Analysis of principal nested spheres

By SUNGKYU JUNG

Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, U.S.A.
sungkyu@pitt.edu

IAN L. DRYDEN

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.
ian.dryden@mailbox.sc.edu

AND J. S. MARRON

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
marron@unc.edu

SUMMARY

A general framework for a novel non-geodesic decomposition of high-dimensional spheres or high-dimensional shape spaces for planar landmarks is discussed. The decomposition, principal nested spheres, leads to a sequence of submanifolds with decreasing intrinsic dimensions, which can be interpreted as an analogue of principal component analysis. In a number of real datasets, an apparent one-dimensional mode of variation curving through more than one geodesic component is captured in the one-dimensional component of principal nested spheres. While analysis of principal nested spheres provides an intuitive and flexible decomposition of the high-dimensional sphere, an interesting special case of the analysis results in finding principal geodesics, similar to those from previous approaches to manifold principal component analysis. An adaptation of our method to Kendall's shape space is discussed, and a computational algorithm for fitting principal nested spheres is proposed. The result provides a coordinate system to visualize the data structure and an intuitive summary of principal modes of variation, as exemplified by several datasets.

Some key words: Dimension reduction; Kendall's shape space; Manifold; Principal arc; Principal component analysis; Spherical data.

1. INTRODUCTION

This paper proposes a general framework for a novel decomposition of a high-dimensional sphere, which is the sample space of directions (Fisher, 1993; Fisher et al., 1993; Mardia & Jupp, 2000) and preshapes in Kendall's statistical theory of landmark shapes (Kendall, 1984; Dryden & Mardia, 1998). The proposed decomposition method, analysis of principal nested spheres, is a flexible extension of principal component analysis to spheres. This method provides an effective means of analysing the main modes of variation of a dataset and also gives a basis for dimension reduction. There have been a number of extensions of principal component analysis to manifold-valued data, most of which find principal geodesics (Fletcher et al., 2004; Huckemann & Ziezold, 2006; Huckemann et al., 2010; Kenobi et al., 2010). A geodesic
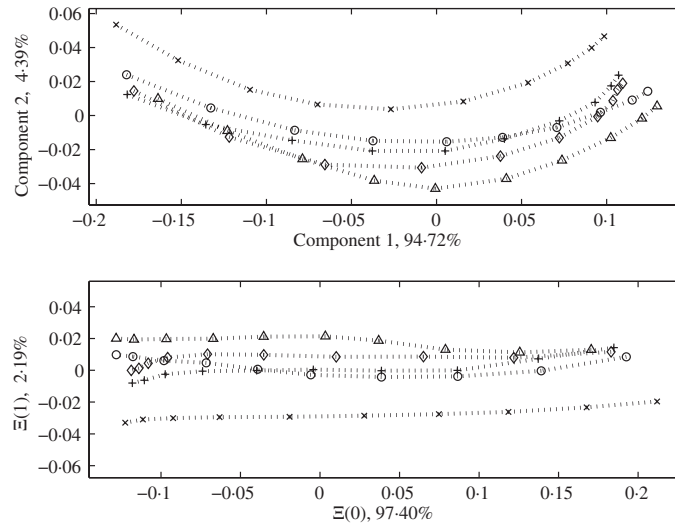
Fig. 1. Plots of scores for some human movement data by two major geodesic components (top), where the different symbols represent different movements and samples for each movement are interpolated, and by the first two lowest dimensional nested spheres (bottom). The number% is the percent variance explained, defined later in §2·4. The one-dimensional principal nested sphere captures a larger proportion of variation than the first geodesic component.

on a manifold is a shortest path between two points and can be understood as the analogue of a straight line in Euclidean space. In particular, a geodesic on a sphere is a great circle path.

There has been a concern that when non-geodesic variation is major and apparent, geodesic-based methods do not give a fully effective decomposition of the space. As an example, a dataset of shapes representing human movements, discussed later in §5·1 and introduced in Kume et al. (2007), is plotted in the top panel of Fig. 1 using the first two principal component directions. In this dataset the major one-dimensional mode of variation is represented as a curve lying in two components, as in the top panel of Fig. 1, and thus at least two dimensions are needed to explain the major variation. Analysis of principal nested spheres decomposes the data space so that the major one-dimensional variation is linearly represented, as shown in the bottom panel.

For a unit d-sphere S^d, which is the set of unit vectors in $\mathbb{R}^{d+1}$, the analysis gives a decomposition of S^d that captures the non-geodesic variation in a lower dimensional subsphere. The decomposition sequentially provides the best k-dimensional approximation $\mathfrak{A}_k$ of the data for each $k = 0, 1, \ldots, d-1$. The sphere $\mathfrak{A}_k$ is called the k-dimensional principal nested sphere and is a submanifold of the higher dimensional principal nested spheres. The sequence of principal nested spheres is

$$\mathfrak{A}_0 \subset \mathfrak{A}_1 \subset \cdots \subset \mathfrak{A}_{d-1} \subset S^d. \tag{1}$$

Since the preshape space of two-dimensional landmark-based shapes is also a sphere, the method can be applied to shape data lifted to and aligned on the preshape space; see §4. The analysis provides approximations to the directional or shape data for every dimension, captures the non-geodesic variation, and provides visualization of the major variability in terms of shape changes.

The procedure of fitting principal nested spheres involves iterative reduction of the data dimension. We first fit a $(d-1)$-dimensional subsphere $\mathfrak{A}_{d-1}$ of S^d that best approximates the data. This subsphere is not necessarily a great sphere, i.e., a sphere with radius 1 analogous to the great circle for S^2, which makes the resulting decomposition non-geodesic. Nevertheless, $\mathfrak{A}_{d-1}$ can

be treated as if it were the unit $(d-1)$-sphere. Each data point has an associated residual, which is the signed geodesic distance to its projection on $\mathfrak{A}_{d-1}$. Then for the data projected onto the subsphere, we continue to search for the best fitting $(d-2)$-dimensional subsphere. These steps are iterated to find lower dimensional nested spheres. For visualization and further analysis, we obtain an Euclidean-type representation of the data, essentially consisting of the residuals of each level. The first two coordinates of this representation, related to $\mathfrak{A}_1$ and $\mathfrak{A}_2$, are plotted in Fig. 1 for the human movement data.

In Fig. 1 the variation of the dataset is represented linearly in a plot of scores given by principal nested spheres in the bottom panel compared to the curvy form of variation in geodesic components in the top panel. The proportion of variance contained in $\mathfrak{A}_1$ is almost the proportion of the sum of the first two geodesic component variances. That is, the variation explained by two geodesic components is almost attained in only the one-dimensional principal nested sphere. Moreover, the graph in the top panel is indeed obtained by a special case of our analysis, which is similar to geodesic-based principal component analysis (Fletcher et al., 2004; Huckemann et al., 2010).

Nonlinear variation like that in Fig. 1 is observed in many other datasets, where the proposed method is most useful. Such situations can be found in landmark shape analysis, examples of which are in § 5, and in computer-aided image analysis. For example, in investigations of shape variations of human organs including the lung, prostate, and hippocampus, applications of our method lead to succinct representations of the data compared to those from geodesic or linear principal component analysis (Jung et al., 2010, 2011; Pizer et al., 2012). The proposed method can also be seen as an extension of Jung et al. (2011) from S^2 to S^d, $d > 2$.

2. PRINCIPAL NESTED SPHERES

2·1. *Geometry of nested spheres*

We begin by describing the essential geometric background for nested spheres. More detailed discussion can be found in the Supplementary Material.

For a unit sphere S^d, a geodesic joining any two points is a great circle joining the two points. A natural distance function on S^d is the geodesic distance function $\rho_d(\cdot, \cdot)$ defined as the length of the shortest great circle segment joining $x, y \in S^d$, $\rho_d(x, y) = \cos^{-1}(x^{\mathrm{T}}y)$. The shortest great circle path is unique unless $x^{\mathrm{T}}y = -1$, i.e., the two points are antipodal.

DEFINITION 1. *A subsphere A_{d-1} of S^d is defined by an axis $v \in S^d$ and a distance $r \in (0, \pi/2]$, as follows*:

$$A_{d-1}(v, r) = \{x \in S^d : \rho_d(v, x) = r\}.$$

The subsphere A_{d-1} can be viewed as an intersection of $S^d \subset \mathbb{R}^{d+1}$ and a d-dimensional hyperplane, $\{x \in \mathbb{R}^{d+1} : v^{\mathrm{T}}x - \cos(r) = 0\}$. In other words, A_{d-1} is identified with a slicing of S^d with the hyperplane, an example of which is illustrated as a shaded plane in Fig. 2. A subsphere A_{d-1} is indeed a $(d-1)$-dimensional nested sphere $\mathfrak{A}_{d-1}$ of S^d.

The subsphere A_{d-1} is identified with S^{d-1}, so we can treat A_{d-1} as a unit sphere S^{d-1}. This is done by a function $f_1 : A_{d-1} \to S^{d-1}$ and its inverse f_1^{-1}, defined in (2) below and depicted in Fig. 2. A subsphere A_{d-2} of S^{d-1} can be obtained by Definition 1 with dimension d reduced by 1. For a general subsphere A_{d-k} of S^{d-k+1} for $k = 1, \ldots, d-1$, we also use the transformation $f_k : A_{d-k} \to S^{d-k}$ and its inverse f_k^{-1}. Let $m = d - k + 1$, so that the subsphere $A_{d-k} \subset S^m \subset$
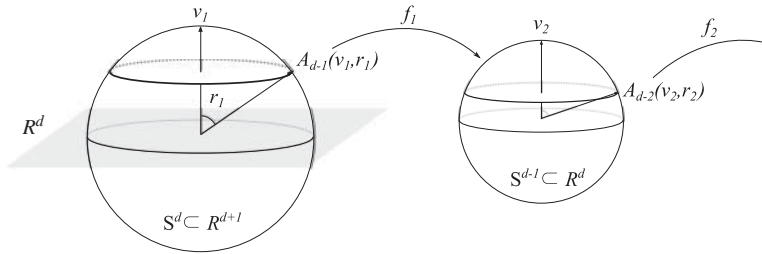
Fig. 2. The subsphere $A_{d-1}(v_1, r_1)$ in S^d and its relation to S^{d-1}, through the transformation f_1. Recursively, $A_{d-2}(v_2, r_2)$ is found in S^{d-1}, and is identified with S^{d-2} by the transformation f_2.

$\mathbb{R}^{m+1}$. The transformations are defined by $v_k \in S^m$ and $r_k \in (0, \pi/2]$ as

$$f_k(x) = \frac{1}{\sin(r_k)} R^-(v_k)x, \quad f_k^{-1}(x^\dagger) = R^{\mathrm{T}}(v_k) \begin{bmatrix} \sin(r_k)x^\dagger \\ \cos(r_k) \end{bmatrix}, \quad x \in A_{d-k}, x^\dagger \in S^{d-k}, \quad (2)$$

where $R(v_k)$ is a $(m+1) \times (m+1)$ rotation matrix that moves v_k to the north pole, and $R^-(v_k)$ is the $m \times (m+1)$ matrix consisting of the first m rows of $R(v_k)$. See (A2) for a precise definition of $R(v_k)$.

The subspheres A_{d-k} are defined in different spaces S^{d-k+1} for each k. A nested sphere is defined by the subsphere located in the original space S^d.

DEFINITION 2. *A $(d-k)$-dimensional nested sphere $\mathfrak{A}_{d-k}$ of S^d is defined as*

$$\mathfrak{A}_{d-k} = \begin{cases} f_1^{-1} \circ \cdots \circ f_{k-1}^{-1}(A_{d-k}) & (k = 2, \dots, d-1), \\ A_{d-1} & (k = 1). \end{cases}$$

The nested spheres $\mathfrak{A}_{d-k}$ ($k = 1, \dots, d-1$) satisfy the relation (1). A $(d-k)$-dimensional nested sphere $\mathfrak{A}_{d-k}$ is indeed identified with a slicing of S^d by a $(d-k+1)$-dimensional hyperplane.

2·2. *The best fitting subsphere*

Let $x_1, \dots, x_n$ be a sample in S^d, $d \geqslant 2$. We first define the residual ξ of x from a subsphere $A_{d-1}(v_1, r_1)$ of S^d as the signed length of the minimal geodesic that joins x to A_{d-1}. Then $\xi = \rho_d(x, v_1) - r_1$. The sign of ξ is negative if x is in the interior of the geodesic ball corresponding to A_{d-1}, and is positive if x is in the exterior.

The best fitting subsphere $\hat{A}_{d-1} = A_{d-1}(\hat{v}_1, \hat{r}_1)$ is found by minimizing the sum of squares of residuals of the data points to $\hat{A}_{d-1}$. In other words, $\hat{v}_1$ and $\hat{r}_1$ minimize

$$\sum_{i=1}^n \xi_i(v_1, r_1)^2 = \sum_{i=1}^n \{\rho_d(x_i, v_1) - r_1\}^2, \quad (3)$$

among all $v_1 \in S^d$, $r_1 \in (0, \pi/2]$.

The method can be extended using other objective functions, e.g., the sum of absolute deviations, for more robust fitting.

Each x_i can be projected on $\hat{A}_{d-1}$ along the minimal geodesic that joins x to $\hat{A}_{d-1}$. The projection of x onto A_{d-1} is defined as

$$P\{x; A_{d-1}(v, r)\} = \frac{\sin(r)x + \sin\{\rho_d(x, v) - r\}v}{\sin\{\rho_d(x, v)\}}. \tag{4}$$

Denote $x^P = P(x; \hat{A}_{d-1}) \in \hat{A}_{d-1}$ for the projected x. We use the isomorphic function $\hat{f}_1 \equiv f(\hat{v}_1, \hat{r}_1)$, as defined in (2), to transform $\hat{A}_{d-1}$ to S^{d-1} so that $\hat{f}_1(x_i^P) \in S^{d-1}$.

2·3. *The sequence of principal nested spheres*

The procedure to find the sample principal nested spheres consists of iteratively finding the best fitting subsphere and mapping to the original space.

The $(d-1)$-dimensional sample principal nested sphere $\hat{\mathfrak{A}}_{d-1}$ is the best fitting subsphere $A_{d-1}(\hat{v}_1, \hat{r}_1)$. The second layer $\hat{\mathfrak{A}}_{d-2}$ is obtained from the subsphere that best fits $\hat{f}_1(x_i^P) \in S^{d-1}$. The best fitting subsphere $A_{d-2}(\hat{v}_2, \hat{r}_2) \subset S^{d-1}$ is then mapped to S^d by the relevant transformation $\hat{f}_1^{-1}$ and becomes $\hat{\mathfrak{A}}_{d-2}$.

In general, we recursively find the sequence of best fitting subspheres from the projected and transformed data points $\hat{f}_k\{P(x; \hat{A}_{d-k})\} \in S^{d-k}$ $(x \in S^{d-k+1})$. At the kth level, where we fit a subsphere from S^{d-k+1}, we denote the best fitting subsphere as $\hat{A}_{d-k} \equiv A_{d-k}(\hat{v}_k, \hat{r}_k)$ and keep residuals defined by $\xi_{i,d-k}$ $(i = 1, \ldots, n)$ for later use as analogues of principal component scores.

The lowest level best fitting subsphere $\hat{A}_1$ is then a small circle isomorphic to S^1. No further sphere or circle can be used to reduce the dimensionality. Instead, we find the Fréchet mean (Fréchet, 1944, 1948; Karcher, 1977; Bhattacharya & Patrangenaru, 2003) $\hat{A}_0$ of $x_1^\dagger, \ldots, x_n^\dagger$, which are the projected and transformed data points in S^1. The Fréchet mean can be thought of as a best zero-dimensional representation of the data in our framework, and is defined as the minimizer of the sum of squared distances to the $x_i^\dagger$s, i.e.,

$$\hat{A}_0 = \operatorname*{argmin}_{x \in S^1} \sum_{i=1}^{n} \rho_1(x, x_i^\dagger)^2.$$

The Fréchet mean is unique when the support of $x_i^\dagger$ is a proper subset of a half-circle in S^1 (Karcher, 1977), which is often satisfied in practice. If there are multiple Fréchet means, then the data must be carefully inspected. A typical case for having multiple means is that the data are uniformly distributed on the circle. If this is the case, then $\hat{A}_0$ can be chosen to be any solution of the above criterion and we may not lay much emphasis on $\hat{A}_0$.

The sequence of best fitting subspheres including $\hat{A}_0$ can be located in the original space S^d, as follows.

DEFINITION 3. *The sequence of sample principal nested spheres in S^d is $\{\hat{\mathfrak{A}}_0, \ldots, \hat{\mathfrak{A}}_{d-1}\}$, where*

$$\hat{\mathfrak{A}}_{d-k} = \begin{cases} \hat{f}_1^{-1} \circ \cdots \circ \hat{f}_{k-1}^{-1}(\hat{A}_{d-k}) & (k = 2, \ldots, d), \\ \hat{A}_{d-1} & (k = 1). \end{cases}$$

We call $\hat{\mathfrak{A}}_0$ the principal nested spheres mean.

2·4. *Euclidean-type representation*

We wish to represent the data in a Euclidean space for visualization and further analysis.

Recall that in the kth level of the procedure, we have collected the signed residuals $\xi_{i,d-k}$ ($i = 1, \ldots, n$). These residuals were measured by the metric ρ_{d-k} of the unit sphere S^{d-k}, which was used to identify $\hat{\mathfrak{A}}_{d-k} \subset S^d$. A natural spherical metric for $\hat{\mathfrak{A}}_{d-k}$ is defined for $x, y \in \hat{\mathfrak{A}}_{d-k}$ as

$$\rho_{d-k}^*(x, y) = \prod_{i=1}^{k} \sin(\hat{r}_i) \rho_{d-k}\{\hat{f}^k(x), \hat{f}^k(y)\} \qquad (k = 1, \ldots, d-1),$$

where $\hat{f}^k(x) = \hat{f}_k \circ \cdots \circ \hat{f}_1(x)$. The distance measured by ρ_{d-k}^* is different from the great circle distance measured by ρ_d whenever any $\hat{r}_i$ is smaller than $\pi/2$, that is, whenever a small sphere was used in the sequence. However, it can be shown that $\rho_{d-k}^*(x, y)$ is the length of the minimal arc in $\hat{\mathfrak{A}}_{d-k}$ that joins x and y.

The residuals at the kth level are now rescaled as if they were measured by ρ_{d-k}^*, so that their magnitudes are commensurate. This is achieved by multiplying $\prod_{i=1}^{k-1} \sin(\hat{r}_i)$ to the residuals. We put the scaled residuals in a row vector

$$\Xi(d-k)_{1 \times n} = \prod_{i=1}^{k-1} \sin(\hat{r}_i)(\xi_{1,d-k}, \ldots, \xi_{n,d-k}).$$

We further define $\xi_{i,0}$ as the ith sample's signed deviation from $\hat{A}_0$ measured by ρ_1^*. Similarly to before, rescale the deviations and let

$$\Xi(0)_{1 \times n} = \prod_{i=1}^{d-1} \sin(\hat{r}_i)(\xi_{1,0}, \ldots, \xi_{n,0}).$$

These commensurate residuals are combined into a $d \times n$ data matrix

$$\hat{X}_{\text{PNS}} = \begin{bmatrix} \Xi(0) \\ \Xi(1) \\ \vdots \\ \Xi(d-1) \end{bmatrix},$$

where each column is the corresponding sample's coordinates in terms of the sample principal nested spheres, and lies in a subset of $E = [-\pi, \pi] \times [-\pi/2, \pi/2]^{d-1}$. Each entry in $\Xi(k)$ has a role like the $(k+1)$th principal component scores.

The data matrix $\hat{X}_{\text{PNS}}$ can be used to visualize the structure of the data. For example, the graph in Fig. 1 is a scatter-plot of $\Xi(0)$ and $\Xi(1)$. The variance of the jth component is $s_j^2 = (n-1)^{-1} \sum_{i=1}^{n} \tilde{\xi}_{i,j-1}^2$, where $\tilde{\xi}_{i,j-1}$ ($i = 1, \ldots, n$) are the elements of $\Xi(j-1)$. The percent of variance exhibited in Fig. 1, and in the rest of the paper, is the proportion of the jth component variance of the sum of all variances, i.e., $s_j^2 / \sum_{\ell=1}^{d} s_\ell^2$.

Moreover, conventional multivariate statistics based on Euclidean geometry can be applied to $\hat{X}_{\text{PNS}}$ for further analysis.

Remark 1. A point in S^d can be mapped to E by utilizing the projection (4) and rescaling, leading to a mapping $h : S^d \to E$. The inverse operation $\tilde{h} : E^* \to S^d$ can also be defined, for a subset $E^* \subset E$. These transformations are built in an obvious way, details of which are contained in the Supplementary Material.

2·5. *Principal arcs*

In analogy with the principal component directions in Euclidean space, or the manifold extension principal geodesics, the principal arcs that represent the directions of major variation are defined. These arcs are small circles in S^d, which frequently are not equivalent to any geodesic.

Given a sequence of principal nested spheres $\hat{\mathfrak{A}}_0, \ldots, \hat{\mathfrak{A}}_{d-1}$, the first principal arc coincides with $\hat{\mathfrak{A}}_1$. This arc may be parameterized by the signed distance from $\hat{\mathfrak{A}}_0$. In the space E of $\hat{X}_{\mathrm{PNS}}$, the first principal arc γ_1 coincides with the direction $e_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$, and is parameterized by $\gamma_1(t) = \tilde{h}(te_1)$, for $t \in \mathbb{R}$. See Remark 1 for the definition of $\tilde{h}$.

The second principal arc γ_2 lies in $\hat{\mathfrak{A}}_2$ and is orthogonal to γ_1 at all points common to both. The first and second principal arcs cross at $\hat{\mathfrak{A}}_0$ and also at the farthest point from $\hat{\mathfrak{A}}_0$ on $\hat{\mathfrak{A}}_1$. The second principal arc is in general a small circle in S^d but is identified with a great circle in S^2, the isomorphic space of $\hat{\mathfrak{A}}_2$. The second principal arc in S^2 must pass through the axis v_{d-1} in order to be orthogonal to the first. This arc may be parameterized by the signed distance from $\hat{\mathfrak{A}}_0$, and coincides with the direction e_2 in the space of $\hat{X}_{\mathrm{PNS}}$.

The higher order principal arcs are defined in the same manner. The kth principal arc γ_k can be defined and identified with the direction e_k in the space of $\hat{X}_{\mathrm{PNS}}$ by $\gamma_k(t) = \tilde{h}(te_k)$. The kth arc is then orthogonal to the principal arcs of order $1, \ldots, k-1$, and passes through $\hat{\mathfrak{A}}_0$.

The term principal arc was used in Jung et al. (2011), which concerns the space of direct products of two-spheres and Euclidean spaces. The principal arc γ_k used here coincides with the term in Jung et al. (2011), and both represent the path of major variation.

2·6. *Principal nested spheres restricted to great spheres*

An important special case of the analysis is obtained by setting $r = \pi/2$ for each subsphere fitting. This restriction leads to the nested spheres being great spheres, and the principal arcs become geodesics. We call this special case principal nested great spheres.

We conjecture that the principal geodesics, found by principal nested great spheres, are more similar to the geodesic principal components of Huckemann et al. (2010) than the usual tangent space projection methods. This is mainly because no predetermined mean, either geodesic or Procrustes, is used in our approach or in that of Huckemann et al. (2010). When $r = \pi/2$, the mean of the principal nested spheres is similar, but not identical, to the notion of mean of Huckemann et al. (2010). Although we have not yet found a significant difference between the principal nested great spheres and the geodesic-based methods in the literature, we point out that our approach as a generalization of principal components is conceptually different from the geodesic-based methods. See § 7 for discussion on this topic.

3. Computational algorithm

The computation of sample principal nested spheres involves iterative applications of minimization, projection and transformation. We have given explicit formulas for the projection (4) and the transformation (2). The least squares problem (3) is a constrained nonlinear minimization problem. It can be solved by the two-step algorithm described in Jung et al. (2011) with some modifications. The outer loop finds a point of tangency to approximate S^d by a tangent space; the inner loop solves an optimization problem in the linear space.

We make use of the exponential map and its inverse for mappings between the manifold and tangent spaces (Helgason, 2001; Buss & Fillmore, 2001). A tangent space at $p \in S^m$, $T_p S^m$, is an affine m-dimensional vector space and can be identified with $\mathbb{R}^m$. Without loss of generality, set the point of tangency as $p = e_{m+1} = (0, \ldots, 0, 1)$, because one can use the rotation operator

$R(p)$ to transform p to e_{m+1} while preserving all the data structure. The exponential map $\mathrm{Exp}_p :
T_p S^m \longrightarrow S^m$ is defined by

$$\mathrm{Exp}_p(z) = \left(\frac{\sin(\|z\|)}{\|z\|} z^{\mathrm{T}}, \cos(\|z\|) \right)^{\mathrm{T}} \in S^m \qquad (z \in \mathbb{R}^m \cong T_p S^m).$$

The inverse exponential map or the log map $\mathrm{Log}_p : S^m \longrightarrow T_p S^m$ is defined by

$$\mathrm{Log}_p(x) = \frac{\theta}{\sin(\theta)} (x_1, \ldots, x_m)^{\mathrm{T}} \in \mathbb{R}^m, \quad x = (x_1, \ldots, x_{m+1})^{\mathrm{T}} \in S^m,$$

where $\cos(\theta) = x_{m+1}$. These mappings preserve the distances to the point of tangency. By using
the exponential mapping and its inverse, a hypersphere with radius r in the tangent space corre-
sponds to a subsphere in S^m with distance r. In particular, $A_{m-1}(v, r)$ is equivalent to the image
of $\{x \in \mathbb{R}^m : \|x\| = r\}$ by Exp_v.

The algorithm finds a suitable point of tangency v, which is also the axis of the fitted sub-
sphere. Given a candidate v_0, the data are mapped to the tangent space $T_{v_0} S^m$ by the log map.
With $x_i^{\dagger} = \mathrm{Log}_{v_0}(x_i)$, the inner loop finds

$$\min_{v^{\dagger}, r} \sum_{i=1}^{n} (\|x_i^{\dagger} - v^{\dagger}\| - r)^2,$$

a nonlinear least-squares problem that can be solved numerically by, e.g., the Levenberg–
Marquardt algorithm (Scales, 1985, Ch. 4). The solution $v^{\dagger}$ is then mapped to S^m by the expo-
nential map, and becomes the updated value of v denoted by v_1. This procedure is repeated until
v converges.

A main advantage of this approach is the reduced difficulty of the optimization task. The
inner loop solves an unconstrained problem in a vector space, which is much simpler than the
original constrained problem on manifolds. We have found that in many real data applications,
including those not contained in the paper, the algorithm converges to at least a local minimum.
The computation times for the examples in this paper on a computer with a $2 \cdot 13$ GHz processor
are less than 2 seconds in MATLAB. In our experience with a number of datasets, the longest
time required was less than 5 seconds. The computation of the principal nested great spheres is
about two times faster than that of the principal nested spheres.

In modern applied problems the sample size is often less than the dimension of the mani-
fold, i.e., $x_1, \ldots, x_n \in S^d$ with $n \leqslant d$, which is frequently referred to as the high-dimension low-
sample-size situation (Hall et al., 2005; Dryden, 2005; Ahn et al., 2007). In Euclidean space, the
dimension of the data can be reduced to n without losing any information. Likewise, the intrin-
sic dimensionality of the data on the hypersphere can be reduced to $n - 1$, where the additional
reduction of 1 occurs because there is no origin in S^d. For the simplest example, let $n = 2$. Then
there is a geodesic joining the two points, which is the submanifold containing all information.
A generalization can be made for any $n > 2$.

Remark 2. For $n < d$, there exists an $(n - 1)$-dimensional nested sphere $\mathfrak{A}_{n-1}$ of S^d satisfying
$x_i \in \mathfrak{A}_{n-1}$ for all $i = 1, \ldots, n$. Moreover, there exist $\mathfrak{A}_{d-1} \supset \cdots \supset \mathfrak{A}_{n-1}$, all of which are great
spheres with radius 1. The singular value decomposition of the data matrix $[x_1, \ldots, x_n]$ gives the
appropriate $\mathfrak{A}_{n-1}$. Let $\mathcal{L}_1^n$ be the vector space of dimension n that all data points span. Then, the
intersection of $\mathcal{L}_1^n$ and S^d is the $(n - 1)$-dimensional great sphere $\mathfrak{A}_{n-1}$.

For faster computation, when $n < d$, we reduce the dimensionality to $\mathfrak{A}_{n-1}$ by the singular value decomposition, and use the proposed algorithm to fit $\mathfrak{A}_{n-2}$, and so on.

4. APPLICATION TO PLANAR SHAPE SPACE

4·1. *Planar shape space*

The shape of an object is what is left after removing location, scale, and rotation. The classical approach in shape analysis is to work with landmarks on the objects; see, e.g., Dryden & Mardia (1998). Each shape determined by a set of landmarks can be represented by a point in Kendall's (1984) shape space. A useful approach to understanding the non-Euclidean shape space is through the preshape space, which is a sphere. We begin by summarizing Kendall's framework for shape data, followed by a discussion of necessary considerations to apply our method to shape space through the preshape space.

Consider a set of $k > 2$ landmarks in $\mathbb{R}^2$ and the corresponding configuration matrix X, which is a $k \times 2$ matrix of Cartesian coordinates of landmarks. The preshape of the configuration X is invariant under translation and scale, which is given by $Z = HX/\|HX\|_F$, where H is the $(k-1) \times k$ Helmert submatrix (Dryden & Mardia, 1998, p. 34) and $\|\cdot\|_F$ is the Frobenius norm. Provided that $\|HX\|_F > 0$, $Z \in S^{2(k-1)-1}$. The unit sphere $S^{2(k-1)-1}$ in $\mathbb{R}^{2(k-1)}$ is the space of all possible preshapes, and is called the preshape space.

The shape of a configuration matrix X can be represented by the equivalence set under rotation, $[Z] = \{Z\Gamma : \Gamma \in \mathrm{SO}(2)\}$, where $\mathrm{SO}(2)$ is the set of all 2×2 rotation matrices. The space of all possible shapes is then a non-Euclidean space called the shape space and denoted by Σ_2^k.

A useful notation for planar shapes, allowing much simpler mathematical development, is to represent a preshape $Z = (z_{ij})$ as a complex vector $z = (z_{11} + iz_{12}, \ldots, z_{k-1,1} + iz_{k-1,2})^{\mathrm{T}}$. Let z^* denote the complex conjugate transpose of z. Then the following facts may be found in Small (1996, Ch. 3).

The horizontal subspace of the tangent space of $S^{2(k-1)-1}$ at w, with respect to the quotient map from $S^{2(k-1)-1}$ to Σ_2^k, is expressed as $H_w = \{v \in T_w S^{2(k-1)-1} : v^*w = 0\}$. The horizontal subspace is isometric to the tangent space of Σ_2^k at $[w]$ (Kendall et al., 1999, p. 137). A horizontal geodesic at w in $S^{2(k-1)-1}$ is a great circle path starting at w with initial direction in H_w. Suppose that v, w are complex preshapes satisfying $v^*w \geqslant 0$, by which we mean $\Re(v^*w) \geqslant 0$ and $\Im(v^*w) = 0$. Then the geodesic joining v and w is a horizontal geodesic and the spherical distance between v and w is the same as the Riemannian distance between the corresponding shapes $[v]$ and $[w]$ in Σ_2^k, where the Riemannian distance is defined as

$$\rho_{\Sigma_2^k}([v], [w]) = \min_{\phi \in (0, 2\pi]} \rho_{2(k-1)-1}(e^{i\phi}v, w) = \cos^{-1}(|v^*w|).$$

For v, w that do not satisfy $v^*w \geqslant 0$, one can optimally rotate v onto w by multiplying $e^{i\phi}$, $\phi = \arg(v^*w)$, and treat $v' = e^{i\phi}v$ as if it was the v.

Since the discussions before this section are based on real vectors, we also write the preshape Z as a vectorized version $z = \mathrm{vec}(Z^{\mathrm{T}})$, where $\mathrm{vec}(A)$ is obtained by stacking the columns of the matrix A on top of one another. Then the assumption $v^*w \geqslant 0$ translates into $v^{\mathrm{T}}w \geqslant 0$ and $v^{\mathrm{T}}Mw = 0$, where M is the $2(k-1) \times 2(k-1)$ skew-symmetric matrix consisting of $k-1$ diagonal blocks

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The geodesic that joins v to w is parametrized by $Q(v \to w, \theta)v$ $(\theta \in \mathbb{R})$ defined in (A1).

For the preshapes $w, z_1, \ldots, z_n$, as long as the shapes of those are of interest, we assume without loss of generality that $w^{\mathrm{T}} z_i \geqslant 0$ and $w^{\mathrm{T}} M z_i = 0$. These z_is are the Procrustes fits to a common base w.

4·2. *Principal nested spheres for planar shapes*

The intrinsic dimension of the shape space is $2k - 4$, since the degrees of freedom are reduced from $2k$ of the set of landmarks by 2 for translation, 1 for scale, and 1 for rotation. This is less than the dimension of the preshape space $d = 2k - 3$. It is thus desired that $\mathfrak{A}_{d-1}$ of S^d leaves no residuals. This is achieved by the theory and practical modifications in this section. In short, the Procrustes fit of configurations or preshapes to a common alignment base w, e.g., the Procrustes mean, results in the desired decomposition of the shape space.

THEOREM 1. *Suppose the preshapes* $w, z_1, \ldots, z_n \in S^d$ *satisfy* $w^{\mathrm{T}} z_i \geqslant 0$ *and* $w^{\mathrm{T}} M z_i = 0$ *for* $i = 1, \ldots, n$. *Let* $w^* = Mw$. *Then* $w, z_i \in A_{d-1}(w^*, \pi/2)$. *Moreover, define* $h A_{d-1} = \{z \in A_{d-1} : z^{\mathrm{T}} w \geqslant 0\}$ *as a sub-hemisphere. Then* $w, z_i \in h A_{d-1}(w^*, \pi/2)$.

Theorem 1 is a direct consequence of definitions and thus the proof is omitted. The $h A_{d-1}(w^*, \pi/2) \subset \mathfrak{A}_{d-1}$ is a great sub-hemisphere of dimension $d - 1$. Since $z_i \in h A_{d-1}(w^*, \pi/2)$ for all $i = 1, \ldots, n$, the dimension of preshape space S^d can be reduced by 1 to $\mathfrak{A}_{d-1} \supset h A_{d-1}(w^*, \pi/2)$ without loss of any shape information. The tangent space of $h A_{d-1}(w^*, \pi/2)$ at w is in fact identical to the horizontal subspace of the tangent space of S^d at w, since the vertical geodesic at w runs through w^*. The metrics ρ_d of preshape space and $\rho_{\Sigma_2^k}$ of the shape space are closely related. In particular,

$$\rho_{d-1}^*(w, z) = \rho_d(w, z) = \rho_{\Sigma_2^k}([w], [z]), \quad (z \in h A_{d-1}).$$

Moreover, $\rho_d(z_1, z_2)$ for $z_1, z_2 \in h A_{d-1}$ is close to $\rho_{\Sigma_2^k}([z_1], [z_2])$ when z_1, z_2 are close to w, which is justified by the following lemma, now using complex notation for simplicity of the statement and proof.

LEMMA 1. *For preshapes* w, z_1, z_2 *in complex notation satisfying* $w^* z_i \geqslant 0$,

$$\cos^2\{\rho_d(z_1, z_2)\} = \cos^2\{\rho_{\Sigma_2^k}([z_1], [z_2])\} + c^2, \tag{5}$$

where $c = \Im\{(z_1 - w)^*(z_2 - w)\}$. *Moreover, the additional term* c *satisfies*

  (i) $0 \leqslant |c| \leqslant 4 \sin\{\rho_d(z_1, w)/2\} \sin\{\rho_d(z_2, w)/2\}$,
 (ii) $0 \leqslant |c| \leqslant |\sin \phi|$, *where* $\phi = \arg(z_1^* z_2)$ *is the amount of the rotation to optimally position* z_1 *onto* z_2.

Proof. By the definitions $\rho_d(z_1, z_2) = \arccos\{\Re(z_1^* z_2)\}$, $\rho_{\Sigma_2^k}([z_1], [z_2]) = \arccos(|z_1^* z_2|) = \arccos[\{\Re(z_1^* z_2)^2 + c^2\}^{1/2}]$, we have (5) with $c = \Im(z_1^* z_2) = \Im\{(z_1 - w)^*(z_2 - w) + w^* z_2 + z_1^* w - w^* w\} = \Im\{(z_1 - w)^*(z_2 - w)\}$. Denote the inner product and norm of complex vectors a, b as $\langle a, b \rangle = a^* b$ and $\|a\| = \langle a, a \rangle^{1/2}$. Then we have

$$c = \|z_1 - w\| \, \|z_2 - w\| \, \Im\left\{ \left(\frac{z_1 - w}{\|z_1 - w\|} \right)^* \frac{(z_2 - w)}{\|(z_2 - w)\|} \right\} \leqslant \|z_1 - w\| \, \|z_2 - w\|.$$

The first inequality is obtained by noting $\rho_d(z, w) = 2 \arcsin(\|z - w\|/2)$. For the second, $z_1^* z_2 = |z_1^* z_2| e^{i\phi}$, $\phi = \arg(z_1^* z_2)$, gives $c = \Im(z_1^* z_2) = |z_1^* z_2| \sin \phi$, which in turn gives $|c| \leqslant |\sin \phi|$. $\qquad\square$

The $h A_{d-1}(w^*, \pi/2)$ is therefore closely related to Σ_2^k, but is not identical to it. As previously noted our method applies to shape data that are aligned and lifted to the preshape sphere. The choice of the alignment base w is an important issue, which we discuss next.

In general, we wish to set the alignment base w as a centre of the data. We recommend using the preshape of the Procrustes mean of the data. Other reasonable candidates for w are the geodesic mean and the principal nested spheres mean $\mathfrak{A}_0$ as defined in Definition 3. We have tested these choices on a number of real and simulated datasets. Setting w as the principal nested spheres mean or the geodesic mean usually takes more time in computation than using the full Procrustes mean, but the resulting decompositions are virtually the same in most cases.

In the following, we describe all candidates for w in more detail, giving the advantages and disadvantages of each option.

We first considered use of the principal nested spheres mean $\mathfrak{A}_0$ as the alignment base. The $\mathfrak{A}_0$ is identified with the origin of the coordinate system for the Euclidean representation of data $\hat{X}_{\mathrm{PNS}}$. Since $\mathfrak{A}_0$ is estimated from the data, we begin with the preshape of the full Procrustes mean as an initial guess for $\mathfrak{A}_0$ and recursively update $\hat{\mathfrak{A}}_0$ on which preshapes are aligned. The algorithm consists of the following steps.

Step 1. Initialize w as the preshape of the Procrustes mean of z_i.

Step 2. Align $z_1, \ldots, z_n$ to w and compute $\hat{\mathfrak{A}}_0$ from aligned z_i.

Step 3. If $\rho_d(w, \hat{\mathfrak{A}}_0) < \epsilon$, then set $w = \hat{\mathfrak{A}}_0$ and stop. Otherwise update $w = \hat{\mathfrak{A}}_0$ and go to Step 2.

In practice, there is no guarantee that this algorithm will converge.

Other candidates for w are the full Procrustes mean preshape and the geodesic mean of the preshapes. These are relevant to the Fréchet mean (Fréchet, 1948; Karcher, 1977; Bhattacharya & Patrangenaru, 2003), where the geodesic mean is the Fréchet mean with the Riemannian distance and the Procrustes mean is using the full Procrustes distance, which is extrinsic to S^d. Recently, it has been observed that the curvature of the manifold S^d sometimes makes the Fréchet mean inadequate, since it can lie outside the main support of the data. See Huckemann et al. (2010). When the Fréchet mean is indeed a useful representation of the data, $\hat{\mathfrak{A}}_0$ is usually found at a point close to the Fréchet mean. Even if the Fréchet mean is far from the data, $\hat{\mathfrak{A}}_0$ is nevertheless located at an appropriate centre of the data.

In the next section, the Procrustes mean is used as the alignment base w.

5. Real data analysis

5·1. *Human movement*

A human movement dataset introduced in Kume et al. (2007) contains 50 samples of $k = 4$ landmarks on the lower back, shoulder, wrist, and index finger. The dataset consists of shape configurations of five different reaching movements projected in the plane of a table, each observed at ten different time-points. The raw data are plotted in Fig. 3(a), as 50 overlaid quadrilaterals, each of which is a shape configuration. Vertices of the quadrilaterals are the locations of the landmarks in two dimensions. These 50 samples have been Procrustes fitted to each other, i.e., translated, scaled, and rotated to each other so that they are as close as possible.

We applied the analysis of principal nested spheres and great spheres, with the alignment base w being the preshape of the Procrustes mean, and compared the results with the geodesic-based
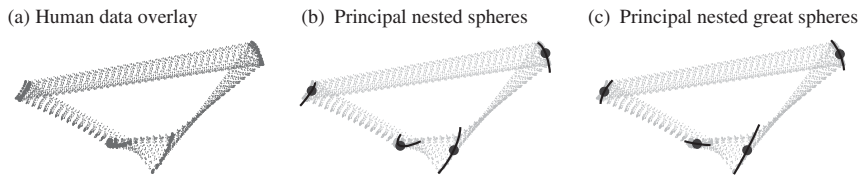
(a) Human data overlay   (b) Principal nested spheres   (c) Principal nested great spheres

Fig. 3. Human movement data. (a) Procrustes fitted human data. (b) The first principal mode of variation by principal nested spheres, showing ± 2 standard deviation from $\hat{A}_0$. (c) The first principal mode of variation by principal nested great spheres. The curving form of variation of the lower left vertex is more precisely captured in (b) than in (c).

Table 1. *Proportions of variance (%), captured in the human movement data*

Component	1	2	3	4	5
Euclidean principal component analysis	94·09	4·71	0·94	0·18	0·08
Tangent coordinates at the extrinsic mean	94·68	4·41	0·75	0·15	0·00
Principal geodesic analysis	94·69	4·40	0·75	0·15	0·00
Principal nested great spheres	94·72	4·39	0·74	0·15	0·00
Principal nested spheres	97·40	2·19	0·29	0·11	0·00

methods. The fitted principal nested spheres have radii 1, 0·7, 0·4, and 0·25 from the $\hat{\mathfrak{A}}_4$ to $\hat{\mathfrak{A}}_1$, respectively. Since the dimension of the corresponding shape space is 4, $\hat{\mathfrak{A}}_4$ is a great sphere and leaves no residuals, as expected.

The quadratic form of variation in the coordinates by principal nested great spheres is captured by $\hat{A}_1$, the one-dimensional principal nested sphere, as illustrated in Fig. 1. The principal mode of variation found by principal nested spheres is plotted in Fig. 3(b), where the four bold dots together represent the shape at $\hat{\mathfrak{A}}_0$, and those curves through $\hat{\mathfrak{A}}_0$ illustrate shape changes captured in $\hat{\mathfrak{A}}_1$. The nonlinear form of variation suggested by these curves can also be discerned in the raw data. Figure 3(c) shows the result of the corresponding principal nested great sphere analysis. This more conventional approach gives a less flexible mode of variation and explains less of the variation in the raw data.

Table 1 shows the proportions of variances captured in each component, as explained in §2·4 to compare the proposed analysis with some previous methods. The first is Euclidean principal component analysis, where the usual principal component analysis is carried out for the Procrustes fitted configuration matrices. The second is called tangent coordinates at the extrinsic mean, where the Procrustes fitted preshapes are orthogonally projected to the tangent coordinates at the extrinsic mean and then the usual principal component analysis is performed for the projections. This is equivalent to carrying out principal component analysis on the partial Procrustes tangent coordinates (Dryden & Mardia, 1998, Ch. 5). The last method is principal geodesic analysis of Fletcher et al. (2004), which is the usual principal component analysis of the data mapped to the tangent space at the intrinsic mean by the inverse exponential map.

Principal nested spheres is clearly the best of these methods in terms of representing maximal variation using as few components as possible, showing the value of analysing this dataset with small spheres. The other manifold methods are comparable to each other, with the intrinsic method of principal nested great spheres slightly better than the two extrinsic methods. As expected, Euclidean principal component analysis, which completely ignores the manifold nature
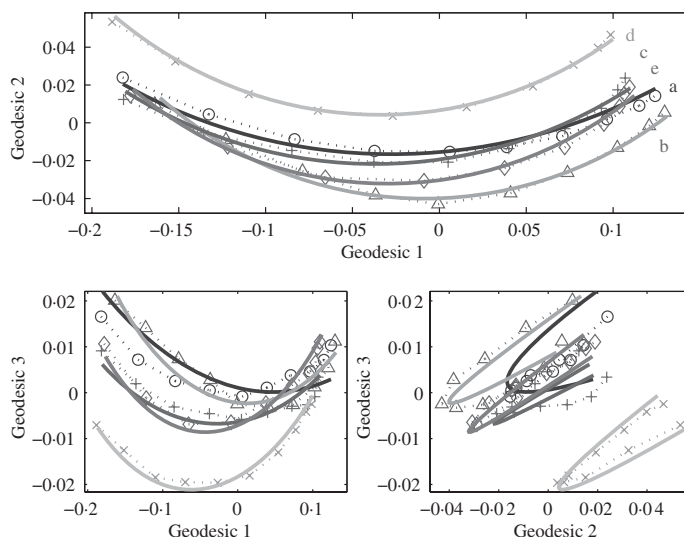
Fig. 4. Human movement data. Mode of variation represented by the small circle $\hat{\mathfrak{A}}_1$ fitted to different tasks labelled a, b, c, d, and e, plotted in coordinates of principal nested great spheres, to reveal the advantage of fitting small spheres as summarized in Table 1.

of the data, gives the worst performance. In particular, more than four components are required, compared to at most four components from manifold extensions.

Each of the five tasks in the dataset can be modelled as a small circle in the preshape space, by separately applying the analysis of principal nested spheres to the samples corresponding to each task. The results are plotted in Fig. 4. At least three geodesic components are needed to capture the variation of each task, which is well approximated by the separately fitted $\hat{\mathfrak{A}}_1$s.

5·2. *Rat skull growth*

The shape and size changes of rat skulls are described in Bookstein (1991) and studied by several other authors including Kenobi et al. (2010). The data are eight landmark locations for skulls of 21 laboratory rats observed at eight ages: days 7, 14, 21, 30, 40, 60, 90, and 150. We discard four X-rays with missing landmarks, and analyse the remaining 164 observations.

The results of our analysis, and comparison with previous methods, are summarized in Table 2 and Fig. 5. Table 2 shows the proportions of variances from our analyses and also from the geodesic-based methods. The corresponding shape space has dimension 12 but we present the first seven components. The analysis of principal nested spheres exhibits the best performance in terms of representing maximal variation using few components. The geodesic-based methods show similar performance to each other. We observe that the amount of variation captured in first two components by principal nested great spheres is about 89·99% which is comparable to 88·67% in the first component of principal nested spheres. Moreover, while the variances in the first three geodesic components sum up to 92·46%, the first two components by principal nested spheres contain 91·99%. This can be explained further by an inspection of Fig. 5. As for the previous example, Euclidean principal component analysis gives noticeably poorer performance than any of the manifold methods.

In Fig. 5(a) and (b), non-geodesic variation curving through three geodesic components is illustrated. This is well captured in $\hat{\mathfrak{A}}_2$, of dimension 2, as in Fig. 5(c). We then inspect the first two principal arcs γ_1 and γ_2 that represent the non-geodesic variation captured by principal nested

Table 2. *Proportions of variance (%), captured in rat skull growth data*

Component	1	2	3	4	5	6	7
Euclidean principal component analysis	81·88	7·94	2·50	1·77	1·19	1·05	0·95
Tangent coordinates at the extrinsic mean	82·17	7·78	2·48	1·77	1·19	1·05	0·95
Principal geodesic analysis	82·19	7·78	2·48	1·77	1·19	1·05	0·95
Principal nested great spheres	82·22	7·77	2·47	1·76	1·19	0·89	0·77
Principal nested spheres	88·68	3·31	1·67	1·38	1·06	0·89	0·77
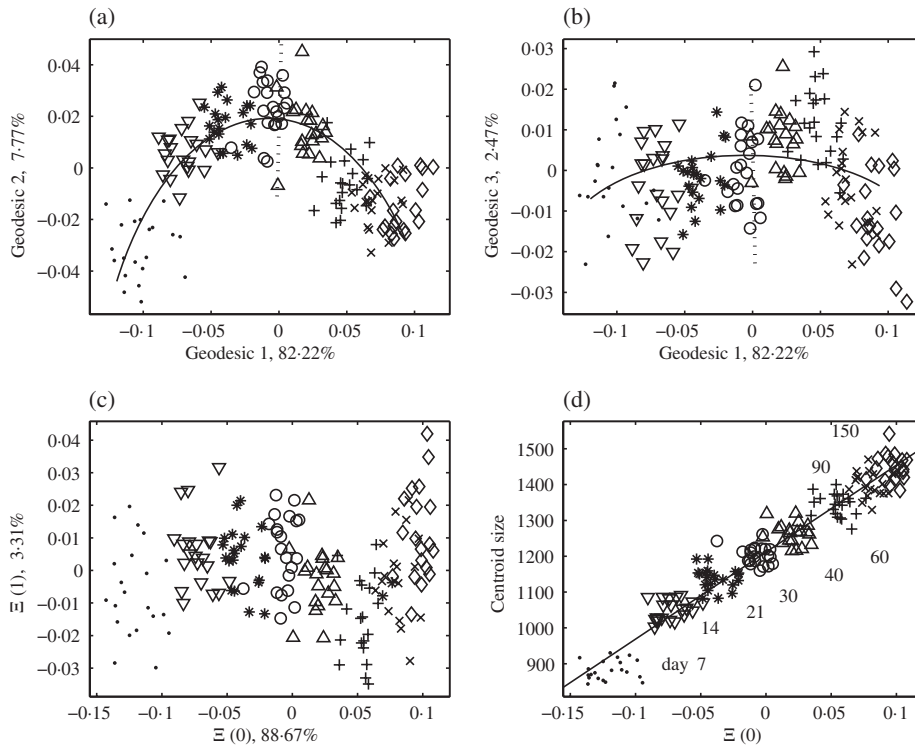
Fig. 5. Rat skull growth: (a–b) Data plotted by the coordinates by principal nested great spheres; (—) represents the first principal arc, and (· · ·) represents the second principal arc. Different symbols represent different age groups. (c) Data plotted by the coordinates by principal nested spheres. (d) Scatter-plot with centroid size and a regression line.

spheres. These principal arcs are plotted in the coordinates by principal nested great spheres in Fig. 5(a–b). The quadratic variation in (a) is captured by γ_1, illustrated by the solid curve. This leads to an understanding that the non-geodesic variation, which sums up to 90% in two geodesic components, is well captured in the one-dimensional $\hat{\mathfrak{A}}_1$, which contains 89% of the variance. The principal nested spheres capture more interesting variability in fewer components and give a concise and useful representation of the data.

Figure 5(d) shows that the shape change due to the growth of the rat is well captured by $\hat{\mathfrak{A}}_1$, which can be checked by inspecting the relation to the size of the rat skulls. The skull size is larger for older rats. The first coordinate by principal nested spheres is strongly associated with the size, and the sample correlation coefficient is 0·97. The shape change due to the different locations of $\hat{\mathfrak{A}}_1$ is illustrated in Fig. 6. The two extreme shape configurations are also overlaid, which shows the typical effect of age on shape.
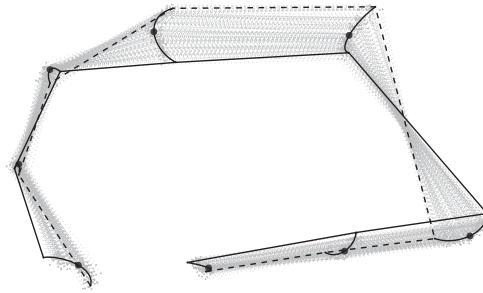
Fig. 6. Rat skull growth: the first principal mode of variation by $\hat{\mathfrak{A}}_1$. The configurations connected by $(\cdots)$ represent the shape of a typical young skull at -2 standard deviations, and (—) represents the shape of a typical old rat skull at $+2$ standard deviations.

6. PROBLEMS RELATED TO OVERFITTING OF SMALL SPHERES

For some datasets, there may be an overfitting issue in using small spheres instead of great spheres. We briefly examine some approaches to deal with this.

We have devised a procedure to apply sequential tests for each level of subsphere fitting, discussed in the Supplementary Material. Similar to the backward regression procedure, sequentially testing the significance of small sphere fits at each level may prevent overfitting. There are some caveats in the application of these tests. For example, they are built upon parametric assumptions. Furthermore, because they are not independent, simultaneous inference is complicated.

Another approach to the choice between small and great spheres is based upon Bayesian information criterion ideas. Formulas for implementing this approach are given in the Supplementary Material. As detailed there, we tried this on the human movement data and rat skull data and obtained reasonable results. These were quite similar to the results of the hypothesis testing approach discussed in the above paragraph.

A referee made the interesting suggestion of a soft decision rule that balances between small and great spheres using a penalization approach. A specific optimization problem for this method is described in the Supplementary Material.

7. DISCUSSION

A classical approach to principal component analysis for manifold data is a tangent plane approach (Fletcher et al., 2004; Dryden & Mardia, 1998, Ch. 5.5). The resulting principal geodesics then pass through the point of tangency, i.e., the Procrustes mean or the geodesic mean. Huckemann et al. (2010) pointed out that the predetermined mean may not be a good representation of the data due to the curvature of the manifold. They find that the best fitting geodesics do not necessarily pass through any predetermined mean. Jung et al. (2011) proposed to use small circles to fit the non-geodesic modes of variation when the sample space is essentially S^2. Our method builds upon these earlier works, where principal nested spheres can be viewed as a high-dimensional extension of Jung et al. (2011), and principal nested great spheres are similar to Huckemann et al. (2010).

As noted in Marron et al. (2010) and Pizer et al. (2012), our method can be viewed as a backward generalization of principal component analysis. A conventional principal component analysis approach begins with lower dimensions and builds up, that is, finding the mean first,

then the least squares line through the mean, then the second line which determines the best fitting plane and so on. We, on the other hand, sequentially reduce the dimensionality from the full data. These two approaches are equivalent in Euclidean space, because of the Pythagorean theorem, but different on manifolds. The backward approach serves as a natural viewpoint for the extension of principal component analysis to manifolds. In fact, Huckemann et al. (2010) partly used the backward approach, since their notion of mean is post-defined by two geodesic components. It should be noted that our method when applied to shape data begins with a forward approximation. This is because we pre-align preshapes with a mean, then approximate the shape space by the preshape sphere. A possible future direction of research is the development of a method for shape spaces that does not begin with any predetermined mean.

A related issue is the optimality of our fit. The fitted principal nested sphere $\hat{\mathfrak{A}}_q$ is not the q-dimensional small sphere that minimizes a squared loss. Specifically, $\hat{\mathfrak{A}}_q$ is not in general equivalent to $\operatorname{argmin}_{\mathfrak{A}_q} \sum_{i=1}^{n} \rho_d(x_i, \mathfrak{A}_q)^2$. We believe that it is also true that other methods in the literature are not optimal under this criterion. We view this optimality issue using the analogy of variable selection in linear regression. In selecting independent variables for the regression model, one can use forward selection, backward elimination or search for the best subset. Approaches of Dryden & Mardia (1998) and Fletcher et al. (2004) can be understood as forward extensions, while the analysis of principal nested spheres can be seen as a backward extension. As the set of variables chosen by backward elimination need not be optimal under some criteria, the q-dimensional principal nested sphere may be suboptimal to a global search. A best subset search will likely be computationally heavy, even when restricting only to great spheres.

A possible issue with the proposed method is model bias. We briefly discuss this in the context of the simple S^2. Suppose data points or random variables are distributed along a small circle on S^2 with an addition of noise: $x = \gamma_1(\lambda) \oplus e \in S^2$, where γ_1 is a parameterized small circle. Model bias occurs when the noise e is added at $\gamma_1(\lambda)$ as an isotropic perturbation. For example, the noise can be added as a bivariate normal random vector with variance $\sigma^2 I_2$ projected onto S^2 by the exponential map or alternatively as a von Mises–Fisher distributed random vector with location parameter $\gamma_1(\lambda)$. Under this noise structure, our procedure will treat this noise in a way that is biased, toward fitting a larger circle (Rivest, 1999). Hastie & Stuetzle (1989) discussed model bias in fitting principal curves to Euclidean data.

An approach to this issue is to use an alternative noise model where our method is unbiased. For example, starting with the radius of a small circle, one could add independent univariate Gaussian noise to the latitude and longitude. This corresponds to modelling e as a bivariate normal random vector with a diagonal covariance matrix in the spherical coordinate system with the centre of the circle at the north pole. The noise model in Jung et al. (2011) is a special case of this model.

Yet another model gives an opposite type of bias. This assumes the small circle distribution proposed by Bingham & Mardia (1978). It can be seen that our procedure results in a radius fit smaller than the true radius parameter, since the marginal distribution along the latitude angle of the small circle distribution is skewed to the centre of the circle.

While these bias issues deserve careful consideration, we believe that, at least in many important cases, they are of secondary concern relative to the central issue of potential large gains, relative to fitting small spheres instead of great spheres. This has been further shown in the examples of Jung et al. (2010) and Pizer et al. (2012).

APPENDIX

Preliminary transformations: rotation matrices

We describe a rotation matrix for moving a dataset on a sphere along a particular minimal geodesic that retains the interpoint geodesic distances after the transformation.

Suppose that a and b are unit vectors in $\mathbb{R}^m$ and we wish to move b to a along the geodesic path on the unit sphere in $\mathbb{R}^m$ that connects b to a. Amaral et al. (2007) showed that a rotation matrix is determined in a natural way.

Define $c = \{b - a(a^\mathsf{T}b)\}/\|b - a(a^\mathsf{T}b)\|$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^m$. Provided that $|a^\mathsf{T}b| < 1$, c is well defined. Let $A = ac^\mathsf{T} - ca^\mathsf{T}$. Note that $A^2 = -aa^\mathsf{T} - cc^\mathsf{T}$ and $A^3 = -A$. Since the matrix A is skew symmetric, its matrix exponential is a rotation matrix, as summarized in the following lemma, proved in Amaral et al. (2007).

LEMMA A1. *Assume that $a, b \in \mathbb{R}^m$ are unit vectors such that $|a^\mathsf{T}b| < 1$, and let A and c be defined as earlier. Then for $\theta \in (0, \pi]$, the matrix*

$$Q(b \to a, \theta) = \exp(\theta A) = I_d + \sum_{j=1}^{\infty} \frac{\theta^j}{j!} A^j$$

has the following properties:

(a) $Q(b \to a, \theta)$ *is an $m \times m$ rotation matrix;*
(b) $Q(b \to a, \theta)$ *can be written as $Q(\theta) = I_d + \sin(\theta)A + \{\cos(\theta) - 1\}(aa^\mathsf{T} + cc^\mathsf{T})$;*
(c) $Q(b \to a, \alpha)b = a$ *for $\alpha = \cos^{-1}(a^\mathsf{T}b)$; and*
(d) *for any $z \in \mathbb{R}^m$ such that $a^\mathsf{T}z = 0$ and $b^\mathsf{T}z = 0$, we have $Q(b \to a, \theta)z = z$ for any $\theta \in (0, \pi]$.*

The path given by $Q(b \to a, \theta)$ is a minimal geodesic on the sphere. If b and a are orthogonal, then

$$Q(b \to a, \theta)b = b\cos(\theta) + a\sin(\theta) \quad (-\pi/2 < \theta \leqslant \pi/2), \tag{A1}$$

which corresponds to a definition of the unit speed geodesic (Kume et al., 2007, p. 515).

We also define $R(v)$, for $v \in \mathbb{R}^m$, as a rotation matrix that rotates v to the north pole $e_m = (0, \ldots, 0, 1)^\mathsf{T}$, i.e.,

$$R(v) = Q(v \to e_m, \theta_v), \tag{A2}$$

where $\theta_v = \cos^{-1}(v^\mathsf{T}e_m)$. The last row of $R(v)$ is v^T. If $v = e_m$, then $R(v) = I_m$. Finally define $R^-(v)$ as the $m \times (m + 1)$ matrix consisting of the first m rows of $R(v)$.

REFERENCES

AHN, J., MARRON, J. S., MULLER, K. M. & CHI, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–6.

AMARAL, G. J. A., DRYDEN, I. L. & WOOD, A. T. A. (2007). Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *J. Am. Statist. Assoc.* **102**, 695–707.

Bhattacharya, R. & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31**, 1–29.

Bingham, C. & Mardia, K. V. (1978). A small circle distribution on the sphere. *Biometrika* **65**, 379–89.

Bookstein, F. L. (1991). *Morphometric Tools for Landmark Data*. Cambridge: Cambridge University Press.

Buss, S. R. & Fillmore, J. P. (2001). Spherical averages and applications to spherical splines and interpolation. *ACM Trans. Graph.* **20**, 95–126.

Dryden, I. L. (2005). Statistical analysis on high-dimensional spheres and shape spaces. *Ann. Statist.* **33**, 1643–65.

Dryden, I. L. & Mardia, K. V. (1998). *Statistical Shape Analysis*. Chichester: Wiley.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Fisher, N. I., Lewis, T. & Embleton, B. J. J. (1993). *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.

Fletcher, P. T., Lu, C., Pizer, S. M. & Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.* **23**, 995–1005.

Fréchet, M. (1944). L'intégrale abstraite d'une fonction abstraite d'une variable abstraite et son application à la moyenne d'un élément aléatoire de nature quelconque. *Rev. Sci.*, 483–512.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10**, 215–310.

Hall, P., Marron, J. S. & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.* B **67**, 427–44.

Hastie, T. & Stuetzle, W. (1989). Principal curves. *J. Am. Statist. Assoc.* **84**, 502–16.

Helgason, S. (2001). *Differential geometry, Lie groups, and Symmetric spaces*, vol. 34 of *Graduate Studies in Mathematics*. Rhode Island: American Mathematical Society.

Huckemann, S., Hotz, T. & Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20**, 1–58.

Huckemann, S. & Ziezold, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. Appl. Prob.* **38**, 299–319.

Jung, S., Foskey, M. & Marron, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. App. Statist*. **5**, 578–603.

Jung, S., Liu, X., Marron, J. S. & Pizer, S. M. (2010). Generalized PCA via the backward stepwise approach in image analysis. In *Brain, Body and Machine: Proc. Int. Symp. 25th Anniversary McGill Univ. Centre Intel. Mach.*, Ed. J. Angeles, B. Boulet, J. Clark, J. Kovecses & K. Siddiqi, vol. 83 of Advances in Intelligent and Soft Computing.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **30**, 509–41.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.* **16**, 81–121.

Kendall, D. G., Barden, D., Carne, T. K. & Le, H. (1999). *Shape and Shape Theory*. Chichester: Wiley.

Kenobi, K., Dryden, I. L. & Le, H. (2010). Shape curves and geodesic modelling. *Biometrika* **97**, 567–84.

Kume, A., Dryden, I. L. & Le, H. (2007). Shape-space smoothing splines for planar landmark data. *Biometrika* **94**, 513–28.

Mardia, K. V. & Jupp, P. E. (2000). *Directional Statistics*. Chichester: Wiley.

Marron, J. S., Jung, S. & Dryden, I. L. (2010). Speculation on the generality of the backward stepwise view of PCA. In *Proc. MIR 2010: 11th ACM SIGMM Int. Conf. Multimedia Info. Retrieval, Assoc. Comp. Mach., Inc., Danvers, MA*, pp. 227–30.

Pizer, S. M., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J. N., Huckemann, S. & Marron, J. S. (2012). Nested sphere statistics of skeletal models. To appear in *Proc. Dagstuhl Workshop on Innovations for Shape Analysis: Models and Algorithms*. Springer Lecture Notes in Computer Science.

Rivest, L.-P. (1999). Some linear model techniques for analyzing small-circle spherical data. *Can. J. Statist.* **27**, 623–38.

Scales, L. E. (1985). *Introduction to Nonlinear Optimization*. New York: Springer.

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