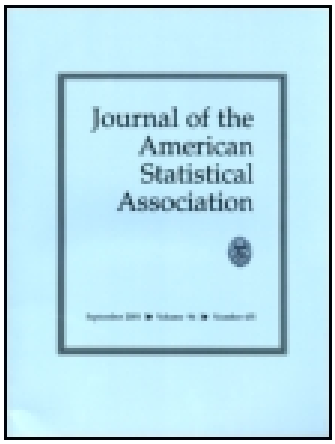


Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Principal Flows

Victor M. Panaretos, Tung Pham & Zhigang Yao

Accepted author version posted online: 14 Nov 2013.Published online: 19 Mar 2014.

To cite this article: Victor M. Panaretos, Tung Pham & Zhigang Yao (2014) Principal Flows, Journal of the American Statistical Association, 109:505, 424-436, DOI: 10.1080/01621459.2013.849199

To link to this article: http://dx.doi.org/10.1080/01621459.2013.849199

Principal Flows

Victor M. PANARETOS, Tung PHAM, and Zhigang YAO

We revisit the problem of extending the notion of principal component analysis (PCA) to multivariate datasets that satisfy nonlinear constraints, therefore lying on Riemannian manifolds. Our aim is to determine curves on the manifold that retain their canonical interpretability as principal components, while at the same time being flexible enough to capture nongeodesic forms of variation. We introduce the concept of a principal flow, a curve on the manifold passing through the mean of the data, and with the property that, at any point of the curve, the tangent velocity vector attempts to fit the first eigenvector of a tangent space PCA locally at that same point, subject to a smoothness constraint. That is, a particle flowing along the principal flow attempts to move along a path of maximal variation of the data, up to smoothness constraints. The rigorous definition of a principal flow is given by means of a Lagrangian variational problem, and its solution is reduced to an ODE problem via the Euler–Lagrange method. Conditions for existence and uniqueness are provided, and an algorithm is outlined for the numerical solution of the problem. Higher order principal flows are also defined. It is shown that global principal flows yield the usual principal components on a Euclidean space. By means of examples, it is illustrated that the principal flow is able to capture patterns of variation that can escape other manifold PCA methods.

KEY WORDS: Covariance tensor field; Euler–Lagrange method; Intrinsic mean; Manifold; Principal component analysis; Tangent space.

1. INTRODUCTION

Much of statistical methodology is deeply rooted in methods resting upon linearity, not necessarily in the sense that the methods are linear in themselves (though these play a prominent role), but in that they exploit the vector structure of the ambient sample space in a fundamental way. Though such sample spaces are able to successfully capture a tremendous variety of statistical problems, there still remain numerous contexts where the sample space is fundamentally nonlinear, and requires a differential-geometric treatment. Manifold sample spaces may arise in a number of ways, but one may isolate certain general themes such as:

1. The manifold being the actual physical space where the data reside. The classical example is that of the sphere, as a model for the surface of the earth in applications ranging from earth sciences (e.g., Watson 1983; Mardia and Jupp 2000) to marine biology (e.g., Brillinger 1997).
2. The manifold being a subset of a classical vector space that arises due to nonlinear constraints, for example, cones of nonnegative definite matrices (e.g., Schwartzman 2006) or Stiefel manifolds of ordered bases (e.g., Mardia and Khatri 1977).
3. The manifold being "isomorphic" (in the appropriate sense) to the representation space that arises as the result of the choice of a particular mathematical formalism to describe a certain data structure. For instance, complex projective spaces when representing planar shapes (e.g., Kendall et al. 1999), or cylinders when representing line processes (e.g., Stoyan, Kendall, and Mecke 1995).

Natural as manifold data may be, they can pose unconventional challenges. Even the simplest statistical procedures may not be well defined in a nonlinear manifold. For instance, the very notion of an average is not straightforward, and definitions of averages such as the Fréchet mean (Fréchet 1948) may suffer from nonuniqueness (Karcher 1977) and have intricate behavior under sampling, even asymptotically (Bhattacharya and Patrangenaru 2003, 2005; Kendall and Le 2011).

Further to describing location, one of course wishes to be able to describe the covariation of the data around their center, whether this may be a Fréchet mean or some other notion of centrality. By analogy to the vector space case, a natural means of doing so is considering how principal component analysis (PCA) methods can be developed on nonlinear manifolds. Methods that generalize PCA to the case of manifolds typically hinge on one of the various equivalent definitions of PCA in Euclidean space and proceed by analogy. For example, principal geodesic analysis and its variants use the fact that the first principal component in a linear space is the line passing through the mean with minimal average squared distance to the points of the sample, and define a corresponding first principal component on the manifold replacing the line by a geodesic (see, e.g., Huckemann and Ziezold 2006; Dryden, Koloydenko, and Zhou 2009; Huckemann, Hotz, and Munk 2010; Jung, Foskey, and Marron 2010). Tangent space PCA considers the interpretation of the first principal component as a direction of maximal variability centered at the mean and yields a vector in the tangent space of the mean (the *tangent principal component*) lying in the direction of maximal variability when the sample points are lifted onto this tangent space by projection or inverse exponentiation (see, e.g., Goodall 1991; Dryden and Mardia 1998, sec. 5.5; Fletcher and Joshi 2007). Using the fact that Euclidean principal components yield a sequence of nested linear subspaces, each providing the optimal embedding of the data in the corresponding dimension, Fletcher et al. (2004) defined a collection of nested geodesic submanifolds locally around the intrinsic mean, by retracting tangent principal components onto the manifold via the exponential map. In the special case of the sphere, Jung, Dryden, and Marron (2012) used the same nested subspace principle to motivate a sequence of nested spheres.

Victor M. Panaretos is Associate Professor (E-mail: *victor.panaretos@epfl.ch*), Tung Pham (E-mail: *tung.pham@epfl.ch*), and Zhigang Yao (E-mail: *zhigang.yao@epfl.ch*) are Postdoctoral Researchers, Section de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland. Research supported by an ERC Starting Grant Award. Our warm thanks go to two referees and the associate editor for a number of constructive comments.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
March 2014, Vol. 109, No. 505, Theory and Methods
DOI: 10.1080/01621459.2013.849199

Other approaches that pursue data reduction are less attached to a "canonical" PCA framework, and attempt to fit curves to the point cloud on the manifold (usually sequentially indexed), often attempting to minimize a sum of squared distances criterion. For instance, Jupp and Kent (1987) (in the case of spheres) and Kume, Dryden, and Le (2007) (in the case of complex projective spaces) unwrap the manifold data onto the tangent space at an appropriate point, fit smoothing splines to obtain a curve on the tangent space, and then wrap the curve back onto the manifold (note that if linear interpolation were used on the tangent space, then the final path obtained would be piecewise geodesic). In contrast, Su et al. (2012) introduced a roughness penalty directly on the manifold with the purpose of fitting "intrinsic smoothing splines." Kenobi, Dryden, and Le (2010) fitted polynomial-type curves on shape spaces, by using their special relationship with the so-called *preshape sphere*.

The first class of approaches has the advantage of being principled, in that they are interpretable through the lens of classical PCA. However, they are often interpretable only locally and can be rather rigid. Though the "curve fitting" approaches may not have as elegant or canonical an interpretation, they can often be superior to PCA-inspired approaches in capturing patterns of variation that are either nonlocal or complex enough to not be well described by geodesics. Nevertheless, they often require that the data points be indexed in some natural order.

The goal of this article is to find a compromise between these two schools of thought, one that will be interpretable more globally, while retaining the interpretability of principled PCA-inspired methods and allowing for the nongeodesic flexibility of curve fitting of manifold data (indeed data that are not even necessarily sequentially indexed). In brief, we consider the problem of defining a smooth curve on the manifold that passes through a given center of the data (e.g., an intrinsic mean), and with the property that, at each point, its derivative (which is tangent to the manifold) is "close" to the first principal component generated by a local tangent PCA at that same point (Section 2.3). In this sense, a flow along this curve always attempts to follow the direction of maximal variability, subject to a smoothness constraint. We call such a curve a principal flow. By varying the size of the neighborhood on which the local PCA yielding the direction of maximal local variability at each point is performed, one can control how rigid or flexible the principal flow will be.

We demonstrate how the problem of obtaining a principal flow can be transformed into one of solving a Euler–Lagrange problem on the manifold (a standard problem of numerical analysis), give conditions for the existence and uniqueness of a solution (Section 3) and describe an algorithm for its practical implementation (Section 3.1). We show that our method remains "canonical," in that it can be proven to reduce to the usual PCA when applied in a Euclidean space, with the usual mean as a starting point, and when the direction of maximal variability at any given point is defined globally (Section 5). Though the principal flow is primarily motivated by curved manifolds, it may also be useful in the case of flat manifolds: when the direction of highest variation around a given point is defined by emphasizing nearest neighbor points, the flow is no longer a line (the principal component line) but a smooth curve, providing a complementary notion to that of a *principal curve*, as introduced by Hastie and Stuetzle (1989).

We illustrate by means of simulated examples and the analysis of seismological data (Section 6) that the principal flow is able to capture patterns of variation that escape methods related to principal geodesics, when the direction of maximal variability at a given point is defined using neighboring points. In principle, our formulation may be implemented on any Riemannian manifold that is defined as the level set of a smooth function, provided that this latter function is explicitly known (though lack of explicit geodesics will certainly increase the computational complexity). This is usually not the case for more flexible curve fitting methods, which are typically studied in special cases of manifolds (and which, in addition, require that the data be indexed in some ordered fashion).

2. PRINCIPAL FLOWS

2.1 Preliminaries

We will throughout work within the context of a complete Riemannian manifold M of dimension r, embedded in $(\mathbb{R}^m, \|\cdot\|)$, $r < m$ (see, e.g., Thorpe (1979) for an accessible overview or Small (1996, sec. 2.2) for a rapid introduction). We will assume that the embedding is explicitly known, in that

$$M := \{x \in \mathbb{R}^m : F(x) = 0\}$$

for a known differentiable function $F : \mathbb{R}^m \to \mathbb{R}^r$. The corresponding bundle of tangent spaces will be denoted by

$$T_x M := \{y \in \mathbb{R}^m : DF\, y = 0\}, \quad x \in M$$

with associated metric tensor $\langle \cdot, \cdot \rangle$ and induced Riemannian metric $d(\cdot, \cdot)$. Here, DF is the $r \times m$ derivative matrix of F evaluated at $x \in M$, assumed to be of full rank everywhere on M. For each $x \in M$, let

$$\exp_x : T_x M \to M$$

be the exponential map, well defined away from the cut locus of x on $T_x M$, and

$$\log_x : M \to T_x M$$

be its inverse, well defined away from the cut locus of x on M. Throughout, $\{x_1, \ldots, x_n\}$ will denote a configuration of n points on M. We will assume that this configuration is localized enough that there exists a connected open set $B \subset M$ covering $\{x_1, \ldots, x_n\}$ and such that $\log_x y$ is well defined for all $x, y \in B$.

2.2 Tangent Space PCA

Before introducing the principal flow, we recall the basic elements of tangent space PCA (e.g., Dryden and Mardia 1998, sec. 5.5), which will play a role in the definition of the principal flow. Let $\{x_1, \ldots, x_n\} \in M$ and $B \subset M$ be as above, and $\bar{x} \in B$ be some other point on the manifold. The point $\bar{x}$ could be a Fréchet mean (Fréchet 1948), that is a minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} d^2(\cdot, x_i),$$

for d the Riemannian metric, or could be some other arbitrary point of interest. We assume that $n > r$, that is, the sample size is larger than the intrinsic dimension of the manifold.

Definition 2.1 (Tangent PCA). The tangent principal components at $\bar{x} \in B$, $\{e_1(\bar{x}), \ldots, e_r(\bar{x})\}$, are a basis for $T_{\bar{x}} M$, given

by the first r eigenvectors of the $m \times m$ matrix

$$\Sigma(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \log_{\bar{x}}(x_i) \otimes \log_{\bar{x}}(x_i), \tag{1}$$

where $y \otimes y := yy^\top$. The corresponding eigenvalues are denoted by $\lambda_1(\bar{x}) \geq \cdots \geq \lambda_n(\bar{x}) \geq 0$, so that:

$$\Sigma(\bar{x})e_1(\bar{x}) = \lambda_1(\bar{x})e_1(\bar{x}).$$

The condition $\bar{x} \in B$ is a technical one, ensuring that the log map is well defined. Note that only r eigenvalues will be nonzero, since $T_{\bar{x}}M$ is of intrinsic dimension r. Intuitively, we lift the configuration of points $\{x_1, \ldots, x_n\}$ from the manifold onto the tangent space $T_{\bar{x}}M$ using the log map, and carry out PCA on the tangent space, with respect to the covariance matrix of the lifted configuration around $\bar{x}$. The tangent principal components are now vectors in $T_{\bar{x}}M$ pointing at the principal directions of variation.

Local versions of the tangent covariance matrix may also be defined, by selecting a scale parameter $h > 0$ and a smooth nonincreasing univariate kernel K on $[0, \infty)$.

Definition 2.2 (Local Tangent Covariance). Letting $h > 0$ and $\kappa_h(x, \bar{x}) = K(h^{-1} \| \log_{\bar{x}} x - \bar{x} \|)$, we define a scale h local tangent covariance matrix at $\bar{x} \in B$ as

$$\Sigma_h(\bar{x}) = \frac{1}{\sum_i \kappa_h(x_i, \bar{x})}\sum_{i=1}^{n} (\log_{\bar{x}}(x_i) \otimes \log_{\bar{x}}(x_i))\kappa_h(x_i, \bar{x}). \tag{2}$$

Alternatively, one may use Mahalanobis tangent neighborhoods, defining $\kappa_h(x, \bar{x}) = K(h^{-1}(\log_{\bar{x}} x - \bar{x})^\top \Sigma^{-1}(\bar{x})(\log_{\bar{x}} x - \bar{x}))$. In the following, we will use the spherical neighborhoods, without loss of generality.

Such local versions can be used as a basis for a localized tangent PCA *at scale h*, since the curvature of the manifold may be such that the tangent space approximation might not yield a reasonable approximation for the entire sample. Setting $h = \infty$ yields the tangent covariance in Equation (1). When there is no danger of confusion, we will omit the subscript h from the local versions of the tangent eigenvectors and eigenvalues.

The tangent principal components $\{e_1(\bar{x}), \ldots, e_r(\bar{x})\}$ (or their local versions) yield the ordered directions of maximal variation (the respective variance components being $\{\lambda_1(\bar{x}), \ldots, \lambda_r(\bar{x})\}$) (correspondingly their local versions), but do not immediately yield a *path* of maximal variation. A local solution is to consider geodesics passing through $\bar{x}$ and tangent, at $\bar{x}$, to the corresponding principal components (such paths exist and are unique locally around $\bar{x}$). This is essentially the approach of Fletcher et al. (2004). The drawback is that these geodesics are interpretable as "paths of high variation" only for a small neighborhood around $\bar{x}$ (depending also on the geometry of the particular manifold). This motivates our definition of a principal flow, a path admitting a more global interpretation.

2.3 Definition of Principal Flows

The heuristic behind the principal flow is that, after we construct the principal tangent vector $e_1(\bar{x})$ at the starting point $\bar{x}$, we want to move infinitesimally in that direction. Once this infinitesimal move is made, it may no longer make sense to keep moving at this same direction (i.e., the geodesic tangent to the initial direction at the starting point $\bar{x}$). Instead, after the infinitesimal transition, we may again ask for the tangent direction $e_1(x')$ of highest variation at the current position x' and move accordingly. Iterating this procedure would approximately yield an integral curve that is always tangent to the direction of maximal variation at any given point it contains. In the terminology of classical mechanics, we will have a *flow* of the vector field $\{e_1(x)\}$ started at the point $\bar{x}$. The problem with this approach is that it remains intrinsically local in its updating rule, and thus greedy and subject to unstable variation under minor perturbations of the sample.

This section provides a principled construction of a path of maximal "cumulative variation" motivated by this heuristic, called a *principal flow*, corresponding to a concrete objective functional that can be interpreted precisely and has a global character (but is locally equivalent with the motivating heuristic). The strategy will be to pose the problem as one of Lagrangian mechanics and solve the corresponding variational problem by the Euler–Lagrange method.

Initially, we concentrate on the first principal flow and denote the first eigenvector (respectively, eigenvalue) of a matrix $\Sigma_h(x)$ simply by $e(x)$ (respectively, $\lambda(x)$), for tidiness. These define an orientation field $\{V(x) : x \in B\}$, comprised of orientations of "maximal variation,"

$$V(x) = \{-\lambda(x)e(x), \lambda(x)e(x)\}. \tag{3}$$

What will be the key for our later derivations is that, at least locally, this orientation field can be transformed into a vector field. That is, within an open neighborhood $N \subset B$, we can pick the eigenvectors $\{e(x) : x \in N\}$ to be "pointing in the same direction."

Proposition 2.1. Let $\{x_1, \ldots, x_n\} \subset M$ and $B \subset M$ be as above, and assume that the tangent covariance matrix $\Sigma_h(x)$ has distinct first and second eigenvalues for all $x \in B$. Given any $\bar{x} \in B$, the tangent vector $\lambda(\bar{x})e(\bar{x})$ can be extended to a tangent vector field $W := \{W(x) : x \in N(\bar{x})\}$ defined over an open neighborhood $N(\bar{x})$ of $\bar{x}$ and such that

1. $W(x)$ is independent of the local coordinates of T_xM and can be expressed in coordinates of the ambient space $\mathbb{R}^m$.
2. $\Sigma_h(x)W(x) = \lambda(x)W(x)$ (i.e., $W(x) \in V(x)$).
3. $W : N(\bar{x}) \to \mathbb{R}^m$ is a differentiable mapping.

With this result, we can now make sense of integrating over the orientation field given by the eigenvectors and may turn to describing the class of functions that will provide the candidates for a principal flow. For $\bar{x}$ a starting point on M, define the class

$$\begin{aligned}\Gamma(\bar{x}, v) = \{\gamma : [0, r] \to M :\ & \gamma \in C^2(M), \quad r \leq 1, \\ & \gamma(s) \neq \gamma(s') \text{ for } s \neq s' \\ & \gamma(0) = \bar{x}, \quad \dot{\gamma}(0) = v, \quad \ell(\gamma[0, t]) = t \\ & \text{for all} \quad 0 \leq t \leq r \leq 1\},\end{aligned}$$

where $\ell(\gamma[0, t])$ denotes the length of the parametric curve γ from $\gamma(0)$ to $\gamma(t)$, for all $0 \leq t \leq 1$, and v is a tangent unit vector at $\bar{x}$. That is, given $\bar{x}$ and a unit vector $v \in T_{\bar{x}}M$, $\Gamma(\bar{x}, v)$ is the set of all twice differentiable unit speed nonintersecting curves on M, starting at $\bar{x}$ with velocity vector v at $\bar{x}$, and length $r \leq 1$. The choice of 1 as an upper bound for the length is,

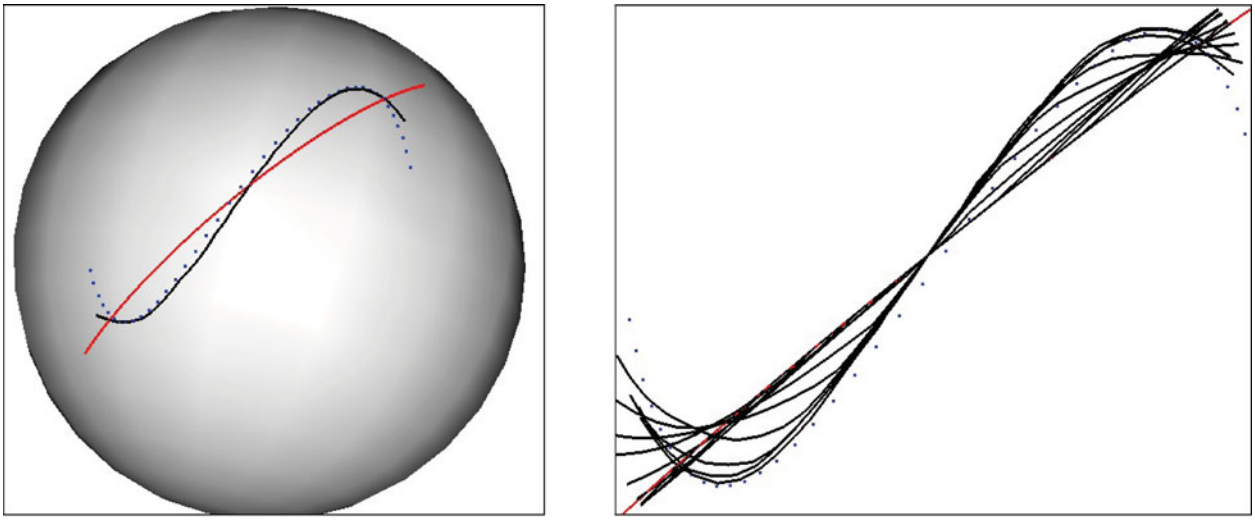


Figure 1. Principal flows for different values of the h parameter for an artificial configuration of points on the unit sphere (plotted in blue). The flows are started at the Fréchet mean (which obviously lies on the center of symmetry of the configuration). Panel (a) contrasts two flows generated by two choices of h. Panel (b) shows a magnification of the region of the sphere where the points lie, illustrating the effect of varying h on the form of the principal flow. (a) The red curve is the principal flow for $h = \infty$; the black curve is a principal flow for a reduced h. (b) The red curve is the principal flow for $h = \infty$; the 16 black curves are the principal flows induced by gradually decreasing the h parameter.

of course, arbitrary. It is chosen as a matter of convenience in later derivations and can be thought of as drawing an analogy to linear PCA, where eigenvectors have length 1. The length of a curve will be denoted as $\ell(\gamma)$.

The principal flow will now be defined as the solution of a variational problem: we will require the principal flow to be a smooth curve γ on the manifold, that passes through a given point $\bar{x}$ (e.g., a Fréchet mean) and whose derivative vector field $\dot{\gamma}$ is maximally compatible with the vector field $\{V(x) : x \in B\}$ of orientations of maximal variation.

Definition 2.3 (Principal Flow at Scale h). Let $\bar{x} \in B$. Assume that $\{\Sigma_h(x) : x \in B\}$ have distinct first and second eigenvalues. Let W be the extension of $\lambda(\bar{x})e(\bar{x})$ described in Proposition 2.1. A principal flow of $\{x_1, \ldots, x_n\}$ through $\bar{x}$ at scale h is the union of two curves γ_1 and γ_2 satisfying

$$\gamma_1 = \arg \sup_{\gamma \in \Gamma(\bar{x}, e(\bar{x}))} \int_0^{\ell(\gamma)} \langle \dot{\gamma}(t), W(\gamma(t)) \rangle dt, \tag{4}$$

$$\gamma_2 = \arg \inf_{\gamma \in \Gamma(\bar{x}, -e(\bar{x}))} \int_0^{\ell(\gamma)} \langle \dot{\gamma}(t), W(\gamma(t)) \rangle dt. \tag{5}$$

In light of Proposition 2.1, when $\{\Sigma_h(x) : x \in B\}$ have distinct first and second eigenvalues, the curve γ_1 starts at $\bar{x}$ and follows the direction of the vector field extending $\lambda(\bar{x})e(\bar{x})$, while the curve γ_2 starts at $\bar{x}$ and follows the opposite direction, $-\lambda(\bar{x})e(\bar{x})$ (which explains why the integral for γ_2 is negative and why the infimum appears: the integrand is always negative, whereas the integrand for γ_1 is always positive). The definition of γ_1 and γ_2 also explains the unit speed constraint: since the integrals defining them are line integrals, they are invariant under reparameterization, and hence a constraint would be required for the principal flow to be uniquely defined.

From a classical mechanics perspective, if the vector field W is thought of as a force field, then the principal flow provides a smooth path that maximizes the *work* done by the force field on a particle traveling along it.

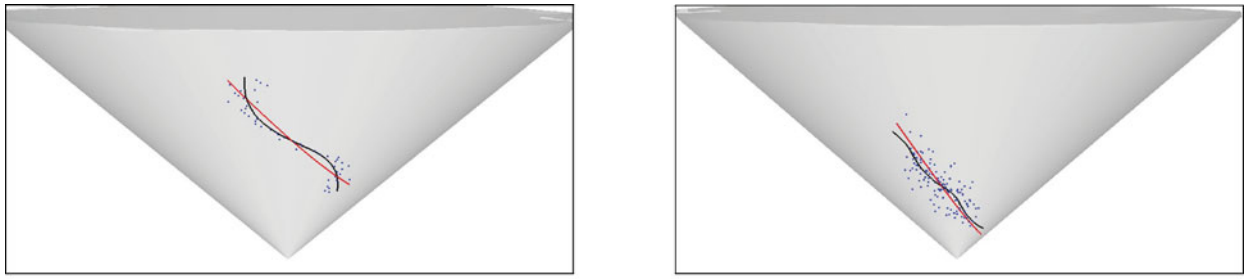
The scale parameter h controls how rigid/flexible the flow will be. For large values of h, the flow will be insensitive to local variation of the sample, and will focus on the global variation properties, resembling more rigid approaches such as principal geodesics. For small values of h, the flow will be more accommodating of local variation, and will capture variation properties at finer scales, resembling more flexible approaches such as polynomial shape curves (Figure 1 provides an example illustrating the behavior of the principal flow as h varies, while Figure 2 gives an example of two flows on the surface of a cone). What is important, though, is that in either case the flow will admit a clear principled PCA-type interpretation. The parameter h need not be interpreted as a bandwidth parameter that needs to be optimized over. Rather, we take a scale-space perspective (Chaudhuri and Marron 2000) and view h as characterizing how locally or globally we wish to describe a path of maximal variation. The behavior of the principal flow induced for different values of h is illustrated in Section 6, and further discussion on the issue is provided in Section 7.

3. DETERMINATION OF PRINCIPAL FLOWS VIA THE EULER–LAGRANGE METHOD

By symmetry, we restrict the discussion to the solution of the optimization problem yielding the curve γ_1. This variational problem is of the Euler–Lagrange form with mixed boundary conditions. The problem is to find a critical point of

$$\mathscr{L}_1(W, \gamma) = \int_0^{\ell(\gamma)} \langle \dot{\gamma}(t), W(\gamma(t)) \rangle dt \tag{6}$$

subject to the constraints defining the set of candidate curves $\Gamma(\bar{x}, W(\bar{x}))$, that is



(a) (b)

Figure 2. Principal flows for different values of the h parameter for two configurations of points on a cone (plotted in blue). In each case, the flows are started at the Fréchet mean. Panels (a) and (b) contrast flows generated by two choices of h for a configuration of points concentrated around a smooth curve and for a diffuse configuration of points, respectively.

1. $\|\dot\gamma(t)\| = 1, \forall t \in [0, r]$,
2. $\int_0^r \|\dot\gamma(t)\|dt = r$,
3. $F(\gamma(t)) = 0, \forall t \in [0, r]$, that is the curve must lie on the manifold,
4. $\gamma(0) = \bar x$ and $\dot\gamma(0) = e(\bar x)$,
5. $\gamma(s) \neq \gamma(s')$ for $s \neq s'$.

Of these constraints, (1) and (2) are awkward to work with (they have been imposed to avoid nonuniqueness issues due to reparameterization). We will use the fact that $\mathscr{L}_1(W, \gamma)$ does not depend on the parameterization of γ, to replace these constraints with constraints that are easier to handle, reformulating the solution set as

$$\Gamma_2(\bar x, e(\bar x)) = \left\{\gamma : [0,1] \to M, \quad \gamma(0) = \bar x,\ \dot\gamma(0) = c_1 e(\bar x), \quad c_1 \in \mathbb{R}, \quad \gamma(s) \neq \gamma(s') \text{ for } s \neq s', \quad \int_0^1 \|\dot\gamma(t)\|^2 dt = 1\right\}.$$

Provided we can show that $\Gamma(\bar x, W(\bar x))$ and $\Gamma_2(\bar x, e(\bar x))$ are equivalent up to reparameterization, it will follow that we may solve the variational problem by replacing Γ by Γ_2. This is indeed the case as is stated below (and proven in the Appendix).

Proposition 3.1 (Reparameterization). The sets $\Gamma(\bar x, W(\bar x))$ and $\Gamma_2(\bar x, e(\bar x))$ are equal modulo curve reparameterization.

We now concentrate on finding a critical point of $\mathscr{L}_1(W, \gamma)$ over the set $\Gamma_2(\bar x, \lambda(\bar x)e(\bar x))$. Using Lagrange multipliers to define the Lagrangian

$$\mathscr{L}(t, \gamma, \dot\gamma) = \langle W(\gamma), \dot\gamma\rangle + \delta[\langle\dot\gamma, \dot\gamma\rangle - 1] + zF(\gamma), \tag{7}$$

our original problem may be restated as finding extreme points $\gamma \in C^2([0,1], \mathbb{R}^m)$ of

$$\int_0^1 \mathscr{L}(t, \gamma, \dot\gamma)dt \tag{8}$$

subject to the boundary conditions $\gamma(0) = \bar x,\ \dot\gamma(0) = c_1 e(\bar x)$, $c_1 \in \mathbb{R}$. Following the standard steps of the Euler–Lagrange method, we calculate

$$\frac{\partial\mathscr{L}}{\dot\gamma} = W(\gamma) + 2\delta\dot\gamma \tag{9}$$

$$\frac{\partial\mathscr{L}}{\gamma} = \dot W(\gamma)^\top\dot\gamma + DF(\gamma)^\top z \tag{10}$$

$$\frac{d}{dt}\frac{\partial\mathscr{L}}{\dot\gamma} = \dot W(\gamma)\dot\gamma + 2\delta\ddot\gamma, \tag{11}$$

and, equating the last two of the above equations, we obtain the Euler–Lagrange equation

$$\big[\dot W(\gamma)^\top - \dot W(\gamma)\big]\dot\gamma - 2\delta\ddot\gamma + DF(\gamma)^\top z = 0. \tag{12}$$

This can be rewritten in autonomous form as

$$\begin{cases} -2\delta\ddot\gamma + DF(\gamma)^\top z = [\dot W(\gamma) - \dot W(\gamma)^\top]\dot\gamma, \\ F(\gamma) = 0, \end{cases} \tag{13}$$

thus reducing the problem of determining the principal flow in practice, to the numerical solution of a system of ordinary differential equations. By defining $Q(\gamma) = -2\delta$ and $G(\gamma, \dot\gamma) = [\dot W(\gamma) - \dot W(\gamma)^\top]\dot\gamma$, we see that the Euler–Lagrange equations are of the form

$$\begin{cases} Q(\gamma)\ddot\gamma + DF(\gamma)^\top z = G(\gamma, \dot\gamma), \\ F(\gamma) = 0. \end{cases} \tag{14}$$

A key result (Rabier and Rheinboldt 1995, theorem 2.2) that is important in our context is that, provided $Q \in C^1$, $F \in C^{2,1}$ and $G \in C^{0,1}$, the initial value problem consisting of Equation (14) and compatible initial conditions $\gamma(0) = \bar x \in M$ & $\dot\gamma(0) = w \in T_{\bar x}M$ admits a C^2 solution γ, that is unique on an open interval containing 0. In other words, if the constraint function F is twice differentiable with a Lipschitz continuous second derivative, and the derivative of the vector field W (whose existence is given in Proposition 2.1) is Lipschitz continuous, then a solution curve to the optimization problem (8) exists, and is in fact unique, up to a certain positive length.

3.1 Numerical Implementation

Solutions to initial value problems of the form (14) have been studied extensively in the numerical analysis literature. The implementation that we used follows the basic steps of Rabier and Rheinboldt (1995), whose numerical algorithm is particularly well-adapted to our setting. The algorithm requires the availability of local coordinate systems $\varphi_x : \mathbb{R}^m \to \mathscr{N}_x$ for open neighborhoods of points $x \in M$, and orthogonal matrices $U(x)$

providing bases for the tangent spaces $T_x M$. In principle, therefore, it may be applied to any Riemannian manifold that arises as a zero set of some smooth function F, using standard methods for the efficient determination of tangential local coordinate systems.

The essence of the approach is that γ is a solution to Equation (14) if and only if it is a solution to

$$\ddot{\gamma} = H(\gamma, \dot{\gamma}), \tag{15}$$

a second-order ODE on M (Rabier and Rheinboldt 1995, theorem 2.1). The function H is explicitly defined through the coefficients of the ODE and the function F. In our case, since $Q(x) = -2\delta$,

$$H(x, w) = -2\delta P(x)[G(x, w) + 2\delta \mathbb{V}_x(w, w)] + \mathbb{V}_x(w, w),$$

where $P(x)$ is the orthogonal projection onto $T_x M$ and $\mathbb{V}_x$ is the second fundamental tensor at $x \in M$ (a known object given the manifold M). The algorithm of Rabier and Rheinboldt (1995) solves Equation (15) using an iteration that employs the local coordinate representation given by $\{\varphi_x\}$. A high-level description of the algorithm is as follows.

Algorithm 1 Numerical Optimization of $\mathscr{L}_1$

(A) Choose a value $\delta < 0$ for the Lagrange multiplier (recall that $Q(x) = -2\delta$);
(B) Let $\bar{x}$ be a starting point on M and choose the first eigenvector $e(\bar{x})$ of the tangent covariance matrix at $\Sigma_h(\bar{x})$ as the initial direction of the curve; let $X_0 = \bar{x}$; let $S_0 = \{\bar{x}\}$
(C) Set $k = 0$;
(D) Find the orthogonal basis $U(X_k)$;
(E) Solve the ODE

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} v \\ U(X_k)^\top H\big(\varphi_{X_k}(u), D\varphi_{X_k}(u)v\big) \end{pmatrix}$$

numerically on $T_{X_k} M$ for the specific value of $\delta = -\frac{1}{2}Q(x)$, where u is expressed in $U(X_k)$ coordinates, obtaining a sequence of $(u_{k,1}, \ldots, u_{k,t}) \subset \varphi_{X_k}^{-1} \mathscr{N}_{X_k} \subset T_{X_k} M$;
(F) Store $x_{k,i} = \varphi_{X_k}(u_{k,i})$ for $1 \le i \le t$; let $X_{k+1} = x_{k,t}$; let $S_k = \{x_{k,i}\}_{i=1}^t$;
(G) If the geodesic length of the piecewise geodesic curve γ_k produced by geodesically interpolating the sequence $\{S_j\}_{j=0}^k$ exceeds length 1, stop and return $\gamma_\delta = \gamma_k$; otherwise, set $k \mapsto k + 1$ and go to step (D).

The algorithm will be executed for a range of choices of δ, yielding a collection of curves, γ_δ. The curve that will then be retained as the optimum of $\mathscr{L}_1$ will be the γ_{δ^*} such that

$$\gamma_{\delta^*} = \arg\sup_\delta \int_0^{\ell(\gamma_\delta)} \langle W(\gamma_\delta), \dot{\gamma}_\delta \rangle dt.$$

If the discretized flow γ_δ is represented as $\gamma_\delta = (z_1, \ldots, z_N)$, where $z_i \in M$, then the *length* of γ_δ and the *work* done by the field W on γ_δ are numerically approximated as

$$\ell(\gamma_\delta) := \sum_{i=1}^{N-1} d(z_i, z_{i+1}),$$

$$\int_0^{\ell(\gamma_\delta)} \langle W(\gamma_\delta), \dot{\gamma}_\delta \rangle dt := \sum_{i=1}^{N-1} \langle W(z_i), \log_{z_i}(z_{i+1}) \rangle,$$

where d represents the Riemannian metric on M.

The computational cost of the method is, in general, related to the generic cost of solving second order ODEs on manifolds (see also Rabier and Rheinboldt 1995). This depends not only on the manifold itself, but also on the form of the ODE which in our case is also influenced by the particular data structure (e.g., its curvature and/or diameter), the density of data points and the value of h (decreasing in h). The 17 flows depicted in Figure 1(b) took a total user CPU time of 2406 sec, corresponding to 2455 sec of real time (processor: 2.5 GHz Intel Core i5; memory: 8GB 1333 MHz DDR3). This means that computing each principal flow from about 40 data points exhibiting substantial curvature took 2.4 min on average. Computing the pairs of flows started at the intrinsic mean that are depicted in Figures 3(a), 3(d), and 3(g) took 2.88 min, 4.39 min, and 2.58 min of real time, respectively.

4. HIGHER ORDER PRINCIPAL FLOWS

Similarly with our definition of the first principal flow, we may define higher order principal flows with an analogous interpretation. By analogy to our earlier development, we now consider the kth eigenvector (respectively, eigenvalue) of the matrix $\Sigma_h(x)$, say $e_k(x)$ (resp. $\lambda_k(x)$) and define the vector field W_k analogously to W, as defined in Proposition 2.1. Then, assuming that the first $k + 1$ eigenvalues of $\Sigma_h(x)$ are distinct, Proposition 2.1 remains valid when applied to $\lambda_k(\bar{x})e_k(\bar{x})$, so that we may define:

Definition 4.1 (kth Principal Flow at Scale h). Let $\bar{x} \in B$. Assume that the first $k + 1$ eigenvalues of $\Sigma_h(x)$ are distinct. Let $\{\lambda_k(x)\}$ and $\{e_k(x)\}$ be the corresponding kth eigenvalues and eigenvectors, and let W_k be the extension of $\lambda_k(\bar{x})e_k(\bar{x})$ as described in Proposition 2.1. The kth principal flow of $\{x_1, \ldots, x_n\}$ through $\bar{x}$ at scale h is the union of two curves γ_1 and γ_2 satisfying

$$\gamma_1 = \arg\sup_{\gamma \in \Gamma(\bar{x}, e_k(\bar{x}))} \int_0^{\ell(\gamma)} \langle \dot{\gamma}(t), W_k(\gamma(t)) \rangle dt, \tag{16}$$

$$\gamma_2 = \arg\inf_{\gamma \in \Gamma(\bar{x}, -e_k(\bar{x}))} \int_0^{\ell(\gamma)} \langle \dot{\gamma}(t), W_k(\gamma(t)) \rangle dt. \tag{17}$$

The determination of the kth principal flow will follow the exact same steps as that of the first principal flow, by means of the Euler–Lagrange method. The interpretation, however, will require a bit more care. For example, consider the case of the second principal flow with $h = \infty$. The heuristic behind Definition (16) with $k = 2$ would be that we search for a smooth curve $\gamma_1(t)$ that passes through $\bar{x}$ and attempts to be normal at each point $\gamma_1(t)$ to the direction of maximal variation at that point, $e_1(\gamma(t))$, while being tangent to the direction of maximal variation in the orthogonal complement of $e_1(\gamma(t))$, that is, $e_2(\gamma(t))$. While t is such that $\gamma_1(t)$ remains relatively close to

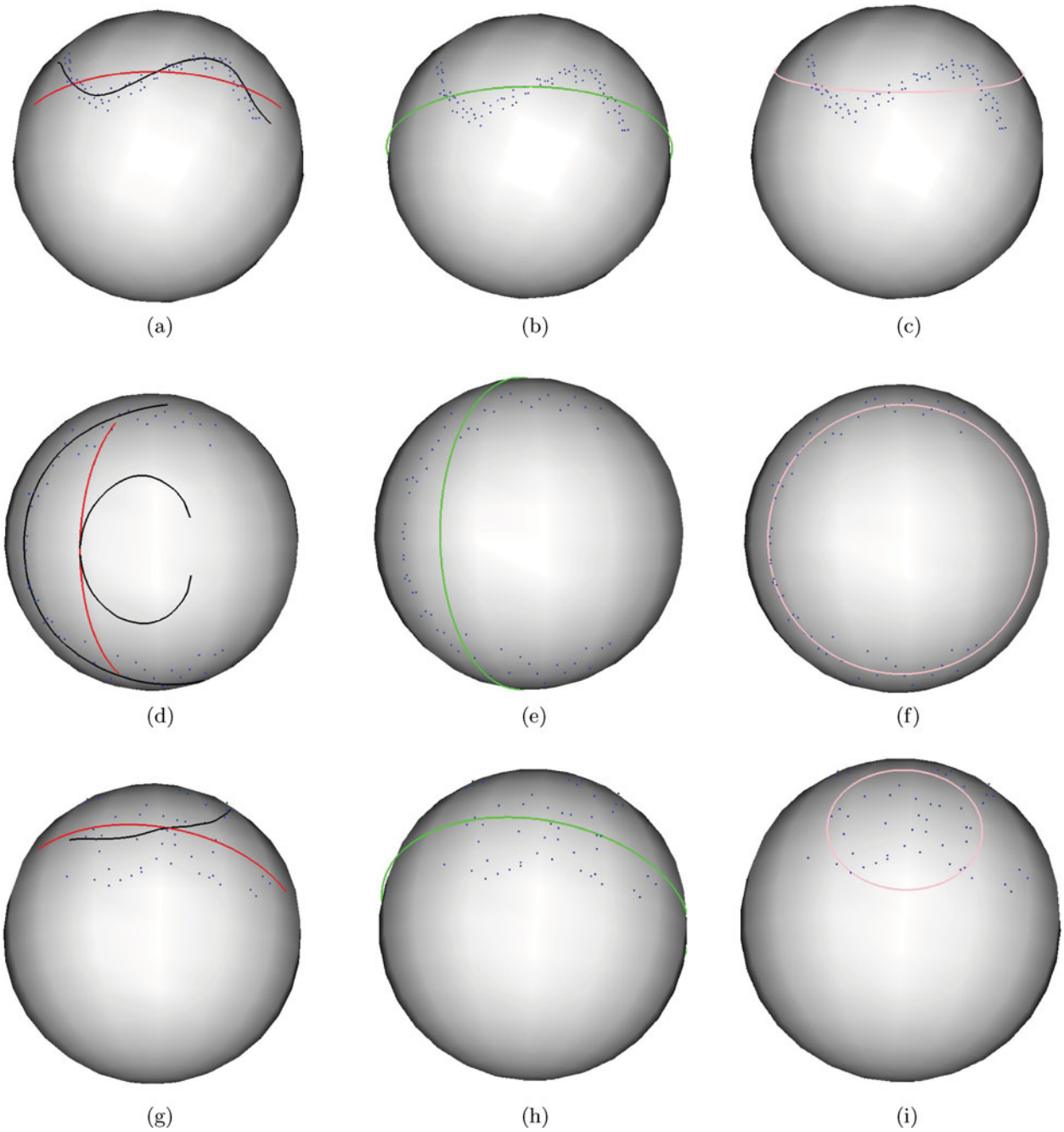


Figure 3. Principal flows, principal geodesics and principal nested spheres for three different sets of data on the unit sphere. (a) Principal flows for $h = \infty$ (red) and $h < \infty$ (black) started at the Fréchet mean. (b) Principal geodesic (principal great circle). (c) Principal nested Sphere (potentially a small circle). (d) Principal flows for $h = \infty$ (started at the mean; in red) and $h < \infty$ (started at the center of the point cloud; in black). (e) Principal geodesic (principal great circle). (f) Principal nested sphere (potentially a small circle). (g) Principal flows for $h = \infty$ (red) and $h < \infty$ (black) started at the Fréchet mean. (h) Principal geodesic (principal great circle). (i) principal nested sphere (potentially a small circle).

$\bar{x}$, it makes sense to attempt to move orthogonally to $e_1(\gamma(t))$. However, beyond some distance from $\bar{x}$ (i.e., for some t large enough), it is no longer clear why one should still attempt to move orthogonally to $e_1(\gamma(t))$. In other words, the canonical interpretation of a kth principal flow seems to remain valid only locally. A precise quantification of "local" will be derived in the next paragraph. In brief, for $h = \infty$, the canonical interpretation of the kth principal flow is retained ($k \geq 2$), provided that we restrict our attention to its segment that is contained in the ball of radius $\sqrt{\lambda_{k-1}(\bar{x}) - \lambda_k(\bar{x})}$ around $\bar{x}$.

5. PRINCIPAL FLOWS IN EUCLIDEAN SPACE

The purpose of this section is to prove that principal flows are canonical, in the sense that they reduce to the usual principal components, locally around the mean, in the context of Euclidean spaces. By making precise the notion of "locally," we also provide a means of interpreting the principal flows of order greater than 1. Suppose that the manifold M is, in fact, a linear subspace of $\mathbb{R}^m$, and that $x_1, x_2, \ldots, x_n$ constitute a configuration of points in M. Let $\bar{x}$ denote

the corresponding mean vector. Initially we will take $h = \infty$.

The proposition below shows that the first principal flow on a flat space, coincides with the first principal direction.

Proposition 5.1 Let $\{\Sigma_h(x) : x \in B\}$, $\{\lambda(x) : x \in B\}$ and $\{e(x) : x \in B\}$ be as in Definition 2.3. Let γ_* be a segment on M such that

$$\gamma_*(0) = \bar{x} \quad \text{and} \quad \gamma_*(t) = \gamma_*(0) + te(\bar{x}), \qquad t \in [0, 1].$$

and let

$$\gamma_1(t) = \gamma_*(t), \qquad \forall\, t \in [0, 1]$$
$$\gamma_2(t) = -\gamma_*(t), \qquad \forall\, t \in [0, 1].$$

Then, γ_1 and γ_2 are the unique optima of the functionals (4) and (5), as defined in Definition (2.3), for $h = \infty$. That is, the unique principal flow is the line segment with starting point $e(\bar{x})$ and endpoint $-e(\bar{x})$.

For higher order principal flows, the result remains true, at least locally around the mean. The locality is not infinitesimal. Rather, it is governed by the corresponding eigenvalue spacings of the covariance matrix $\Sigma(\bar{x})$. Recall here that $e_k(\bar{x})$ and $\lambda_k(\bar{x})$ denote the kth eigenvector and eigenvalue of $\Sigma(\bar{x})$, respectively.

Proposition 5.2 Let $\{\Sigma_h(x) : x \in B\}$, $\{\lambda_k(x) : x \in B\}$ and $\{e_k(x) : x \in B\}$ be as in Definition 4.1, with $k \geq 2$. Let γ_* be a segment on M such that

$$\gamma_*(0) = \bar{x} \quad \text{and} \quad \gamma_*(t) = \gamma_*(0) + te_k(\bar{x}), \qquad t \in [0, \sqrt{\lambda_{k-1}(\bar{x}) - \lambda_k(\bar{x})}].$$

Then, any optima γ_1 and γ_2 of the functionals 16 and 17 in Definition (4.1), must satisfy

$$\gamma_1(t) = \gamma_*(t), \qquad \forall\, t \in [0, \sqrt{\lambda_{k-1}(\bar{x}) - \lambda_k(\bar{x})}]$$
$$\gamma_2(t) = -\gamma_*(t), \qquad \forall\, t \in [0, \sqrt{\lambda_{k-1}(\bar{x}) - \lambda_k(\bar{x})}].$$

when h is taken to be ∞.

When $h < \infty$, we are potentially using local versions of the tangent covariance matrices, $\Sigma_h(x)$ and the principal flows on a flat manifold M will no longer necessarily reduce to the principal components, thus no longer admitting a global PCA interpretation. Nevertheless, they do retain the local interpretation that motivated them in the case of curved M, which has a potential to be of interest even in the case of linear spaces. It provides a compromise between the somewhat rigid notion of a principal component and the more flexible notion of a principal curve (Hastie and Stuetzle 1989): a curve that is flexible enough to capture nonlinear variation, but nevertheless retains the interpretation of being a "local principal component" at each one of its points. Therefore, for a flat manifold, principal flows can be viewed as a counterpart to the notion of a principal curve—one that is motivated by maximal variation considerations rather than optimal one-dimensional reductions.

6. EXAMPLES

6.1 Simulated Data

To probe the behavior of the principal flow as dependent on the configuration of the dataset, the starting point, and the choice of scale parameter h, we considered three examples on the sphere (Figure 5). We chose the sphere as a "test manifold" since it represents one of the most commonly encountered sample spaces for manifold data, and since it provides a manifold for which we can compare the principal flow with other notions of principal components. We contrast our results to those obtained when using principal geodesics (principal great circles) and principal spheres (which potentially yield small circles); see and Jung, Dryden, and Marron (2012). We remark here that the flow is not defined through a least-squares principle; regardless of the value of h, the flow is interpreted as a curve whose derivative approximately points at the direction of maximum local variation—how local being regulated by the value of h. On the other hand, principal circles are motivated through a least-squares objective function, attempting to minimize the overall squared lengths of geodesic projections. Since the principal flow and the principal circles correspond to optima of different objective functions, we do not compare them on a quantitative basis, but rather we contrast them qualitatively. Throughout, we used $K(y) = \mathbf{1}\{y \leq 1\}$, $y \geq 0$ as the kernel for the local covariance.

We generated three datasets representing different types of manifold variation. The first dataset is such that the point cloud is concentrated around an "S"-like curve on the manifold. The data thus present a global variation pattern (along the geodesic connecting the tip and the bottom of the "S") as well as a more localized variation pattern, along the "S"-like curve itself. The symmetry of the "S" means that the Fréchet mean represents a good starting point for the principal flow. On the other hand, the fact that its curvature is nonconstant means that the local variation cannot be well explained by an arc. The configuration of the second dataset is also such that the point cloud is concentrated around a curve, this time a curve resembling a "C" (see Figure 3(d)). Once again, there are two patterns of variation, a large scale one (along the normal to the axis of symmetry of the "C" shape) and a local one, along the "C"-like curve itself. The sign of the curvature is constant in this case. This time, however, the Fréchet mean is not a good descriptor of the center of the data. The last configuration is more diffuse: the point cloud lies on the northern hemisphere and is characterized by a very weak global elongation pattern, not being obviously concentrated around a particular smooth curve.

For the S-pattern, we fitted the principal flow through the Fréchet mean, using $h = \infty$ and a reduced value for h (Figure 3(a)). The results indicate that the corresponding flows perform well in capturing the global and local variation patterns, respectively. The flow with $h = \infty$ seems very similar to the principal great circle (Figure 3(b)). The principal nested sphere yields a small sphere in this case (Figure 3(c)), which also performs well in capturing the global structure, though with the opposite curvature. We note that neither the great nor the small sphere can capture the finer local variation structure. However, a principal flow with a reduced value for h seems to describe this finer structure with success.

In the case of the "C"-pattern, the Fréchet mean is no longer a good descriptor of the center of the dataset. A flow started at this mean with $h = \infty$ (Figure 3(d)) seems to capture the global elongation of the data (along the normal to the axis of symmetry of the "C"), and indeed appears to have a structure very similar to the principal great circle (Figure 3(e)). Arguably, though, it is the

principal nested small circle that provides the best description of the variation of the data in this case (Figure 3(f)). This example shows the impact of the choice of the starting point on the principal flow. If the flow is started at a more natural center of the data cloud, it can be seen to perform as well as the small circle in terms of capturing the variation pattern of the data (Figure 3(d) shows such a flow with a reduced choice of h).

When the data are more diffuse, one expects that a large value of h will yield a principal flow that is more sensible. The principal flow for $h = \infty$ (started at the mean) provides very similar results with the principal great circle (Figures 3(g) and 3(h)). Using a smaller h appears to yield results that are not as appealing. Though the flow still is interpretable as a path of maximal local variation, the fact that the data are diffuse means that looking at local scales does not adequately portray the structure of the configuration. The principle nested sphere (which in this case yields a small circle) also appears to be hard to interpret in this case (Figure 3(i)).

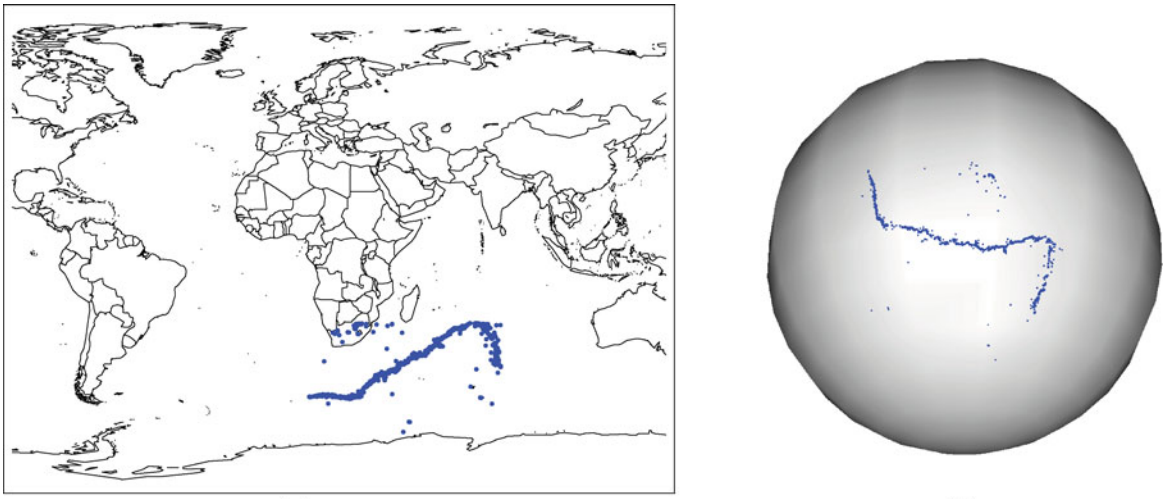
6.2 Seismological Data

To illustrate the use of the principal flow in a concrete example, we consider a spherical dataset comprised of earthquake epicenters. Figure 4(a) displays the epicenters of earthquakes that occurred between January 1, 1995 and May 28, 2013 within a large region around the meet of the African, Antarctic, and Australian tectonic plates (from -26° to -70° latitude, and from 10° to 80° longitude, negative numbers denoting latitudes in the southern hemisphere) as recorded by the U.S. Geological Survey (U.S. Geological Survey (in progress)). Though Figure 4(a) represents the points on a flat atlas, the region considered is large enough that the curvature of the Earth should be taken into consideration (see Figure 4(b)).

As can be expected, the earthquakes tend to concentrate around the tectonic plate boundaries, which provide a natural description of an underlying one-dimensional pattern carrying most of the variation of the point cloud. The shape of the plate boundary in this particular geographical region exhibits nonconstant curvature and thus manifests localized variation patterns (not unlike the "S" pattern in Figure 3(a)) in addition to the large-scale geodesic variation from southwest to northeast. This dual-scale variation pattern of the data is reflected on the behavior of their principal flow, depending on the choice to the scale parameter h. Using $K(y) = \mathbf{1}\{y \leq 1\}$ as the kernel for the local covariance again, and starting the flow at a point within the main cloud of the data, we can see that more local covariance fields (e.g., with $h = 400$ miles) generate a flow that tends to follow the main pattern of variation very closely, almost outlining the tectonic plate boundaries (Figure 5(a)). Moving to a larger scale of covariance field ($h = 800$ miles) yields a flow that attempts to compromise the local and global patters of variation (Figure 5(b)). The case $h = \infty$ yields a rigid curve that describes solely the large-scale pattern of variation (Figure 5(c)). All three flows are depicted in superposition in Figure 5(g).

This particular dataset can also serve to illustrate the effect that a poor choice of starting point may have on the success (or lack of it) of the principal flow in capturing the main pattern of variation in the data. In particular, the dataset features a number of outlying points that are located far from the main cloud in a direction essentially "orthogonal" to the plate boundary. As with linear PCA, these outliers can have a detrimental effect by: (a) moving the (intrinsic) mean away from the true center of the point cloud (and thus producing spurious directions of variation); (b) interchanging the first and second main directions of variation (if the outliers fall on a (geodesic) axis "orthogonal" to the main direction of variation); (c) a combination of both.

These effects are apparent in the case of the principal flow. In these cases, when h is relatively small (400 miles), the principal flow is pulled away from the bulk of the data, since the local covariance pattern at that scale is orthogonal to the more global pattern, at least in one direction (Figure 5(d)). On the other hand, doubling h (800 miles) leads to a less pronounced effect, but still the flow is struggling to compromise the two variation patterns: the one formed by the bulk of the data, and the one appearing due to the outliers (Figure 5(e)). When $h = \infty$, the global pattern of variation prevails, and the principal flow of the data provides a more sensible description of the data variation. All three flows are depicted in superposition in Figure 5(h).



(a) (b)

Figure 4. Earthquake epicenter data (coordinates courtesy of the U.S. Geological Survey U.S. Geological Survey (in progress)). (a) Earthquake epicenters (in blue) on a flat world atlas. (b) Earthquake epicenters (in blue) plotted on a globe.

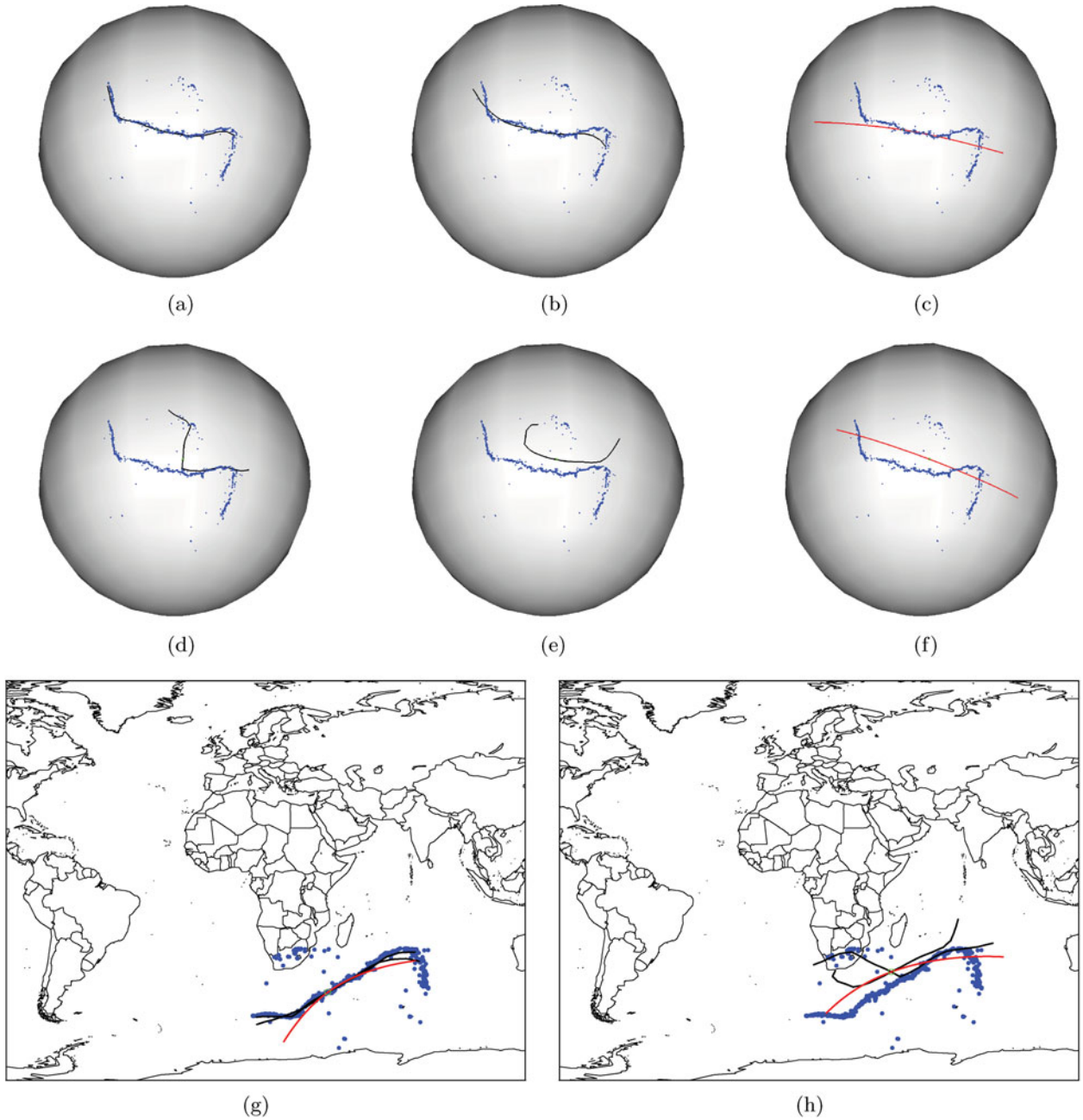


Figure 5. Principal flows of the earthquake data with different starting points and different scale parameters. (a) Principal flow for $h = 400$ miles (in black), started at center of symmetry (in green). (b) Principal flow for $h = 800$ miles (in black), started at center of symmetry (in green). (c) Principal flow for $h = \infty$ (in red), started at center of symmetry (in green). (d) Principal flow for $h = 400$ miles (in black), started at the Fréchet mean (in green). (e) Principal flow for $h = 800$ miles (in black), started at the Fréchet mean (in green). (f) Principal flow for $h = \infty$ (in red), started at the Fréchet mean (in green). (g) Superimposed principal flows for $h = 400$ miles, $h = 800$ miles, and $h = \infty$ (red) and $h < \infty$ (black) started at a center of symmetry (green). (h) Superimposed principal flows for $h = 400$ miles, $h = 800$ miles, and $h = \infty$ (red) and $h < \infty$ (black) started at the Fréchet mean (green).

7. DISCUSSION

With the aim of defining a manifold extension of principal components that retains its canonical PCA interpretation while allowing for more flexible reductions of nongeodesic patters of variation, we introduced the notion of a principal flow. This was seen to be interpretable as a curve passing through an intrinsic mean whose tangent velocity at any point attempts to fit the eigenvector of the corresponding tangent local covariance matrix at that point, subject to smoothness constraints. Depending on how locally or globally the tangent covariance matrix at any point was defined, the principal flow was seen to be more rigid or more flexible. In this sense, it can be thought of as a manifold extension of the principal curves introduced by Hastie and Stuetzle (1989). In fact, it was seen to reduce to PCA in the case of a Euclidean space, provided that the covariances employed were defined globally.

As is the case with principal components in Euclidean space, principal flows on a manifold require the determination of a

starting point. This can be taken to be an intrinsic mean or any other point of significance chosen by the researcher. However, similarly to Euclidean space PCA, when the starting point is not a good descriptor of location, a principal flow may not be a good descriptor of the variation of the sample. This points to an interesting direction for future work. Our principal flow was defined as the optimum of an Euler–Lagrange problem with mixed boundary conditions: the starting point and velocity of the curve were prescribed, but not the endpoint. An interesting question would be to study the free boundary version of the problem, where neither a starting nor an ending point of the curve is prescribed.

Depending on the choice of the scale parameter h, the principal flow captures local or global covariation features of the data. Our view is that h should not be seen as a bandwidth parameter to be optimized over, but rather that a whole spectrum of h should be considered in parallel, each h yielding a reduction of variation at a particular scale. Indeed, we claim that it does not strictly make sense to speak of an "optimal" h, since, depending on the choice of h, the principal flow attempts to describe a different orientation field (that generated by the first eigenspaces of the local covariance field $\{\Sigma_h(x) : x \in M\}$); so speaking of an optimal h is more about speaking of an optimal orientation field, a question that is rather subtle and requires a deeper study of the local tangent covariance field, seen as a process indexed by h.

Perhaps a more appropriate question, as one referee aptly remarked, is to ask: at which scale are the features of interest present? That is, at which scale is the principal flow picking up the most strongly one-dimensional variation? There are potentially several ways to quantify this statement depending on how one makes it precise. An informal approach followed in our illustrations was that of visual inspection. More objective criteria can be envisaged, for example one may conceive a criterion of the form:

$$R(h) = \int_a^b \frac{1}{\operatorname{trace}[\Sigma_h(\gamma(t))]} \sum_{i=1}^{n} \left\langle \dot{\gamma}_h(t), \log_{\gamma_h(t)} x_i \right\rangle^2 \times \frac{\kappa_h(x_i, \gamma(t))}{\sum_i \kappa_h(x_i, \gamma(t))} dt.$$

Here, $\gamma_h(t) : [a, b] \to M$ is the principal flow at scale h, under some curve parameterization and we recall that $\Sigma_h(x)$ is the local tangent covariance at $x \in M$. This criterion measures what percentage of the total tangent variance at scale h along the path of the flow is captured by the flow itself (it is most easily interpreted when $K(y) = \mathbf{1}\{y \le 1\}$). It is inspired by the usual approach to judging the success of the first principal component in capturing the total variation, by looking at the percentage of variance explained. Low values of the criterion could arise either when $\dot{\gamma}(t)$ is not close to being parallel to $e_1(\gamma(t))$ and/or

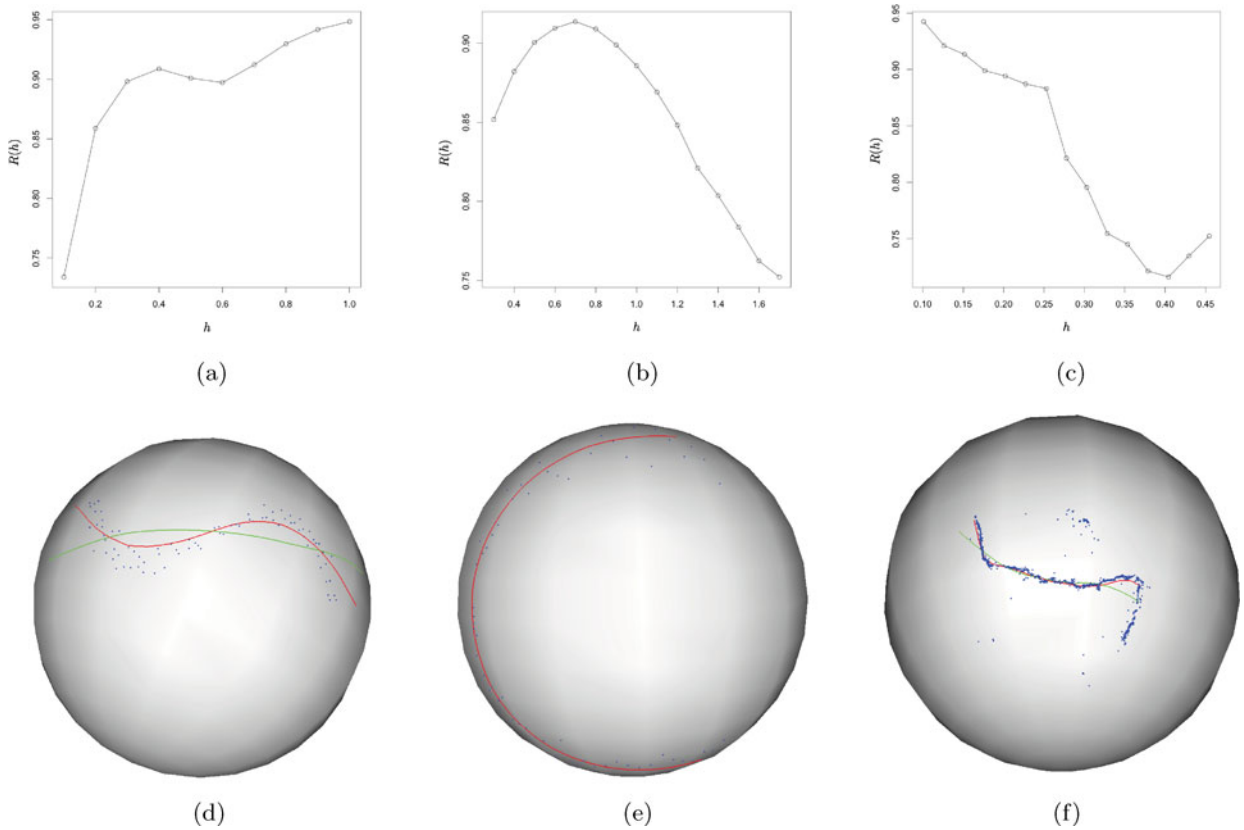


Figure 6. Plots of the R criterion (normalized so that $R \le 1$) and the flows corresponding to local maxima and/or elbows thereof for three different datasets. (a) Plot of $R(h)$ for the "S"-like data with the flows started at center of symmetry. (b) Plot of $R(h)$ for the "C"-like data with the flows started at center of symmetry. (c) Plot of $R(h)$ for the earthquake data with the flows started at center of symmetry. (d) Principal flows for the "S"-like data with $h = 0.4$ (in red) and $h = 1$ (in green), started at center of symmetry. (e) Principal flow for the "C"-like data with $h = 0.7$ (in red), started at center of symmetry. (f) Principal flows for the earthquake data with $h = 0.1$ (about 396 miles, plotted in red) and $h = 0.25$ (about 990 miles, plotted in green), started at center of symmetry.

because $\Sigma_h(\gamma(t))$ is not close to being of rank 1. Thus, R attempts to simultaneously measure how strongly one-directional the variation at scale h along the flow is, as well as how well the flow captures this variation. The maximum possible value would be attained at a certain h, if the data were such that all the h-local covariance tensors along the curve were precisely of rank 1 and with their first eigenvector tangent to the flow. Therefore, it must be that $0 \le R \le \ell$, where ℓ is the total length of the flow, which is fixed by construction.

In the examples we considered, values of h corresponding to local maxima and elbows of R yielded visually appealing flows, that seemed to capture particularly relevant features at different scales of variation (these are depicted in Figure 6, which also includes the corresponding plots of $R(h)$). In fact, we note that recent work on density estimation by Schmidt-Hieber, Munk, and Dümbgen (2013) shows that it is not even necessary to assume a known minimal scale h as was the case in Chaudhuri and Marron (2000): by allowing h to tend to zero at an appropriate rate, essentially all scales can be examined simultaneously (be it through R or another criterion), and the opposite extreme of letting $h \to 0$ instead of $h = \infty$ could be an interesting case in itself. Though R appears to be a useful means of guiding the practitioner's choice, it is certainly not the only choice and it seems that the issue of assessing which are the most informative scales h—whether by $R(\cdot)$ or another criterion—is one deserving further scrutiny, beyond the confines of the present article.

APPENDIX

Proofs of Formal Statements

Proof of Proposition 2.1. Let $y_i = \log_x(x_i)$, $x, x_i \in B$, and $\kappa_i = \kappa_h(x, x_i)$ and consider the set of linear operators from T_xM to T_xM, say $\mathrm{lin}(T_xM, T_xM)$. Depending on the coordinate system chosen for T_xM, $\sum_i (y_i \otimes y_i)\kappa_i$ has a different $r \times r$ matrix representation of an element in $\mathrm{lin}(T_xM, T_xM)$. Nevertheless, its action on T_xM is invariant to changes of the coordinate system, and hence its eigenvectors and eigenvalues are well defined and independent of the choice of local coordinates.

Next, we note that the map

$$T(x) := \sum_{i=1}^{n} (\log_x x_i \otimes \log_x x_i)\kappa_h(x, x_i) : \mathbb{R}^m \to \mathbb{R}^{m \times m} \qquad \text{(A.1)}$$

is differentiable on B. Given $x_c \in B$, let $N(X_c)$ be a neighborhood of $X_c = T(x_c)$ in $\mathbb{R}^{m \times m}$, the space of m by m matrices. For any $X \in N(X_c)$, let $\lambda(X)$ be the largest eigenvalue of X and $u(X)$ be the leading eigenvector of X. Following Magnus (1985), use of the implicit function theorem yields that λ and u may be differentiably extended on $N(X_c)$. Combining this fact with the differentiability of T, we conclude that the function $W = (\lambda \circ T) \times (u \circ T)$ is well defined and differentiable on an open neighborhood $N(x_c)$. □

Proof of Proposition 3.1. First we show that a curve in Γ may be reparameterized to yield a curve in Γ_2. Given $\gamma \in \Gamma(\bar{x}, W(\bar{x}))$, define

$$f : [0, 1] \to M, \quad f(t) = \gamma(rt), \text{ for all } 0 \le t \le 1.$$

Then $\|\dot{f}(t)\| = \|\dot{\gamma}(rt)\| r = r$ and so f is of length r. Furthermore,

$$\int_0^1 \|\dot{f}(t)\|^2 dt = r^2.$$

Now define

$$\gamma_* : [0, 1] \to M, \quad \gamma_*(t) = f(t^\alpha)$$

so that $\dot{\gamma}_*(t) = \alpha t^{\alpha-1} \dot{f}(t^\alpha)$. It follows that

$$\|\dot{\gamma}_*(t)\|^2 = \alpha^2 t^{2\alpha-2} \|\dot{f}(t^\alpha)\|^2 = \alpha^2 t^{2\alpha-2} r^2.$$

Integrating over both sides,

$$\int_0^1 \|\dot{\gamma}_*(t)\|^2 dt = \int_0^1 \alpha^2 t^{2\alpha-2} r^2 dr = \frac{\alpha^2}{2\alpha - 1} r^2.$$

Since $r \le 1$, we may choose $\alpha \ge 1$ such that $\alpha^2 r^2 = (2\alpha - 1)$, and so $\gamma_* \in \Gamma_2(\bar{x}, e(\bar{x}))$ provides the reparameterization of $\gamma \in \Gamma$.

To show that any curve $\gamma_2 \in \Gamma_2(\bar{x}, e(\bar{x}))$ can be reparameterized into a curve in $\Gamma(\bar{x}, W(\bar{x}))$, we observe that the Cauchy–Schwarz inequality yields

$$\left(\int_0^1 \|\dot{\gamma}_2\| dt \right)^2 \le \int_0^1 \|\dot{\gamma}_2\|^2 dt = 1.$$

This implies that the length of curve γ_2 is less than or equal to 1. Hence, we can reparameterize the curve γ_2 so that it have unit vector speed. □

Proof of Proposition 5.1. By symmetry, it suffices to prove that $\gamma_1(t) = \gamma_*(t)$ for all $t \in [0, 1]$. We begin by observing that, for any $x \in M$, we may write

$$\begin{aligned} \Sigma(x) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - x)(x_i - \bar{x} + \bar{x} - x)^\top \\ &= \Sigma(\bar{x}) + (x - \bar{x})(x - \bar{x})^\top. \end{aligned}$$

Clearly, the first eigenvector and eigenvalue of the rank 1 matrix $(x - \bar{x})(x - \bar{x})^\top$ are

$$\frac{(x - \bar{x})}{\|x - \bar{x}\|} \quad \text{and} \quad \|x - \bar{x}\|^2,$$

respectively. Therefore, using Weyl's inequality for the first eigenvalue of $\Sigma(x)$ and that of $\Sigma(\bar{x})$, we obtain

$$\lambda(x) \le \lambda(\bar{x}) + \|x - \bar{x}\|^2.$$

We note that the above inequality becomes an equality only when the first eigenvector of $\Sigma(\bar{x})$ coincides with the first eigenvector of $(x - \bar{x})(x - \bar{x})^\top$. Equivalently, we have equality if and only if x is on the line containing the first principal component of $\{x_1, \ldots, x_n\}$

Now, let $\gamma \in \Gamma_2$. Our inequality implies that

$$\lambda(\gamma(t)) \le \lambda(\bar{x}) + \|\gamma(t) - \bar{x}\|^2 \le \lambda(\bar{x}) + \|\gamma_*(t) - \bar{x}\|^2 = \lambda(\gamma_*(t)). \qquad \text{(A.2)}$$

Here, we have used the fact that

$$\|\gamma_*(t) - \bar{x}\| = \ell(\gamma_*[0, t]) = t = \ell(\gamma[0, t]) \ge \|\gamma(t) - \bar{x}\|,$$

for all $t \in [0, 1]$ and $\gamma \in \Gamma_2$, since γ_* is the segment of a straight line (traveling for a distance t along a straight line gets us further than traveling for a distance t along a curve). Since $|\dot{\gamma}(t)| = 1$ and $|e(\gamma(t))| = 1$, we obtain

$$\begin{aligned} |\langle \dot{\gamma}(t), \lambda(\gamma(t))e(\gamma(t)) \rangle| &\le \lambda(\gamma(t)) \le \lambda(\gamma_*(t)) \\ &= |\langle \dot{\gamma}_*(t), \lambda(\gamma_*(t))e(\gamma_*(t)) \rangle|, \end{aligned}$$

the last equality coming from the fact that $\dot{\gamma}_*(t)$ and $e(\gamma_*(t))$ are two parallel unit vectors.

Therefore,

$$\int_0^1 |\langle \dot{\gamma}(t), \lambda(\gamma(t))e(\gamma(t)) \rangle| dt \le \int_0^1 |\langle \dot{\gamma}_*(t), \lambda(\gamma_*(t))e(\gamma_*(t)) \rangle| dt.$$

Since the inequality is strict, unless the points $\{x_1 - \bar{x}, \ldots, x_n - \bar{x}\}$ are collinear, the proof is complete. □

Proof of Proposition 5.2. The proof is essentially identical to that of Proposition 5.1, where instead of the first eigenvalue, we apply Weyl's inequality for the kth eigenvalue of $\Sigma(x)$. All the remaining steps follow through in the same manner. Note that the restriction $t \in [0, \sqrt{\lambda_{k-1}(\bar{x}) - \lambda_k(\bar{x})}]$ is required to guarantee that the last equality in the sequence of inequalities (A.2) is preserved. □

[Received December 2012. Revised July 2013.]

REFERENCES

Bhattacharya, R., and Patrangenaru, V. (2003), "Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. I," *The Annals of Statistics*, 31, 1–29. [424]

——— (2005), "Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. II," *The Annals of Statistics*, 33, 1225–1259. [424]

Brillinger, D. R. (1997), "A Particle Migrating Randomly on the Sphere," *Journal of Theoretical Probability*, 10, 429–443. [424]

Chaudhuri, P., and Marron, J. S. (2000), "Scale Space View of Curve Estimation," *The Annals of Statistics*, 28, 408–428. [427,435]

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009), "Non-Euclidean Statistics for Covariance Matrices, With Applications to Diffusion Tensor Imaging," *Annals of Applied Statistics*, 3, 1102–1123. [424]

Dryden, I. L., and Mardia, K. V. (1998), *Statistical Shape Analysis*, New York: Wiley. [424,425]

Fletcher, P. T., and Joshi, S. (2007), "Riemannian Geometry for the Statistical Analysis of Diffusion Tensor Data," *Signal Processing*, 87, 250–262. [424]

Fletcher, P. T., Lu, C. Pizer, S. M., and Joshi, S. (2004), "Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape," *IEEE Transactions on Medical Imaging*, 23, 995–1005. [424,426]

Fréchet, M. (1948), "Les éléments Aléatoires de Nature Quelconque Dans un Espace Distancié," *Annales de l'Institut Henri Poincaré*, 10, 215–310. [424,425]

Goodall, C. R. (1991), "Procrustes Methods in the Statistical Analysis of Shape" (with discussion), *Journal of the Royal Statistical Society,* Series B, 53, 285–339. [424]

Hastie, T., and Stuetzle, W. (1989), "Principal Curves," *Journal of the American Statistical Association*, 84, 502–516. [425,431,433]

Huckemann, S., Hotz, T., and Munk, A. (2010), "Intrinsic Shape Analysis: Geodesic PCA for Riemannian Manifolds Modulo Isometric Lie Group Actions" (with discussion), *Statistica Sinica*, 20, 1–100. [424]

Huckemann, S., and Ziezold, H. (2006), "Principal Component Analysis for Riemannian Manifolds, With an Application to Triangular Shape Spaces," *Advances in Applied Probability*, 38, 299–319. [424]

Jupp, P. E., and Kent, J. T. (1987), "Fitting Smooth Paths to Spherical Data," *Journal of the Royal Statistical Soceity,* Series C, 36, 34–46. [425]

Jung, S., Foskey, M., and Marron, J. S. (2010), "Principal arc Analysis on Direct Product Manifolds," *Annals of Applied Statistics*, 5, 578–603. [424]

Jung, S., Dryden, I. L., and Marron, J. S. (2012), "Analysis of Principal Nested Spheres," *Biometrika*, 99, 551–568. [424,431]

Karcher, H. (1977), "Riemannian Center of Mass and Mollifier Smoothing," *Communications in Pure Applied Mathematics*, 30, 509–541. [424]

Kendall, D. G., Barden, D., Carne, T. K., and H. Le, (1999), *Shape and Shape Theory*, New York: Wiley. [424]

Kendall, W. S., and Le, H. (2011), "Limit Theorems for Empirical Fréchet Means of Independent and Non-Identically Distributed Manifold-Valued Random Variables," *Brazilian Journal of Probability and Statistics*, 25, 323–352. [424]

Kenobi, K., Dryden, I. L., and Le, H. (2010), "Shape Curves and Geodesic Modelling," *Biometrika*, 97, 567–584. [425]

Kume, A., Dryden, I. L., and Le, H. (2007), "Shape-Space Smoothing Splines for Planar Landmark Data," *Biometrika*, 94, 513–528. [425]

Magnus, J. R. (1985), "On Differentiating Eigenvalues and Eigenvectors," *Econometric Theory*, 1, 179–191. [435]

Mardia, K. V., and Jupp, P. E. (2000), *Directional Statistics*, Chichester: Wiley. [424]

Mardia, K. V., and Khatri, C. G. (1977), "Uniform Distribution on a Stiefel Manifold," *Journal of Multivariate Analysis*, 7, 468–473. [424]

Rabier, P. J., and Rheinboldt, W. C. (1995), "On the Numerical Solution of the Euler-Lagrange Equations," *SIAM Journal on Numerical Analysis*, 32, 318–329. [428,429]

Schmidt-Hieber, J., Munk, A., and Dümbgen, L. (2013), "Multiscale Methods for Shape Constraints in Deconvolution: Confidence Statements for Qualitative Features," *The Annals of Statistics*, 41, 1299–1328. [435]

Schwartzman, A. (2006), *Random Ellipsoids and False Discovery Rates: Statistics for Diffusion Tensor Imaging Data*, Ph.D. Thesis, Stanford University. [424]

Small, C. G. (1996), *The Statistical Theory of Shape*, New York: Springer. [425]

Stoyan, D., Kendall, W. S., and Mecke, J. (1995), *Stochastic Geometry and Its Applications*, Chichester: Wiley. [424]

Su, J., Dryden, I. L., Klassen, E., Le, H., and Srivastava, A. (2012), "Fitting Smoothing Splines to Time-Indexed, Noisy Points on Nonlinear Manifolds," *Image and Vision Computing*, 30, 428–442. [425]

Thorpe, J. (1979), *Elementary Topics in Differential Geometry*, New York: Springer-Verlag. [425]

U.S. Geological Survey, Global Earthquake Search. Available at: *http://earthquake.usgs.gov/earthquakes/eqarchives/epic/*, in progress. [432]

Watson, G. S. (1983), *Statistics on Spheres*, New York: Wiley. [424]