. *For small $\delta > 0$ there is a unique second principal component geodesic and, thus, a unique principal component mean, given by (4.2). The second principal component geodesic is of the form*

$$\gamma_2(s) = (1, 0, 0)\cos s + (0, 0, 1)\sin s.$$

Proof. For sufficiently small $\delta > 0$, let γ_1 be the first principal component geodesic determined by (4.5). For a suitable $t \in [-\pi, \pi)$, any second principal component geodesic will be of the form

$$\gamma_2(s) = (\cos\varepsilon\cos t, \sin t, \sin\varepsilon\cos t)\cos s + (-\sin\varepsilon, 0, \cos\varepsilon)\sin s,$$

as it orthogonally intersects the first principal component geodesic at some point $\gamma_1(t)$. Moreover, from Proposition 3.1 and

$$\begin{aligned} D \equiv D_\varepsilon(t) &:= (\cos\varepsilon\cos\alpha\cos t + \sin t\sin\alpha)^2 + \sin^2\varepsilon\cos^2\alpha, \\ E \equiv E_\varepsilon(t) &:= (\cos\varepsilon\cos\alpha\cos t - \sin t\sin\alpha)^2 + \sin^2\varepsilon\cos^2\alpha, \\ H \equiv H_\varepsilon(t) &:= \cos^2 t\cos^2(\delta-\varepsilon) + \sin^2(\delta-\varepsilon), \end{aligned}$$

we have

$$f_\varepsilon(t) := \sum_{i=1}^{3} d(p_i, \gamma_2)^2 = \arccos^2\sqrt{D} + \arccos^2\sqrt{E} + \arccos^2\sqrt{H}.$$

Note that $f_\varepsilon(0) = 2\alpha^2$ for all $\varepsilon \in \mathbb{R}$, that $f_0(t) = 3t^2 + 2\alpha^2$, and that $(\mathrm{d}/\mathrm{d}t) f_\varepsilon(0) = 0$ as $D_\varepsilon(0) = E_\varepsilon(0)$, $H_\varepsilon(0) = 1$, and

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\varepsilon(0) = 2\cos\alpha\cos\varepsilon = -\frac{\mathrm{d}}{\mathrm{d}t} E_\varepsilon(0).$$

This implies that, for small $\delta > 0$ (i.e. small $\varepsilon > 0$), the minimum of $f_\varepsilon(t)$ over $t \in [-\pi, \pi)$ is uniquely attained at $t = 0$. This yields the assertion.

We are now in position to prove Theorem 4.1.

Proof of Theorem 4.1. In conjunction with Lemma 4.1 and Lemma 4.2, all we need to prove is the assertion in (4.4). Suppose that $\bar{p}$ is the intrinsic mean given by $\epsilon = \mu(\delta)$ in (4.3). Recalling (3.1), consider the derivative of the squared distances for the intrinsic mean,

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\sum_{i=1}^{3} d(p_i, \bar{p})^2\right) = -2(\delta - \epsilon) + 4a\sin\epsilon\,\frac{\arccos(a\cos\epsilon)}{\sqrt{1 - a^2\cos^2\epsilon}},$$

which vanishes if and only if

$$\delta = \epsilon + 2\sin\epsilon\cos\alpha\,\frac{\arccos(\cos\alpha\cos\epsilon)}{\sqrt{1 - \cos^2\alpha\,\cos^2\epsilon}}.$$

For small ϵ, the right-hand side is

$$\left(1 + 2\frac{a\arccos a}{\sqrt{1-a^2}}\right)\epsilon + O(\epsilon^3) = (1 + 2\alpha\cot\alpha)\epsilon + O(\epsilon^3).$$

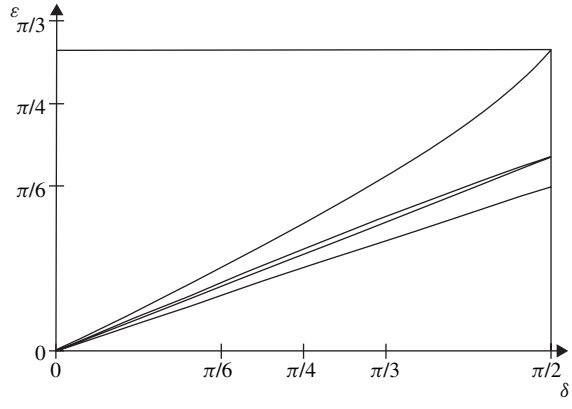

FIGURE 1: The four means, for $\alpha = \pi/4$: from top to bottom we display the function $\nu(\delta)$, yielding the principal component mean for small $\delta > 0$; the Euclidean mean (see (4.7)); the function $\mu(\delta)$, yielding the intrinsic mean; and the intrinsic mean on the second principal component geodesic.

which, in conjunction with $\alpha < \tan\alpha$ for $0 < \alpha < \pi/2$, yields the first part of the assertion. By series expansion of (4.6) we further obtain

$$\delta = (1 + 2a^2)\epsilon + O(\epsilon^3).$$

This proves the rest of the assertion.

Remark 4.1. Note that the linear approximation in 4.4 is rather good, i.e. that $\mu(\delta)$ and $\nu(\delta)$ are almost linear in δ. It turns out that – unlike in Euclidean geometry – even in the equilateral case ($\alpha = \delta/2 = \pi/4$), the horizontal geodesic $\gamma_{\mathrm{h}}(t) = (\cos\varepsilon, 0, \sin\varepsilon)\cos t + (0, 1, 0)\sin t$ is a better approximation than the vertical geodesic $\gamma_{\mathrm{v}}(t) = (1, 0, 0)\cos t + (0, 0, 1)\sin t$, with

$$\sum_{i=1}^{3} d(p_i, \gamma_{\mathrm{h}})^2 < 1.14 < 1.23 < 2\left(\frac{\pi}{4}\right)^2 = \sum_{i=1}^{3} d(p_i, \gamma_{\mathrm{v}})^2.$$

See also Table 1, below. In Figure 1 we illustrate the case $\alpha = \pi/4$.

In conclusion we note that the intrinsic mean on γ_2 of the points p_1, p_2, and p_3 is given by (see Section 5.5)

$$\epsilon = \frac{\delta}{3},$$

and the Euclidean mean (the mean of the three data points in $\mathbb{R}^3$ projected to the unit sphere) takes the value

$$(\cos\epsilon, 0, \sin\epsilon)$$

with

$$\epsilon = \arctan\frac{\sin\delta}{2\cos\alpha + \cos\delta}. \tag{4.7}$$

From Theorem 4.1 we infer that all four means disagree for small δ (in Figure 1 we display them as functions of δ). Indeed, for $0 < \alpha < \pi/2$ and small $\delta > 0$,

$$\frac{\delta}{3} < \mu(\delta) < \arctan\frac{\sin\delta}{2\cos\alpha + \cos\delta} < \nu(\delta).$$

5. Algorithms for geodesic PCA and means on spheres

In this section we apply the method of principal component analysis based on geodesics, as put forth in Section 2.3, to a unit sphere $S := S^m \subset \mathbb{R}^{m+1}$ and N data points $p_1, \dots, p_N \in S$. The unit sphere is defined by

$$\phi(x) = \langle x, x\rangle - 1 = 0,$$

and every tangent space is given by $T_x S = \{v \in \mathbb{R}^n : \mathrm{d}\phi(x)\, v = 2\langle x, v\rangle = 0\}$. Thus,

$$\Phi_1(x, v) := (\langle x, x\rangle - 1, 2\langle x, v\rangle, \langle v, v\rangle - 1)^\top$$

for $x, v \in \mathbb{R}^{m+1}$. Observe that

$$\gamma_{x,v}(t) := x \cos t + v \sin t$$

is a geodesic on S if and only if $\Phi_1(x, v) = (0, 0, 0)^\top$. Moreover (see Proposition 3.1), we have the following two distance functions:

$$F(x, v) = \sum_{i=1}^N d(p_i, \gamma_{(x,v)})^2 = \sum_{i=1}^N \arccos^2 \sqrt{\langle x, p_i\rangle^2 + \langle v, p_i\rangle^2},$$
$$G(x) = \sum_{i=1}^N d(p_i, x)^2 = \sum_{i=1}^N \arccos^2 \langle x, p_i\rangle.$$

5.1. The first principal component great circle

Note that if (x^*, v^*) is a solution to (2.8), then any $(ax^* + bv^*, cx^* + dv^*)$ is also a solution when

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{O}(2).$$

This ambiguity can be overcome most simply by requiring that $x^{j_0} = 0$ and $x^{j_1} \geq 0$ (or $x^{j_1} < 0$) for suitable component indices j_0 and j_1, $1 \leq j_0 \neq j_1 \leq m + 1$.

By using a Lagrange multiplier $\lambda = (\lambda_1, \lambda_2, \lambda_3)^\top$ and writing

$$\zeta_i := \sqrt{\langle x, p_i\rangle^2 + \langle v, p_i\rangle^2}, \qquad \xi_i := -\frac{1}{2\zeta_i} \frac{\mathrm{d}}{\mathrm{d}\zeta_i} \arccos^2 \zeta_i = \frac{\arccos \zeta_i}{\zeta_i \sqrt{1 - \zeta_i^2}},$$

the Lagrange equation (2.9) can be rewritten as

$$\sum_{i=1}^N \xi_i \langle x, p_i\rangle p_i = \lambda_1 x + \lambda_2 v, \qquad \sum_{i=1}^N \xi_i \langle v, p_i\rangle p_i = \lambda_3 v + \lambda_2 x.$$

Note that $\zeta_i \neq 1$ unless p_i lies on $\gamma_{(x,v)}$, in which case we replace ξ_i by 1. Moreover, $\zeta_i \neq 0$ if x and v are such that p_i is sufficiently close to $\gamma_{x,v}$. We want to assume the latter. By solving for x and v we obtain the fixed-point problem

$$\begin{aligned} \sum_{i=1}^N \xi_i (\lambda_3 \langle x, p_i\rangle - \lambda_2 \langle v, p_i\rangle) p_i &= (\lambda_1 \lambda_3 - \lambda_2^2) x, \\ \sum_{i=1}^N \xi_i (\lambda_2 \langle x, p_i\rangle - \lambda_1 \langle v, p_i\rangle) p_i &= (\lambda_2^2 - \lambda_1 \lambda_3) v, \end{aligned} \tag{5.1}$$

with

$$\sum_{i=1}^{N} \xi_i \langle x, p_i\rangle^2 = \lambda_1, \qquad \sum_{i=1}^{N} \xi_i \langle x, p_i\rangle\langle v, p_i\rangle = \lambda_2, \qquad \sum_{i=1}^{N} \xi_i \langle v, p_i\rangle^2 = \lambda_3.$$

Denoting by $\Psi_1(x, v)$ and $\Psi_2(x, v)$ the two left-hand sides of the fixed-point problem (5.1), in a natural way we define the algorithm

$$\begin{aligned} &(x_n, v_n) \mapsto (x_{n+1}, v_{n+1}), \text{ where} \\ &\qquad x_{n+1} = \frac{\Psi_1(x_n, v_n)}{\|\Psi_1(x_n, v_n)\|}, \\ &\qquad v_{n+1} = \frac{\Psi_2(x_n, v_n) - \langle \Psi_2(x_n, v_n), x_{n+1}\rangle x_{n+1}}{\|\Psi_2(x_n, v_n) - \langle \Psi_2(x_n, v_n), x_{n+1}\rangle x_{n+1}\|}. \end{aligned} \tag{5.2}$$

In practice (see Sections 6.4 and 6.5) the natural ambiguity in the above-mentioned parametrization of spherical geodesics seems irrelevant. Nevertheless, to compare with an intrinsic mean $\bar{p}$, for example, after each iteration reparametrize the geodesic $t \mapsto x\cos t + v\sin t$ such that $x^{j_0} = \bar{p}^{j_0}$ and $\operatorname{sgn}(x^{j_1}) = \operatorname{sgn}(\bar{p}^{j_1})$ for suitable indices j_0 and j_1, $0 \le j_0 \ne j_1 \le m+1$.

As starting point x_0 choose either one of the p_i or, more subtly, the spherical mean, $\bar{p}$, of $p_1, \ldots, p_N$, which can be computed using the algorithm of [15] or, in our special case, using the algorithm introduced below. For the starting direction v_0 either take the normalized part of any vector $x_0 - p_i$ orthogonal to x_0 or choose among the vectors $p_i - \bar{p}$ of maximal spherical length and again normalize that part orthogonal to $\bar{p}$.

5.2. The second principal component great circle

Having found a first principal component geodesic, $\gamma_1 = \gamma_{x,v}$, determined by $x, v \in S$, $\langle x, v\rangle = 0$, suppose that $\gamma_2(t) = \gamma_{y,w}(t) = y\cos t + w\sin t$, with $y = y(\tau) = x\cos\tau + v\sin\tau$ for a suitable $\tau \in \mathbb{R}$, is a second principal component geodesic. According to Section 2.3, we consider

$$\mathrm{d}\phi(\gamma_{x,v}(\tau))w = \langle 2(x\cos\tau + v\sin\tau), w\rangle = 0,$$

$$\langle \dot{\gamma}_{x,v}(\tau), w\rangle = \langle -x\sin\tau + v\cos\tau, w\rangle = 0.$$

This equation system is equivalent to

$$\langle x, w\rangle = 0, \qquad \langle v, w\rangle = 0.$$

Thus, instead of $\Phi_2(\tau, w) = 0$, we consider the equivalent constraint

$$\tilde{\Phi}_2(w) = (\langle x, w\rangle, \langle v, w\rangle, \langle w, w\rangle - 1)^\top = 0,$$

under which we want to minimize the function

$$\begin{aligned} F_2(\tau, w) &:= \sum_{i=1}^{N} d(p_i, \gamma_{y(\tau),w})^2 \\ &= \sum_{i=1}^{N} \arccos^2 \sqrt{(\langle x, p_i\rangle\cos\tau + \langle v, p_i\rangle\sin\tau)^2 + \langle w, p_i\rangle^2}. \end{aligned}$$

Writing

$$\begin{aligned} a_i &= a_i(\tau) := \langle x, p_i\rangle \cos\tau + \langle v, p_i\rangle \sin\tau, \\ b_i &= b_i(\tau) := \langle v, p_i\rangle \cos\tau - \langle x, p_i\rangle \sin\tau, \\ \zeta_i &= \zeta_i(\tau, w) := \sqrt{a_i^2 + \langle w, p_i\rangle^2}, \\ \xi_i &= \xi_i(\tau, w) := \frac{\arccos\zeta_i}{\zeta_i\sqrt{1-\zeta_i^2}} \end{aligned}$$

and defining a Lagrange multiplier $\lambda = (2\lambda_1, 2\lambda_2, \lambda_3)^\top$, from the Lagrange equation (2.10) we obtain

$$\sum_{i=1}^N \xi_i a_i b_i = 0, \qquad \sum_{i=1}^N \xi_i \langle w, p_i\rangle p_i = \lambda_1 x + \lambda_2 v + \lambda_3 w,$$

which implies that

$$\sum_{i=1}^N \xi_i \langle w, p_i\rangle\langle x, p_i\rangle = \lambda_1, \qquad \sum_{i=1}^N \xi_i \langle w, p_i\rangle\langle v, p_i\rangle = \lambda_2, \qquad \sum_{i=1}^N \xi_i \langle w, p_i\rangle^2 = \lambda_3.$$

By letting

$$\Psi_1(\tau, w) := \frac{\sum_{i=1}^N \xi_i \langle w, p_i\rangle p_i}{\sum_{i=1}^N \xi_i \langle w, p_i\rangle^2}, \qquad \Psi_2(\tau, w) := \sum_{i=1}^N \xi_i a_i b_i$$

we obtain the following algorithm: start with suitable initial values $\tau^{(0)}$ and $w^{(0)}$, e.g.

$$(\tau^{(0)}, w^{(0)}) = \left(0, \frac{p_1 - p_0 - \langle p_1 - p_0, x\rangle x - \langle p_1 - p_0, v\rangle v}{\|p_1 - p_0 - \langle p_1 - p_0, x\rangle x - \langle p_1 - p_0, v\rangle v\|}\right),$$

and compute $(\tau^{(n+1)}, w^{(n+1)})$ from $(\tau^{(n)}, w^{(n)})$ by setting

$$\begin{aligned} z^{(n+1)} &:= \Psi_1(\tau^{(n)}, w^{(n)}), \\ w^{(n+1)} &:= \frac{z^{(n+1)} - \langle z^{(n+1)}, x\rangle x - \langle z^{(n+1)}, v\rangle v}{\|z^{(n+1)} - \langle z^{(n+1)}, x\rangle x - \langle z^{(n+1)}, v\rangle v\|}, \\ \tau^{(n+1)} &:= \text{solution to } \Psi_2(\tau, w^{(n+1)}) = 0 \text{ in } [-\pi/2, \pi/2). \end{aligned} \tag{5.3}$$

5.3. Higher-order principal component great circles

For simplicity, set $x := \hat{x}$. Suppose that we have found principal component geodesics $\gamma_{x,v_1}, \ldots, \gamma_{x,v_{j-1}}$, $3 \le j \le m$. Introducing

$$\zeta_i := \sqrt{\langle x, p_i\rangle^2 + \langle v, p_i\rangle^2}, \qquad \xi_i := \frac{\arccos\zeta_i}{\zeta_i\sqrt{1-\zeta_i^2}},$$

and Lagrange multipliers $\lambda_0, \ldots, \lambda_j \in \mathbb{R}$, we can rewrite the Lagrange equation (2.11) as

$$\sum_{i=1}^N \xi_i \langle v, p_i\rangle p_i = \lambda_0 x + \sum_{s=1}^{j-1} \lambda_s v_s + \lambda_j v. \tag{5.4}$$

Starting with a suitable $v^{(0)}$, we thus compute $v^{(n+1)}$ from $v^{(n)}$ using the following algorithm, which follows in a natural way from (5.4):

$$
\begin{aligned}
z^{(n+1)} &:= \sum_{i=1}^{N} \xi_i^{(n)} \langle v^{(n)}, p_i \rangle p_i, \\
\lambda_0^{(n+1)} &:= \langle z^{(n+1)}, x \rangle, \\
\lambda_s^{(n+1)} &:= \langle z^{(n+1)}, v_s \rangle, \qquad 1 \le s < j, \\
\lambda_j^{(n+1)} &:= \sum_{i=1}^{N} \xi_i^{(n)} \langle v^{(n)}, p_i \rangle^2, \\
v^{(n+1)} &:= \operatorname{sgn}(\lambda_j) \frac{z^{(n+1)} - \lambda_0^{(n+1)} x - \sum_{s=1}^{j-1} \lambda_s^{(n+1)} v_s}{\| z^{(n+1)} - \lambda_0^{(n+1)} x - \sum_{s=1}^{j-1} \lambda_s^{(n+1)} v_s \|}.
\end{aligned}
$$

5.4. The spherical mean

Here we set $\zeta_i := \langle x, p_i \rangle$ and $\xi_i := \arccos \zeta_i / (1 - \zeta_i^2)^{1/2}$ and consider a single Lagrange multiplier $\lambda \in \mathbb{R}$, to obtain

$$\sum_{i=1}^{N} \xi_i p_i = \lambda x,$$

with

$$\sum_{i=1}^{N} \xi_i \langle p_i, x \rangle = \lambda,$$

from the Lagrange equation (2.12). Thus, with $\Psi(x) := (1/\lambda) \sum_{i=1}^{N} \xi_i p_i$ we have the following algorithm for the intrinsic mean:

$$x_n \mapsto x_{n+1} = \frac{\Psi(x_n)}{\| \Psi(x_n) \|}.$$

5.5. The intrinsic mean on a great circle

Suppose that we have specified a spherical geodesic, $t \mapsto \gamma_{x,v}(t)$, determined by $\gamma_{x,v}(0) = x$ and $\dot{\gamma}_{x,v}(0) = v$. With

$$\alpha_i := \arctan \frac{\langle v, p_i \rangle}{\langle x, p_i \rangle} \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right)$$

and

$$[-\pi, \pi) \ni t_i := \begin{cases} \alpha_i \bmod 2\pi & \text{if } \langle v, p_i \rangle, \langle x, p_i \rangle > 0 \text{ or } \langle v, p_i \rangle < 0 < \langle x, p_i \rangle, \\ \alpha_i + \pi \bmod 2\pi & \text{otherwise,} \end{cases}$$

from Corollary 3.1 we see that the geodesic projections of the data points $p_1, \ldots, p_N$ onto $\gamma_{x,v}$ are given by

$$q_i = x \cos t_i + v \sin t_i, \qquad i = 1, \ldots, N.$$

The function G_1 (see 2.13) is then given by

$$G_1(t) = \sum_{i=1}^{N} \arccos^2(\cos t \cos t_i + \sin t \sin t_i) = \sum_{i=1}^{N} \left(\epsilon_i \delta_i (t - t_i) + \frac{1 - \epsilon_i}{2} 2\pi \right)^2,$$

where $\delta_i = \operatorname{sgn}(t - t_i)$ and $\epsilon_i = \operatorname{sgn}(2\pi - |t - t_i|)$. This quantity is uniquely minimized by

$$t = \frac{1}{N}\sum_{i=1}^{N} t_i - \frac{2\pi}{N}\sum_{i=1}^{N} \epsilon_i \delta_i \frac{1-\epsilon_i}{2}.$$

The second sum can assume any integer values between $-N$ and N. In practice, we will determine $t^* := (1/N)\sum_{i=1}^{N} t_i$ and check which of the values

$$t^* + \frac{2\pi k}{N}, \qquad k = 0, \ldots, N-1,$$

minimizes the function G_1. This value yields an intrinsic mean on the geodesic.

In the following section the above algorithms are applied to planar triangles whose shape space is a two-dimensional sphere.

6. Geodesic PCA for planar triangular shape spaces

The consideration of spheres in matrix spaces modulo suitable rotation groups leads to *Kendall's* shape spaces. Denote by $M(m,k)$ the set of all real matrices with m rows and k columns with the inner product $\langle a, b\rangle := \operatorname{tr}(ab^\top)$ and $\|a\| = \sqrt{\langle a, a\rangle}$, where $\operatorname{tr}(\cdot)$ is the trace function, and by $\mathrm{SO}(m)$ the special orthogonal group in $M(m,m)$.

6.1. Kendall's shape spaces

Shape analysis is based on configurations consisting of k labelled vertices in $\mathbb{R}^m$, called *landmarks*, that do not all coincide. Each configuration is a point in $M(m,k)$. Disregarding location and size, these configurations are mapped by a Helmert matrix to *preshape space* (see, e.g. [3])

$$S \equiv S_m^k := \{s \in M(m, k-1) \colon \|s\| = 1\}.$$

Additionally, disregarding rotation leads to the definition of *shape space*. Define on S a smooth action of $\mathrm{SO}(m)$ by

$$gs := (gs_1, \ldots, gs_{k-1}) \in S,$$

for $g \in \mathrm{SO}(m)$ and $s = (s_1, \ldots, s_{k-1}) \in S$. Then the orbit $[s] = \{gs \colon g \in \mathrm{SO}(m)\}$ is the *shape* of $s \in S$ and the topological quotient

$$\Sigma \equiv \Sigma_m^k := S/\mathrm{SO}(m)$$

is called the *shape space*.

Shape spaces of one-dimensional objects are just the corresponding preshape spheres, as $\mathrm{SO}(1) = \{\mathrm{id}\}$ is trivial. For $m = 2$ and $k = 3$, i.e. planar triangular shapes, the above projection will be explicitly given below: it is the *Hopf fibration* projecting the preshape sphere of radius 1 in four-dimensional Euclidean space onto the two-dimensional shape space sphere $S^2(\frac{1}{2})$ of radius $\frac{1}{2}$ in three-dimensional Euclidean space.

6.2. Euclidean PCA for shape spaces

Two preshapes $p, x \in S$ are in *optimal position* to each other if

$$\|p - x\| = \inf_{g \in \mathrm{SO}(m)} \|gp - x\|.$$

As $\mathrm{SO}(m)$ is compact, any preshape p can be rotated into optimal position to a given preshape x. We will denote the optimally rotated version of p with respect to x by p^x, also called the *partial Procrustes fit* of p onto x.

Given preshapes $p_1, \ldots, p_N \in S$, call a preshape $\bar{x}_{\mathrm{Eucl.}} \in S$ a *preshape of an extrinsic mean shape* (or a *preshape of an extrinsic Fréchet mean shape*) if

$$\min_{g_1,\ldots,g_N \in \mathrm{SO}(m)} \sum_{i=1}^{N} \|g_i p_i - \bar{x}_{\mathrm{Eucl.}}\|^2 = \min_{x \in S} \left(\min_{g_1,\ldots,g_N \in \mathrm{SO}(m)} \sum_{i=1}^{N} \|g_i p_i - x\|^2 \right).$$

In [8] Gower proposed an algorithm to find a preshape of an extrinsic mean shape; see also [22].

Euclidean PCA for Kendall's shape spaces is performed as follows (see, e.g. [2], [3, p. 96], [7], and [11]). Having found a preshape $\bar{x} = \bar{x}_{\mathrm{Eucl.}} \in S$ of an extrinsic mean shape, all data points are brought into optimal position to $\bar{x}$ and projected onto the tangent space $T_{\bar{x}}S$ at the preshape of the extrinsic mean shape. Standard PCA is then performed with respect to the residuals

$$r_i := \frac{p_i^{\bar{x}}}{\langle p_i^{\bar{x}}, \bar{x} \rangle} - \bar{x} \in \mathbb{R}^{m(k-1)}, \qquad i = 1, \ldots, N.$$

6.3. Planar triangular shapes

Now consider n triangles, $Q_1, \ldots, Q_n \in M(2, 3)$, in the plane ($m = 2$), each determined by $k = 3$ landmarks. For any such triangle $Q = (q_1, q_2, q_3)$, $q_1, q_2, q_3 \in \mathbb{R}^2$, a corresponding point P in preshape space $S^3(1)$ is given by

$$P = \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{QH}{\|QH\|}, \qquad H = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{6} \\ 0 & -2/\sqrt{6}, \end{pmatrix}$$

where H is a so-called *Helmert submatrix* (see, e.g. [3]). In the complex notation $z = a + \mathrm{i}b$ and $w = c + \mathrm{i}d$, the Hopf fibration is then the composition of the map

$$(z, w) \mapsto \zeta = \frac{z}{w} = \frac{z\bar{w}}{\|w\|^2},$$

the inverse stereographic projection

$$\zeta \mapsto \frac{1}{|\zeta|^2 + 1} \begin{pmatrix} 2\,\mathrm{Re}(\zeta) \\ 2\,\mathrm{Im}(\zeta) \\ |\zeta|^2 - 1 \end{pmatrix},$$

and the following 'halving' (in order to have an isometry):

$$\begin{aligned} S^3(1) &\to S^2(\tfrac{1}{2}), \\ (z, w) &\mapsto \left(\mathrm{Re}(z\bar{w}), \mathrm{Im}(z\bar{w}), \frac{|z|^2 - |w|^2}{2} \right), \\ \begin{pmatrix} a & c \\ b & d \end{pmatrix} &\mapsto \left(ac + bd, bc - ad, \frac{a^2 + b^2 - c^2 - d^2}{2} \right). \end{aligned}$$

The spherical metric of the preshape sphere then naturally pushes forward to the spherical metric of the shape sphere $S^2(\frac{1}{2}) \cong \Sigma_2^3$.

For planar triangular shapes, having found a first principal component geodesic $\gamma_1(t) = x\cos t + v\sin t$ using (5.2), any second principal component geodesic will be of the form $\gamma_2(t) = y\cos t + w\sin t$, where $y = x\cos\tau + v\sin\tau$ for some suitable τ, $-\pi < \tau \leq \pi$, and x, v, and w form an orthonormal basis of $\mathbb{R}^3$. Hence, we only need to determine τ from (5.3). If x was obtained from starting at $\bar{p}$, τ will be close to 0. In fact, the algorithms converge rather quickly. We illustrate our method in two examples.

6.4. Example 1: An isosceles triangles family

Reconsider the family of isosceles triangles introduced in Section 4, in particular the 11 data triples $(p_1, p_2, p_3^{(n)})$ with the following points on the shape sphere:

$$p_1 = \frac{1}{2}\left(\cos\frac{\pi}{4}, \sin\frac{\pi}{4}, 0\right),$$
$$p_2 = \frac{1}{2}\left(\cos\frac{\pi}{4}, -\sin\frac{\pi}{4}, 0\right),$$
$$p_3^{(n)} = \frac{1}{2}\left(\cos\frac{n\pi}{20}, 0, \sin\frac{n\pi}{20}\right), \qquad n = 0, 1, \ldots, 10.$$

Every triple $(p_1, p_2, p_3^{(n)})$ corresponds to the shapes of the three planar triangles respectively having the points p_1, p_2, and $p_3^{(n)}$ as representatives.

At this point we note that a preshape of an extrinsic mean shape as defined in Section 6.2 is different from the Euclidean mean computed directly on the shape sphere. For the above isosceles data it turns out that the projection of the preshape of an extrinsic mean shape onto the shape sphere lies between the Euclidean mean and the intrinsic mean (which are already fairly close to each other; see Figure 1) both computed on the shape sphere itself.

In Table 1, for the data triples above we present the relative amount of variance explained by the first principal component using the classical Euclidean method in the tangent space of the preshape sphere (see Section 6.2), and the respective computations for the three different definitions of geodesic variance proposed in (2.4), (2.5), and (2.6). In Table 2 we display the relative improvement of the fit with respect to shape distances. The first Euclidean PC in the tangent space of the preshape sphere S^3 is first projected onto the preshape sphere and then onto a curve, δ, on the shape sphere. The percentages given compare the improvement of data fit when the sum of the squared distances of the data shapes to that projection is related to the sum of the squared distances of the data shapes to the first geodesic PC γ: we calculate

$$1 - \frac{\sum_{i=1}^3 d(p_i, \gamma)^2}{\sum_{i=1}^3 d(p_i, \delta)^2}.$$

From the bottom rows of Table 1 we can observe the effect of curvature on the various geodesic definitions for highly curved triangles. From the top rows of Table 1 it seems that the first Euclidean PC approximates the data nearly as well as the first geodesic PC. In comparing the two fits on the space directly in Table 2, however, we note that the first geodesic PC is also considerably closer to the data in the neighborhood of a great circle.

For $0 \leq n < 10$, the Euclidean and geodesic first PCs are 'horizontal' in the sense that they intersect the respective 'vertical' second PCs $t \mapsto (\cos t, 0\sin t)$, which describe a meridian, with a direction parallel to the equator.

TABLE 1.

	Relative variance explained by the first principal component			
	Euclidean	Geodesic PCA (%)		
n	PCA (%)	By projection	By residuals	Mixed
0	100.00	100.00	100.00	100.00
1	98.96	98.53	99.01	99.01
2	95.99	94.41	96.16	96.22
3	91.43	88.41	91.78	92.05
4	85.76	81.46	86.31	87.06
5	79.48	74.36	80.22	81.82
6	73.02	67.69	73.94	76.80
7	66.67	61.77	67.77	72.35
8	60.66	56.77	61.97	68.76
9	55.09	52.76	56.69	66.37
10	50.00	49.88	52.05	65.87

TABLE 2.

n	Relative improvement of geodesic data fit (%)
0	0.00
1	8.93
2	9.05
3	9.25
4	9.56
5	9.97
6	10.51
7	11.23
8	12.18
9	13.48
10	15.29

Finally, more subtly, consider the last row in each table, which corresponds to an equilateral triangle. In this case, any (of the infinitely many) first and second Euclidean PCs in the tangent space of a preshape of an extrinsic mean shape explains equally well half the projected data variation. In contrast, numerical investigation reveals that there are three first geodesic PCs, one of them 'horizontal', and three second geodesic PCs, one of them 'vertical'. Any first geodesic PC explains almost two-thirds of geodesic variance in the mixed sense and more than half of the variance explained by residuals. In the case of variance explained by projection, the roles are reversed and any second geodesic PC explains a little more than half of the data variance.

6.5. Example 2: Rat cranium growth

Finally, we consider the well-known Bookstein 'rat data' (see [1, pp. 408ff.] and, e.g. [17]), eight landmarks of 18 rat skulls measured on eight different days in their early lives: days 7, 14, 21, 30, 40, 60, 90, and 150. Working with triangles only we picked three out of the eight landmarks, namely landmarks 1, 5, and 6. Specifying different triples leads to qualitatively similar results. For every specimen we sought a first PC best approximating the data.

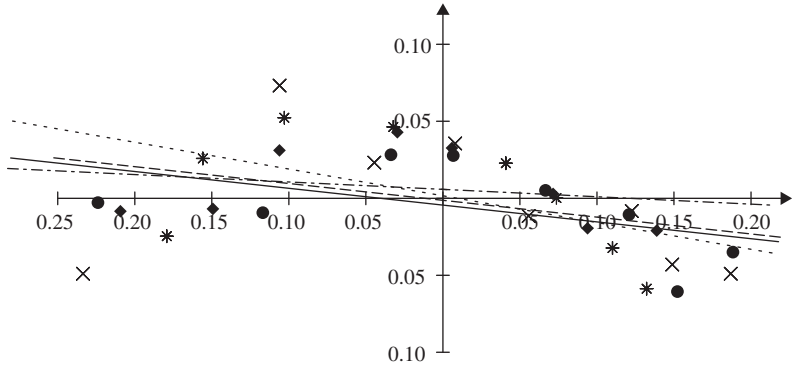

FIGURE 2: A simultaneous depiction of the first principal component geodesics to the temporal evolution of the shapes of the first four rats. The distances are measured in radians. Filled circles and the solid geodesic represent the first rat, crosses and the long-dashed geodesic represent the second rat, diamonds and the dash–dot geodesic represent the third rat, and asterisks and the short-dashed geodesic represent the fourth rat. All first PCs have approximately the same direction and, as they nearly pass through the origin, they are mapped to almost straight lines.

In Figure 2 we have simultaneously depicted the shapes of several rats (each having eight shapes) using their proper first principal component geodesics, calculated using our method. The image is a projection of shape space onto tangent space under the inverse Riemannian exponential taken at the intrinsic mean over all principal component geodesic means of the 18 rats. During growth, the rat shape data move from the left-hand side of the figure to the right-hand side. Obviously, the first PCs of the four rats point in approximately the same direction. Such a vizualisation is possible only for triangular shapes.

The data cover only a very small portion of the shape sphere (the scaling in Figure 2 along radial rays from the origin is precisely shape distance; the maximum possible shape distance is $\pi/2$). On average, the amount of variance explained by the first Euclidean PCs is 95.48%, whereas going by the mixed geodesic definition explains 95.52%, and geodesic first PCs fit the data about 0.8% better than do Euclidean PCs.

Upon closer visual inspection it seems that the rat data follow a parabola rather than a line. We might pursue this observation in a later paper.

Acknowledgements

The numerics were carried out with the computer algebra tool MuPAD. Many thanks go to Walter Oevel at the MuPAD group for suggesting the above algorithmic approach. We would also like to thank the two referees for their instructive comments, which led to considerable improvements to the paper.

References

[1] BOOKSTEIN, F. L. (1991). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

[2] COOTES, T. F., TAYLOR, C. J., COOPER, D. H. AND GRAHAM, J. (1992). Training models of shape from sets of examples. In *Proc. British Mach. Vision Conf.*, eds D. C. Hogg and R. D. Boyle, Springer, Berlin, pp. 9–18.

[3] DRYDEN, I. L. AND MARDIA, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester.

[4] DUCHAM, T. AND STUETZLE W. (1996). Extremal properties of principal curves in the plane. *Ann. Statist.* **24,** 1511–1520.

[5] Fletcher, P. T., Joshi, S., Lu, C. and Pizer, S. (2004). Gaussian distributions on Lie groups and their applications to statistical shape analysis. Preprint.
[6] Fletcher, P. T., Lu, C. and Joshi, S. (2003). Statistics of shapes via principal geodesic analysis on Lie groups. *Proc. Computer Vision and Pattern Recognition 2003*, Vol. 1, IEEE, Piscataway, NJ, pp. 95–101.
[7] Goodall, C. R. and Lange N. (1998). Growth curve models for correlated triangular shapes. In *Proc. 21st INTERFACE Symp.*, Interface Foundation, Fairfax Station, VA, pp. 445–454.
[8] Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika* **40,** 33–51.
[9] Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **30,** 509–541.
[10] Kendall, D. G., Barden, D., Carne, T. K. and Le, H. (1999). *Shape and Shape Theory*. Wiley, Chichester.
[11] Kent, J. T. (1994). The complex Bingham distribution and shape analysis. *J. R. Statist. Soc. B* **56,** 285–299.
[12] Klassen, E., Srivastava, A., Mio, W. and Joshi, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intellig.* **26,** 372–383.
[13] Kobayashi, S. and Nomizu, K. (1969). *Foundations of Differential Geometry*, Vol. II. Wiley-Interscience, New York.
[14] Kume, A. and Le, H. (2003). On Fréchet means in simplex shape space. *Adv. Appl. Prob.* **35,** 885–897.
[15] Le, H. (2001). Locating Fréchet means with application to shape spaces. *Adv. Appl. Prob.* **33,** 324–338.
[16] Le, H. and Barden, D. (2001). On simplex shape spaces. *J. London Math. Soc.* **64,** 501–512.
[17] Le, H. and Kume, A. (2000). Detection of shape changes in biological features. *J. Microscopy* **200,** 140–147.
[18] Le, H. and Small, C. G. (1999). Multidimensional scaling of simplex shapes. *Pattern Recognition* **32,** 1601–1613.
[19] Nash, J. (1956). The imbedding problem for Riemannian manifolds. *Ann. Math.* **63,** 20–63.
[20] Small, C. G. (1996). *The Statistical Theory of Shape*. Springer, New York.
[21] Ziezold, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Trans. 7th Prague Conf. Inf. Theory, Statist. Decision Functions, Random Process.*, Vol. A, Reidel, Dordrecht, pp. 591–602.
[22] Ziezold, H. (1994). Mean figures and mean shapes applied to biological figure and shape distributions in the plane. *Biometrical J.* **36,** 491–510.