Biometrika (2012), **xx**, x, *pp*. 1–24

Printed in Great Britain

Supplementary Material for Analysis of Principal Nested Spheres

BY SUNGKYU JUNG

Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.
sungkyu@pitt.edu

IAN L. DRYDEN

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA.
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, USA.
marron@unc.edu

1. PREVENTION OF OVERFITTING BY SEQUENTIAL TESTS

1·1. *Significance of small sphere fitting*

The significance of small spheres against a great sphere is discussed. We propose a test procedure consisting of two different tests for each level of subsphere fitting. Similar to the backward regression procedure, sequentially testing small spheres at each layer may prevent overfitting.

There are two cases where a great sphere provides more a appropriate fit to the data, yet the sum of squared residuals is minimized by a small sphere. The first case is where the true major variation is along a great sphere, an example of which on S^2 is illustrated in Fig. 1a. The second case is when the underlying distribution is rotationally symmetric with a single mode, so that there is no major variation along any direction. An example of such a distribution is $N(0, I_k)$ in a linear space, or the von Mises–Fisher distribution on S^d (Fisher, 1953; Mardia & Jupp, 2000), as illustrated in Fig. 1b. In the latter situation, small spheres centered at the point of isotropy are frequently obtained, which do not give a useful decomposition.

We have developed two different tests to handle these cases. The first is a likelihood ratio test for the detection of the first case as exemplified in Fig. 1a, which tests the significance of the reduction of residual variances. The second is a parametric bootstrap test aimed at the second case as depicted in Fig. 1b, which tests the isotropy of the underlying distribution. A detailed description of the tests is given in the following subsections. A procedure to apply these tests in principal nested spheres fitting is then discussed in Section 1·4. As a reviewer pointed out, the hierarchical structure of test hypotheses makes the control of type I error difficult, which is also discussed in Section 1·4. Attempts to avoid the dependent test structure are contained in Section 1·5 to Section 1·7.

1·2. *Likelihood ratio test*

We define a likelihood ratio statistic for each level to test sequentially the significance of the small sphere fitting against the great sphere.

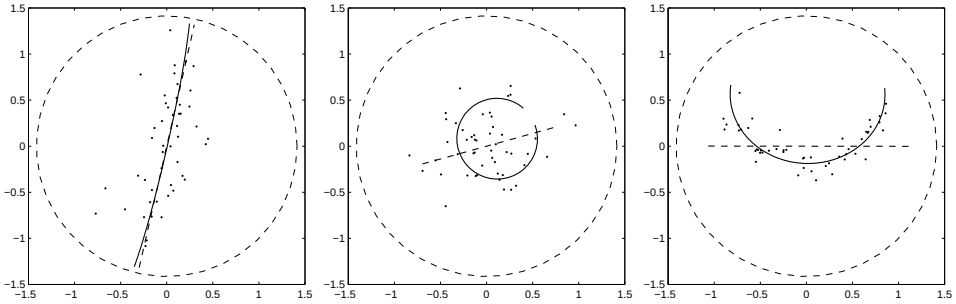

Fig. 1. Simulated data examples on S^2 projected with an equal area projection, and the fitted best small (solid) and great (dashed) spheres, which are arcs in this two-dimensional case. (a) The likelihood ratio test gives the p-value 0.565, while the p-value of the bootstrap test is 0. The likelihood ratio test detects the overfitting. (b) The likelihood ratio test leads to p-value ≈ 0, and the bootstrap p-value is 0.82. The bootstrap test detects the overfitting. (c) When the fitted small sphere (circle) is not overfitted, both tests give very small p-values. This ensures that the small sphere is not overly fitted.

For the kth level of the procedure, where A_{d-k} is fitted to $x_1, \ldots, x_n \in S^{d-k+1}$, we assume that the deviations $\xi_i \equiv \xi_{i,d-k}$ of the samples x_i from the subsphere $A_{d-k}(v, r)$ are independently distributed as $N(0, \sigma^2)$. It makes more sense when a truncated normal distribution on a range $[-\pi/2, \pi/2]$ is assumed. However unless the data spread too widely, the distribution will be approximately normal. Thus we use the approximate likelihood function of (v, r, σ^2), given by

$$L(v, r, \sigma^2 | x_1^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \{\rho(x_i, v) - r\}^2 \right],$$

where ρ is the geodesic distance function on S^{d-k+1}. The approximate maximum likelihood estimator of (v, r) coincides with $(\hat{v}, \hat{r})$, the solution of (3) in the main article, and the approximate maximum likelihood estimator of σ^2 is given by $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \xi_i(\hat{v}, \hat{r})^2$, which is obtained by differentiating the log-likelihood function and setting the derivative equal to zero.

The null hypothesis H_{0a} is that the true model is a great sphere, while the alternative is that the true model is a small sphere. In other words,

$$H_{0a} : r = \pi/2 \text{ vs } \; H_{1a} : r < \pi/2,$$

which is tested by a likelihood ratio test. The maximum likelihood estimator of (v, r, σ^2) under H_{0a} is given by $(\hat{v}^0, \pi/2, \hat{\sigma}_0^2)$, where $\hat{v}^0$ minimizes the sum of squared residuals of (3) in the main article with $r = \pi/2$, and $\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^{n} \xi_i(\hat{v}^0, \pi/2)^2$. The log-likelihood ratio is $(\hat{\sigma}^2/\hat{\sigma}_0^2)^{-n/2}$. Then using Wilks' theorem, for large samples $n \log(\hat{\sigma}^2/\hat{\sigma}_0^2) \sim \chi_1^2$ under H_{0a}, and the test rejects H_{0a} in favor of H_{1a} for large values of $n \log(\hat{\sigma}^2/\hat{\sigma}_0^2)$.

1·3. *Parametric bootstrap test*

For each level of principal nested spheres fitting, suppose $X \in S^m$ has a distribution function F_X. We wish to test for the underlying distribution F_X,

$$H_{0b} : F_X \text{ is an isotropic distribution with a single mode, vs } \; H_{1b} : \text{ not } H_{0b}.$$

We develop a parametric bootstrap test with an assumption of the von Mises–Fisher distribution. The von Mises–Fisher distribution is an analogue of the normal distribution on the unit sphere with concentration parameter κ and directional parameter μ, denoted as $\text{vMF}(\mu, \kappa)$.

We build a test statistic that is large when F_X is neither isotropic nor having a single mode. For this purpose we derive the following test statistic. Given $x_1, \ldots, x_n \in S^m$, estimate the best fitting subsphere $A(\hat{v}, \hat{r})$ as done in (3) in the main article. Let $\zeta_i = \rho_d(x_i, \hat{v}) = \cos^{-1}(x_i^T \hat{v})$ be the radial distances from the axis of the subsphere. Then the test statistic is the coefficient of variation of ζ,

$$Z = Z(x_1, \ldots, x_n) = \frac{\frac{1}{n}\sum_{i=1}^n \zeta_i}{\left\{\frac{1}{n-1}\sum_{i=1}^n (\zeta_i - \bar{\zeta})^2\right\}^{1/2}}.$$

The next step is to estimate the null distribution of this statistic. Under H_{0b}, F_X is $\text{vMF}(\mu, \kappa)$. The unknown parameters are estimated from the data. The location parameter μ is estimated by a standard likelihood approach (Mardia & Jupp, 2000). For an estimate of κ, Banerjee et al. (2005) empirically derived an approximation of the maximum likelihood estimate of κ. The estimates are

$$\hat{\mu} = \frac{r}{\|r\|} = \frac{\sum_{i=1}^n x_i}{\|\sum_{i=1}^n x_i\|},\ \hat{\kappa} \approx \frac{\bar{r}(d+1) - \bar{r}^3}{1 - \bar{r}^2},$$

where $\bar{r} = \|r\|/n$. We generate $B \geq 100$ random samples of size n from $\text{vMF}(\hat{\mu}, \hat{\kappa})$ and calculate $Z_1, \ldots, Z_B$. The test rejects H_{0b} with a significance level α if

$$\frac{1}{B}\sum_{i=1}^B 1_{\{Z_i > Z\}} < \alpha.$$

1·4. *Application procedure*

A sequence of sample principal nested spheres is obtained by iterative fitting of subspheres. In each layer of subsphere fitting, both of the tests developed in the previous sections will be used, due to the observation in Fig. 1. We first illustrate how these tests are applied to the examples in Fig. 1, then propose a procedure to apply the tests to the principal nested spheres fitting procedure.

Some typical data examples on the 2-sphere and the results of the two tests are illustrated in Fig. 1. When the true major variation is along a great circle, as in Fig. 1a, the likelihood ratio test works well and accepts H_{0a}, but the bootstrap test rejects H_{0b}. On the other hand, when the underlying distribution is von Mises–Fisher, the likelihood ratio test rejects H_{0a} in favor of H_{1a}. However, the best fitting small sphere is frequently inappropriate, as shown in Fig. 1b. The bootstrap test accepts H_{0b} and thus can be used to detect such a case. Therefore, in order to prevent overfitting, we proposed to apply sequentially both tests at each level of subsphere fitting. In a case where a true variation is along a small sphere, both tests reject the null hypotheses, and we ensure that the small subsphere is not overfitted.

At each level of subsphere fitting, we use the following testing procedure to test the significance of small subsphere fitting.

1. Test H_{0a} versus H_{1a} by the likelihood ratio test. If H_{0a} is accepted, then fit a great sphere with $r = \pi/2$ and proceed to the next layer.

2. If H_{0a} is rejected, then test the isotropy of the distribution by the parametric bootstrap test. If H_{0b} is accepted, then use great spheres for all further subsphere fittings.

3. If both tests reject the null hypotheses, then use the fitted small sphere for the decomposition.

Note that in Step 2, when H_{0b} is accepted, we use great sphere fitting not only for the corresponding level, but also for all further levels with smaller dimensions. This is because once H_{0b} is accepted, the underlying distribution at the level is assumed to be von Mises–Fisher. An analogy in Euclidean space is $N(0, I_k)$ where a non-linear mode of variation is meaningless. Therefore, great spheres are used for all further nested spheres, without further application of tests.

Note that for S^d, we test at most $2(d-1)$ hypotheses, and the hypotheses are hierarchically structured. The hypotheses and test statistics of the $(m+1)$th tests are dependent on the result of the mth tests. Moreover, the first hypotheses are far more important than the subsequent ones in a sense that when an error, of either type I or type II, is committed in the first hypothesis, all subsequent hypotheses are false. A simple Bonferroni's method controlling family-wise error rate in a multiple testing problem is not applied to our tree-structured hypotheses problem. While we do not attempt to properly analyze the tree-structured hypotheses problem, some literature related to the problem is pointed out, as follows.

A Markov decision process (Puterman, 1994) can be used to analyze the errors in the tree-structured hypotheses problem. A Markov decision process model, translated into our problem, contains the state space S, which contains the data $x_i \in S^d$ $(i = 1, \ldots, n)$ or the data projected on small or great spheres $x^{\dagger}_{i(k)} \in S^{d-k+1}$, a set of two actions $A = \{$accept great sphere, reject great sphere, i.e., accept small sphere$\}$, and a real-valued reward function $R : S \times A \to \mathbb{R}$ that accounts for the type I or type II error of the tests. The effect of the action, rejection or acceptance of the null hypotheses, depends only on the current state $x^{\dagger}_{i(k)} \in S^{d-k+1}$ and not on the prior history of actions. Thus, the Markov assumption is sensible.

Another related testing procedure is the so-called gatekeeping testing procedure (Bauer et al., 1998; Westfall & Krishen, 2001; Dmitrienko et al., 2007, 2008). The gatekeeping procedure provides a way to test multiple hypotheses in a hierarchical manner where the subsequent hypotheses are only examined if all or some hypotheses in the primary set of hypotheses are rejected. In our case, the families of hypotheses are summarized in Table 1, in the case of $d = 3$.

Table 1. *Families of hypotheses in the sequential testing procedure for the $d = 3$ case.*

$S^3 \to S^2$:	H^1_{0a}	H^1_{0b}
$S^2 \to S^1$:	H^2_{0a}	H^2_{0b}

The secondary family of hypotheses, H^2_{0a} and H^2_{0b}, in the second row is only examined if the H^2_{0b} in the first row is rejected. When considering only the second hypotheses H_{0b} in each family, the procedure is similar to the serial gatekeeping procedure (Bauer et al., 1998; Westfall & Krishen, 2001). However, the second family is examined regardless of the rejection of the first hypothesis H^1_{0a}. To the best of our knowledge, such structure has not been studied.

1·5. *Bayesian information criterion*

Another approach is to utilize the Bayesian information criterion under Gaussian assumption. This approach still assumes a distribution, but is convenient enough to use as a rule-or-thumb method to indicate whether a small circle is overfitted. For the kth level of the small sphere fitting, where A_{d-k} is fitted to $x_1, \ldots, x_n \in S^{d-k+1}$, we assume that the deviations ξ_i from the subsphere $A_{d-k}(v, r)$ are independent $N(0, \sigma^2)$. Then the appropriate Bayesian information criterion would be $-2\log\{L(v, r, \sigma^2 | x_1^n)\} + k\log(n)$ and by removing constant terms we compare

the Bayesian information criterion from small sphere fitting

$$\mathrm{BIC}_s = n\log(\hat{\sigma}^2) + (d+1)\log(n)$$

with the Bayesian information criterion from great sphere fitting

$$\mathrm{BIC}_g = n\log(\hat{\sigma}_0^2) + d\log(n),$$

where $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\xi_i(\hat{v},\hat{r})$ is the maximum likelihood estimator of σ^2 under the small sphere model and $\hat{\sigma}_0^2 = n^{-1}\sum_{i=1}^{n}\xi_i(\hat{v}_0,\pi/2)$ is the maximum likelihood estimator under the great sphere model.

1·6. *Data Example*

The two proposed methods, the sequential tests and the Bayesian information criteria, are compared in some data sets. In the sequential tests, we set bootstrap sample size $B = 100$ and the significance level $\alpha = 0.05$. Two datasets, the human movement data contained in Kume et al. (2007) and the rat skull growth data in Bookstein (1991) are analyzed and the numerical results are presented in Table 2.

In Table 2, when the human movement data are analyzed, the tests are applied to each of three layers. Both tests result in p-values less than the significance level, therefore we chose to use the small sphere. The Bayesian information criteria of the small sphere model BIC_s are always smaller than the Bayesian information criteria of the great sphere model BIC_g, which leads to the choice of small spheres. In the simple human movement data analysis, the two methods lead to the same decision.

Table 2 also contains the results from rat skull data analysis. The dimension of the preshape sphere is 13, and is higher than the dimension of the human movement data case where the dimension is five, thus leading to a much more complicated result. Let us focus on the left columns for the sequential tests. When fitting the 11-sphere from the 12-sphere, both tests reject the null hypotheses in favor of the small sphere. Now when fitting the 10-sphere, the likelihood ratio test results in a p-value greater than 0.05. As described in Section 1·4, a great sphere is chosen and the parametric bootstrap test is skipped. Note that in the last three layers, the tests choose small spheres. To compare this result with the decisions using the Bayesian information criterion, note that when a 9-sphere is fitted, the Bayesian information criterion chooses the great sphere but the test results are in favour of the small sphere fitting. Because the principal nested spheres are nested, any further fitting will be different in these two decision rules. However, we find a pattern in both of the decisions. In fact, the difference is only found when fitting the 9-sphere.

We inspected this difference by taking a closer look at Table 3. Table 3 shows the radii of principal nested spheres when using the two different rules. Note that the radii of $\mathfrak{A}_1, \mathfrak{A}_2$, and $\mathfrak{A}_3$ are similar, as seen in the last three rows. These three components contain more than 90% of the variance, and we see that there is virtually no difference in scatter plots of the principal scores, which we omit.

1·7. *Soft decision rule*

The test procedure and the Bayesian Information Criterion approach we discussed before have some caveats. Both are based on a specific parametric assumption, which may be seen as inappropriate. Moreover, simultaneous inference is complicated because the tests are not independent. In this section, an interesting approach, which does not depend on any complicated assumptions, is introduced.

Table 2. *Decisions by sequential tests and the Bayesian information criterion*

Human movement

	p_{LRT}	p_{vMF}	Decision	BIC_s	BIC_g	Decision
4-sphere to 3-sphere	0.00	0	SMALL	-542.68	-536.18	SMALL
3-sphere to 2-sphere	3.2×10^{-10}	0	SMALL	-464.40	-428.78	SMALL
2-sphere to 1-sphere	2.8×10^{-7}	0	SMALL	-310.64	-288.17	SMALL

Rat skull growth

	p_{LRT}	p_{vMF}	Decision	BIC_s	BIC_g	Decision
12-sphere to 11-sphere	0.01	0	SMALL	-1762.39	-1759.71	SMALL
11-sphere to 10-sphere	0.10		GREAT	-1612.67	-1615.10	GREAT
10-sphere to 9-sphere	0.03	0	SMALL	-1587.57	-1587.96	GREAT
9-sphere to 8-sphere	0.01	0	SMALL	-1560.01	-1556.32	SMALL
8-sphere to 7-sphere	0.77		GREAT	-1400.23	-1405.22	GREAT
7-sphere to 6-sphere	2.1×10^{-4}	0	SMALL	-1374.63	-1365.39	SMALL
6-sphere to 5-sphere	0.06		GREAT	-1212.67	-1214.02	GREAT
5-sphere to 4-sphere	0.82		GREAT	-1190.21	-1195.24	GREAT
4-sphere to 3-sphere	1.0×10^{-6}	0	SMALL	-1153.95	-1134.76	SMALL
3-sphere to 2-sphere	0.00	0	SMALL	-947.29	-942.59	SMALL
2-sphere to 1-sphere	0.00	0	SMALL	-824.96	-742.40	SMALL

The p-values from the likelihood ratio test and the parametric bootstrap test are denoted by p_{LRT} and p_{vMF}.

Table 3. *Rat skull data: difference between decisions by sequential tests and the Bayesian information criterion.*

	Tests		BIC	
	decision	radii	radii	decision
13-sphere		1.000	1.000	
12-sphere		1.000	1.000	
11-sphere	SMALL	0.762	0.762	SMALL
10-sphere	GREAT	0.762	0.762	GREAT
9-sphere	SMALL	0.604	0.762	GREAT
8-sphere	SMALL	0.474	0.499	SMALL
7-sphere	GREAT	0.474	0.499	GREAT
6-sphere	SMALL	0.322	0.324	SMALL
5-sphere	GREAT	0.322	0.324	GREAT
4-sphere	GREAT	0.322	0.324	GREAT
3-sphere	SMALL	0.187	0.188	SMALL
2-sphere	SMALL	0.179	0.179	SMALL
1-sphere	SMALL	0.129	0.129	SMALL

A family of important variations of principal nested spheres can be developed by either adding a penalty term or modifying the least squares term in (3) in the main article. In particular, we write a general optimization problem for each subsphere fitting as

$$n^{-1}\sum_{i=1}^{n} L(x_i, v, r) + \lambda j(r), \tag{1}$$

where L is loss function, j is a penalty function and $\lambda > 0$ is a tuning parameter. The original least squares problem (3) in the main article is a special case of (1) given by $L(x_i, v, r) = \{\rho_d(x_i, v) - r\}^2$ and $j(r) = 0$. One may replace the squared loss by the absolute loss, $L(x_i, v, r) = |\rho_d(x_i, v) - r|$, for more robust fitting.

A soft decision rule between small and great spheres fitting can be developed by the penalization approach. A reasonable choice of the penalty functions would be $j_2(r) = (\pi/2 - r)^2$ or $j_1(r) = |\pi/2 - r|$. Note that indeed $j_1(r) = \pi/2 - r$ because $0 < r \leq \pi/2$. These penalty functions penalize the departure of r from $\pi/2$, i.e. the departure of the fitted sphere from a great sphere. As a referee pointed out, this penalized approach will lead to a compromised fit of small and great spheres, leading to a soft decision between small and great spheres.

The algorithm we proposed to solve the optimization problem, (3) in the main article, may be adapted to solve (1), for a special case of $L(x_i, v, r) = \{\rho_d(x_i, v) - r\}^2$ and $j(r) = (\pi/2 - r)^2$. Starting with an initial value, one can alternate the update of v and r. In particular, given $\hat{r}$, the objective function is numerically minimized with respect to v (by e.g. the Levenberg-Marquert algorithm), then given $\hat{v}$, the minimization of the objective function with respect to r is simple, since the objective function is quadratic in r.

We believe a thorough investigation in this direction will lead to important and significant findings.

2. Additional data analysis examples

2·1. *Migration path of an elephant seal*

The use of the proposed method for spherical data is illustrated. Since the sample space is now S^d, we do not need to pre-align the data points as done in the shape data analysis in Section 5 of the main article. The data analysis in this section utilizes the proposed test procedure.

As the simplest example, consider a dataset on the usual sphere S^2. The dataset consists of $n = 73$ daily location measurements of a migrating female elephant seal, presented in Brillinger & Stewart (1998) and also discussed in Rivest (1999). The seal migrates from the southern California coast to the eastern mid-north Pacific Ocean. Of interest is to investigate whether the seal migrates along a great circle path, i.e. the shortest distance path. Note that Brillinger & Stewart (1998) and Rivest (1999) have analyzed this dataset in greater detail. We briefly re-analyze this data set with the proposed hypothesis test.

Fig. 2 shows the path of the migration, including both the forward journey and the return trip. Since the dataset in latitude-longitude coordinates can be converted to points on the unit sphere, it is viewed as a set of points on S^2 and we fitted principal nested spheres, with only one nested circle. We fit the best fitting great circle and small circle with fitted distance $\hat{r} = 75.45°$. The likelihood ratio test developed in Section 1·2 results in p-value 0.0851 (with $H_{0a} : r = 90°$, the great circle). Therefore, the migration is not significantly different at the level $\alpha = 0.05$ from a great circle path, which is consistent with the results from Brillinger & Stewart (1998) and Rivest (1999).

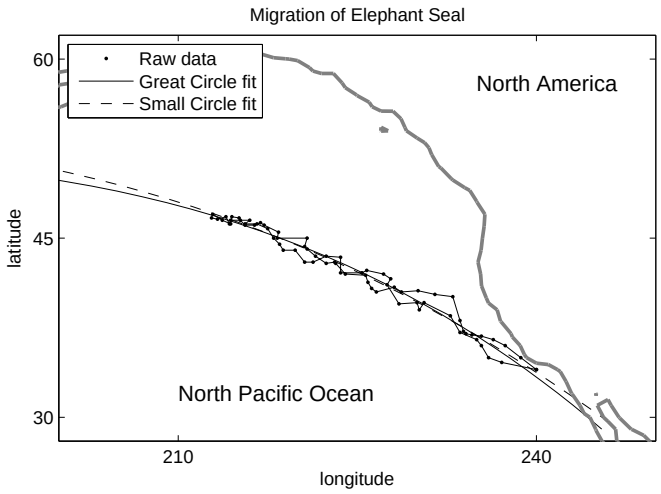

Fig. 2. Daily observations of migration path of an elephant seal, in the latitude-longitude coordinates, and the great circle and small circle fit of the data.

2·2. *River and sea sand grains*

We consider sand grain outlines that can be parameterized as a set of points in a hypersphere. The dataset was originally analyzed in Kent et al. (2000), and consists of outlines of sand grains from two-dimensional views. There are $n_1 = 25$ river and $n_2 = 24$ sea sand grains. We illustrate an application of principal nested spheres, and use of the Euclidean-type representation to test for group mean difference.

The outline of each sand grain is represented in polar coordinates $(r_1, \ldots, r_k)$ at each equally spaced angle $(\theta_1, \ldots, \theta_k)$, with $k = 20$. The scale is removed so that $\sum_{i=1}^{k} r_i^2 = 1$. The origin for each sand grain is its center of gravity, and we keep the grains fixed in the orientation that they were recorded, rather than removing rotation as in the shape analysis of Kent et al. (2000). With θ_i fixed throughout the samples as $\theta_i = (i-1)2\pi/k$, $r = (r_1, \ldots, r_k)$ on the unit $(k-1)$-sphere represents the scale invariant profile of a registered sand grain. Note that the size of river sand grains are typically larger than that of sea sand (Kent et al., 2000), but this analysis focuses on the variability in the scale invariant profiles of sand grains.

To the 49 $(= n_1 + n_2)$ data points on the 19-sphere, we fitted sample principal nested spheres, with significance level $\alpha = 0.05$ for every test applied. The small sphere fitting is significant for only three layers of the procedure, when fitting A_{18}, A_{17} and A_{11}, with both p-values less than 0.05. The one-dimensional principal nested sphere $\mathfrak{A}_1$ has radius 0.8738, suggesting that the captured principal variation is not much more curved than geodesics. The principal nested sphere leads to the Euclidean-type representation X_{PNS} of the dataset, in a way that the curved principal arcs are flattened. The first three coordinates in X_{PNS} are used for visualization of major variation as in Fig. 3.

To test the group mean difference between river and sea sand grains, we can use any Euclidean space based test procedure applied to X_{PNS}. Since we do not have any prior information on the underlying distribution, it makes sense to use a nonparametric permutation test. In particular, we use the Direction-Projection-Permutation test, described in Wichers et al. (2007), that is used to test the group mean difference for multivariate data. The test finds a direction vector pointing from one group to the other, and computes a t-statistic of the projected values onto the direction.

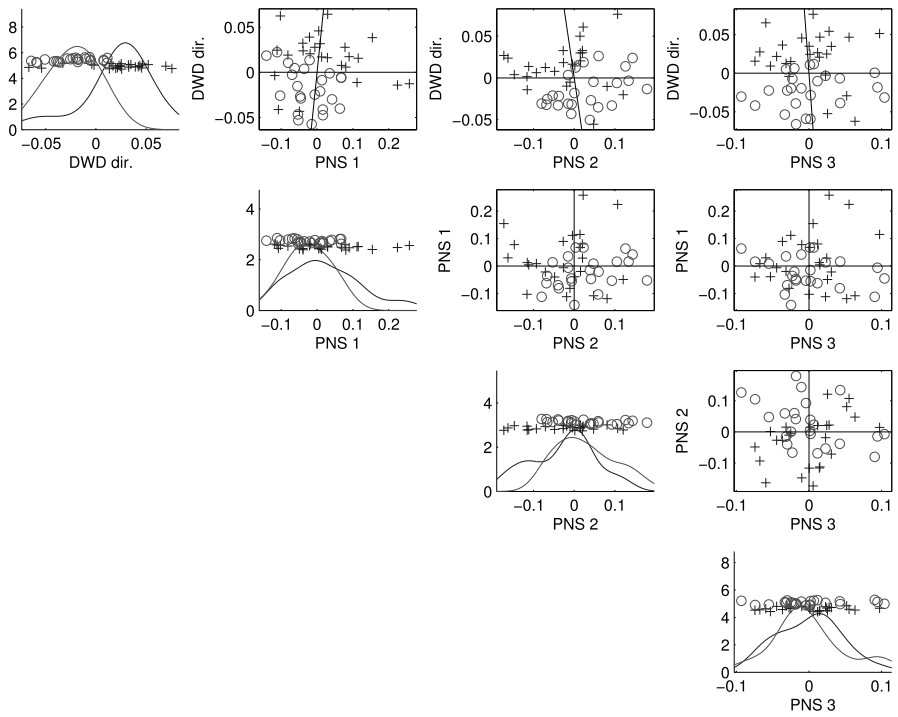

Fig. 3. Scatterplot matrix of sand grain data set, by the distance weighted discrimination direction, labeled as DWD, and the first three coordinates of X_{PNS}. (+: river sand grains, ∘: sea sand grains) Diagonal entries are jitter plots of one-dimensional projections with kernel density estimates for each group. The distance weighted discrimination direction separating the two groups is found in the Euclidean space, X_{PNS}.

The null distribution of the t-statistic is found by permutation of group labels. We have used the test with the direction vector u as the distance weighted discrimination direction (Marron et al., 2007). The distance weighted discrimination is a classification tool that separates two groups with more generalizability than, e.g., the support vector machine (Vapnik, 1995). The subspace formed by u and first three coordinates of X_{PNS} is illustrated as a scatterplot matrix in Fig. 3. Although the first three coordinates of X_{PNS} do not give a visual separation between the groups, X_{PNS} turns out to be a useful Euclidean space for linear classification methods such as distance weighted discrimination.

In our analysis, the test with 1000 permutations rejects the null hypothesis of equal group means with p-value 0.0292. The difference of shapes in the overlay of the outlines of sand grains, illustrated in Fig. 4, is statistically significant.

3. Geometry of nested spheres

3·1. *Rotation matrices*

For $a, b \in S^{m-1} \subset \mathbb{R}^m$ such that $|a^T b| < 1$, the rotation matrix that rotates b to the direction of a with angle θ is denoted by $Q(b \to a, \theta)$, and the rotation matrix that rotates v to the north pole $e_m = (0, \ldots, 0, 1)^T$ is denoted by $R(v)$ as defined in the Appendix.

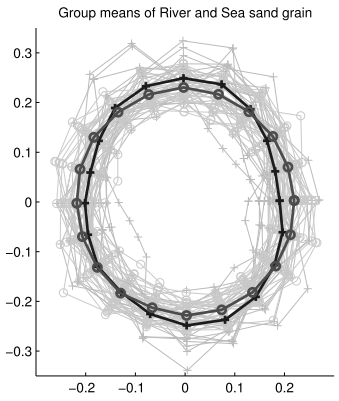

Fig. 4. Overlaid outlines of 25 river sand grains (+) and 24 sea sand grains (∘) with the group means (thick outlines) identified with the geodesic mean of each group. The DiProPerm test rejects a null hypothesis of equal group means with p-value 0.0292

The following lemma is useful in developing further the geometry of nested spheres.

LEMMA 1. *Assume that $a, b \in \mathbb{R}^m$ are unit vectors such that $|a^T b| < 1$, and let $\theta \in (0, 2\pi]$.*
(a) Let R be an $m \times m$ rotation matrix. Then, $Q(Rb \to Ra, \theta) = RQ(b \to a, \theta)R^T$. Equivalently, $Q(R^T b \to R^T a, \theta) = R^T Q(b \to a, \theta)R$.
(b) Let $a' = (a^T, 0)^T$, $b' = (b^T, 0)^T$. Then the $(m+1) \times (m+1)$ rotation matrix that moves b' to a' is given by

$$Q(b' \to a', \theta) = \begin{bmatrix} Q(b \to a, \theta) & 0_{m\times 1} \\ 0_{1\times m} & 1 \end{bmatrix},$$

where $0_{m\times n}$ is the $m \times n$ matrix of zeros.

3·2. *Geometry of Subspheres*

The nested spheres of S^d are lower dimensional submanifolds of S^d, each of which is isomorphic to the unit spheres in different dimensions. The subsphere A_{m-1} of S^m ($m \geq 2$) induces the nested spheres.

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

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

The $A_{m-1}(v, r)$ is the boundary of the geodesic ball in S^m with center v and radius r. The v is said to be orthogonal to A_{m-1} in a sense of the following lemma.

LEMMA 2. *(a) For any $x, y \in A_{m-1}$, $(x - y)^T v = 0$.*
(b) $x \in A_{m-1}$ if and only if $v^T\{x - \cos(r)v\} = 0$ and $\|x\| = 1$.

A subsphere A_{m-1} is essentially an $(m-1)$-dimensional sphere. The following properties of subspheres give the mathematical background to treat A_{m-1} as S^{m-1}.

PROPOSITION 1. *Let $A_{m-1}(v, r)$ be a subsphere in S^m. Then*

(a) (A_{m-1}, ρ_m) *is isomorphic to* (S^{m-1}, ρ_{m-1}) *with an isomorphism* $f : A_{m-1} \longrightarrow S^{m-1}$ *defined by*

$$f(x) = \frac{1}{\sin(r)} R^-(v)x,\ x \in A_{m-1}$$

with inverse

$$f^{-1}(x^\dagger) = R^T(v) \begin{bmatrix} \sin(r) \cdot x^\dagger \\ \cos(r) \end{bmatrix},\ x^\dagger \in S^{m-1},$$

where $R(v)$ *is the* $(m+1) \times (m+1)$ *rotation matrix that moves* v *to the north pole,* $R^-(v)$ *is the* $m \times (m+1)$ *matrix consisting of the first* m *rows of* $R(v)$.

(b) Let $\rho^*_{m-1}(x, y) = \sin(r)\rho_{m-1}\{f(x), f(y)\}$. *Then* ρ^*_{m-1} *is a metric on* A_{m-1}.

(c) (A_{m-1}, ρ^*_{m-1}) *is isometric to* $(S^{m-1}, \sin(r)\rho_{m-1})$.

(d) The two metrics ρ_m *and* ρ^*_{m-1} *are equivalent, in a sense that the following inequalities*

$$\rho_m(x, y) \le \rho^*_{m-1}(x, y) \le \frac{\pi \sin(r)}{2r} \rho_m(x, y)$$

hold for all $x, y \in A_{m-1}$ *and both equalities hold if and only if* $r = \pi/2$ *or* $x = y$.

(e) $\rho^*_{m-1}(x, y) - \rho_m(x, y) \le \pi \sin(r) - 2r$ *for all* $x, y \in A_{m-1}$.

The $\rho^*_{m-1}(x, y)$ can be interpreted as the length of a minimal arc in A_{m-1} that joins x, y. Precisely, the minimal arc is the image by f^{-1} of the minimal geodesic segment joining $f(x)$ and $f(y)$. Let $x^\dagger = f(x)$, $y^\dagger = f(y)$. Then the geodesic segment is given by

$$\Gamma = \{\gamma(\theta) = Q(x^\dagger \to y^\dagger, \theta)x^\dagger : \theta \in [0, \cos^{-1}(x^{\dagger T} y^\dagger)]\}.$$

By Lemma 1, we have for any $\theta \in [0, \cos^{-1}(x^{\dagger T} y^\dagger)]$,

$$\begin{aligned} f^{-1}(\gamma(\theta)) &= R(v)^T \begin{bmatrix} \sin(r) Q(x^\dagger \to y^\dagger, \theta) x^\dagger \\ \cos(r) \end{bmatrix} \\ &= R(v)^T \begin{bmatrix} Q(x^\dagger \to y^\dagger, \theta) & 0_{1\times m} \\ 0_{m\times 1} & 1 \end{bmatrix} R(v) R(v)^T \begin{bmatrix} \sin(r) x^\dagger \\ \cos(r) \end{bmatrix} \\ &= Q(x_p \to y_p, \theta)x, \end{aligned}$$

where

$$x_p = R(v)^T \begin{bmatrix} x^\dagger \\ 0 \end{bmatrix} = \frac{x - \cos(r)v}{\sin(r)},$$

and y_p is defined similarly. One can check that $\rho_m(x_p, y_p) = \rho_{m-1}(x^\dagger, y^\dagger)$ and $Q(x_p \to y_p)x = y$. Thus the arc $\{Q(x_p \to y_p, \theta)x : \theta \in [0, \cos^{-1}(x_p^T y_p)]\}$ joins x to y and is minimal in A_{m-1} because it is isomorphic to the minimal geodesic Γ.

The difference between ρ_m and ρ^*_{m-1} is due to the fact that the minimal arc for ρ^*_{m-1} is not a geodesic in S^m. If $r < \pi/2$, then the geodesic segment joining x, y is always shorter than the minimal arc in A_{m-1}. Since the difference is relatively small for close points as seen in Proposition 1(d-e) later, the difference does not obscure much of the underlying structure in the data points in S^m.

3·3. *Geometry of Nested Spheres*

We now define a sequence of *nested spheres* $\{\mathfrak{A}_{d-1}, \ldots, \mathfrak{A}_1\}$ of S^d $(d \ge 2)$ with decreasing intrinsic dimensions. We first introduce a sequence of subspheres $A_{d-1}, \ldots, A_1$ of S^d, which are

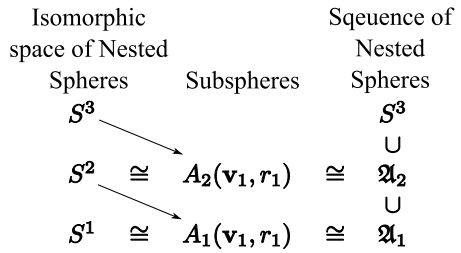

Fig. 5. Hierarchical structure of the sequence of nested spheres of the 3-sphere.

in different spaces. The $(d-1)$-dimensional subsphere A_{d-1} of S^d, defined in Definition 1, is in $S^d \in \mathbb{R}^{d+1}$. The second subsphere A_{d-2} is defined from the isomorphic space S^{d-1} of A_{d-1}. Similarly, the lower dimensional subspheres are defined recursively.

DEFINITION 2. *A sequence $\{A_{d-1}, \ldots, A_1\}$ of subspheres is defined recursively as follows:*

(i) *A_{d-1} is defined as the subsphere with $v_1 \in \mathbb{R}^{d+1}$, $r \in (0, \pi/2]$ by Definition 1.*
(ii) *For each $k = 2, \ldots, d-1$, A_{d-k} is the subsphere defined with $v_k \in \mathbb{R}^{d-k+2}$, $r_k \in (0, \pi/2]$ from S^{d-k+1}, which is isomorphic to A_{d-k+1}.*

A nested sphere is defined by the subsphere located in the original space S^d. For a general subsphere A_{d-k} of S^{d-k+1} $(k = 1, \ldots, d-1)$, we also use the isomorphic 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 $A_{d-k} \subset S^m \subset \mathbb{R}^{m+1}$. The transformations are defined by $v_k \in S^m$ and $r_k \in (0, \pi/2]$ as

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

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 $R(v_k)$.

DEFINITION 3. *A $(d-k)$-dimensional nested sphere $\mathfrak{A}_{d-k}$ of S^d is defined as*

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

The geometric interpretation and hierarchical structure of the nested spheres are illustrated in Fig. 5. The nested sphere $\mathfrak{A}_{d-k}$ can be understood as a shifted $(d-k)$-sphere, which is orthogonal to k orthogonal directions in the sense of Lemma 3. The following properties summarize some geometric facts of the nested spheres.

LEMMA 3. *Let $\mathfrak{A}_{d-1}, \ldots, \mathfrak{A}_1$ be nested spheres of S^d from a sequence of subspheres $A_{d-k}(v_k, r_k)$. Then, there exists an orthogonal basis $v_1^*, \ldots, v_{d-1}^* \in \mathbb{R}^{d+1}$ such that for each $k = 1, \ldots, d-1$,*

(a) *$(x - y)^T v_i^* = 0$ for all $i = 1, \ldots, k$ and $x, y \in \mathfrak{A}_{d-k}$,*

(b) $x \in \mathfrak{A}_{d-k}$ *if and only if* $x_{p,k}^T v_j^* = 0$*, for all* $j = 1, \ldots, k$*, and* $\|x_{p,k}\| = \prod_{i=1}^k \sin(r_i)$ *where*

$$x_{p,k} = x - \cos(r_1)v_1^* - \sin(r_1)\cos(r_2)v_2^* - \cdots - \prod_{i=1}^{k-1} \sin(r_i)\cos(r_k)v_k^*.$$

Moreover, an explicit expression for v_j^* *can be obtained from* $v_1, \ldots, v_j$ *as*

$$v_j^\dagger = f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}(v_j) \in \mathfrak{A}_{d-j+1}, \tag{3}$$

$$v_j^* = \prod_{i=1}^{j-1} \sin^{-1}(r_i) \left\{ v_j^\dagger - \cos(r_1)v_1^* - \cdots - \prod_{i=1}^{j-2} \sin(r_i)\cos(r_{j-1})v_{j-1}^* \right\} \tag{4}$$

Note that $x_{p,k}$ in the lemma can be understood as the projection of x onto the subspace that is orthogonal to $v_1^*, \ldots, v_{d-k}^*$. A direct consequence of Lemma 3 is that for any nested sphere $\mathfrak{A}_{d-k}$ of S^d can be understood as the intersection of a hyperplane $\mathcal{H}_k$ and S^d. The hyperplane $\mathcal{H}_k$ is a $(d-k)$-dimensional affine subspace that is orthogonal to $v_1^*, \ldots, v_{d-k}^*$.

PROPOSITION 2. *Let* $\mathfrak{A}_{d-1}, \ldots, \mathfrak{A}_1$ *be nested spheres of* S^d *from subspheres* $A_{d-k}(v_k, r_k)$*. Then,*

(a) $\mathfrak{A}_1 \subsetneq \mathfrak{A}_2 \subsetneq \cdots \subsetneq \mathfrak{A}_{d-1} \subsetneq S^d$*, where* $A \subsetneq B$ *means that* A *is a proper subset of* B*,*
(b) Let $\rho_{d-k}^*(x, y) = \prod_{i=1}^k \sin(r_i)\rho_{d-k}(x', y')$*, where* $x' = f_k \circ \cdots \circ f_1(x)$*. Then* ρ_{d-k}^* *is a metric on* $\mathfrak{A}_{d-k}$*.*
(c) $(\mathfrak{A}_{d-k}, \rho_{d-k}^*)$ *is isometric to* $(S^{d-k}, \prod_{i=1}^k \sin(r_i)\rho_{d-k})$*.*

The ρ_d and ρ_{d-k}^* are indeed equivalent metrics. Moreover, one can show that $\rho_{d-k}^*(x, y)$ is the length of a minimal arc in $\mathfrak{A}_{d-k}$ that joins x and y.

3·4. *Proofs and Additional Lemmas*

Proof of Lemma 1. (a) Let $a_0 = Ra$, $b_0 = Rb$ and $c_0 = \{b_0 - a_0(a_0^T b_0)\} / \left\|b_0 - a_0(a_0^T b_0)\right\|$. Then $c_0 = Rc$, where $c = \{b - a(a^T b)\} / \left\|b - a(a^T b)\right\|$, since $RR^T = R^T R = I_m$. Then,

$$Q(Rb \to Ra, \theta) = R[I_d + \sin(\theta)(ac^T - ca^T) + \{\cos(\theta) - 1\}(aa^T + cc^T)]R^T = RQ(b \to a, \theta)R^T.$$

(b) Let c' be defined similarly for a', b'. We have

$$a'a'^T = \begin{bmatrix} a \\ 0 \end{bmatrix} \begin{bmatrix} a^T & 0 \end{bmatrix} = \begin{bmatrix} aa^T & 0_{m\times 1} \\ 0_{1\times m} & 0 \end{bmatrix},$$

and $c'c'^T$,$a'c'^T$ and $c'a'^T$ can be expressed in a similar fashion. Then the expression of Q in Lemma 1(b) gives the desired result. □

Proof of Lemma 2. For $x \in \mathbb{R}^m$ such that $\|x\| = 1$, $x \in A_{m-1}$ if and only if $\rho_m(v, x) = \cos^{-1}(v^T x) = r$. This is equivalent to $v^T x - \cos(r)v^T v = 0$ since $v^T v = 1$. This proves (b). Write $x - y = \{x - \cos(r)v\} - \{y - \cos(r)v\}$, then the result (a) follows from (b). □

Proof of Proposition 1. We first show that f is a well-defined bijective function. Proofs for (b-e) will follow. (a) is then given by (c) and (d).

First note that since $\sin(r) > 0$, f is well defined. For any $x \in A_{m-1}$, let $x^\dagger = f(x)$. Then $x^\dagger \in \mathbb{R}^m$, and since $R^-(v)^T R^-(v) = I_{m+1} - vv^T$, we get

$$\|x^\dagger\|^2 = \frac{1}{\sin^2(r)} \left\|R^-(v)x\right\|^2 = \frac{1}{\sin^2(r)}\{x^T x - (x^T v)^2\} = \frac{1}{\sin^2(r)}\{1 - \cos^2(r)\} = 1$$

Thus, $x^\dagger \in S^{m-1}$. Conversely, for any $x^\dagger \in S^{m-1}$, let $x = f^{-1}(x^\dagger)$. Then $\|x\| = 1$ and

$$v^T x = \{R(v)v\}^T \begin{bmatrix} \sin(r)x^\dagger \\ \cos(r) \end{bmatrix} = \cos(r).$$

By Lemma 2(b), $x \in A_{m-1}$. One can easily show that $f \circ f^{-1}(x^\dagger) = x^\dagger,\ f^{-1} \circ f(x) = x$. Therefore, f is a well defined bijective function.

Since ρ_{m-1} is a metric and $\sin(r) > 0$, the metric ρ^*_{m-1} is nonnegative and symmetric, and the triangle inequality holds. In addition, since f is bijective, we have $\rho^*_{m-1}(x, y) = 0$ if and only if $x = y$. This proves (b). With the metric ρ^*_{m-1}, f is an isometry and (c) follows.

To prove (d) and (e), the difference between two metrics for a fixed $r \in (0, \pi/2]$ is given by

$$\rho^*_{m-1}(x, y) - \rho_m(x, y) = \sin(r)\cos^{-1}\left[\frac{\cos\{\rho_m(x, y)\} - \cos^2(r)}{\sin^2(r)}\right] - \rho_m(x, y) := h_r\{\rho_m(x, y)\},$$

for $x, y \in A_{m-1}$. Note that $\max_{x,y} \rho_m(x, y) = 2r$. Then h_r is a strictly increasing function on $[0, 2r]$ with minimum $h_r(0) = 0$ and the maximum $h_r(2r) = \pi \sin(r) - 2r$. This proves (e) and leads to the first inequality of (d). The second inequality is obtained from observing that

$$\frac{\pi \sin(r)}{2r}\rho_m(x, y) - \rho^*_{m-1}(x, y)$$

is nonnegative and is zero if and only if $\rho_m(x, y) = 0$ or $2r$. □

The following lemmas are useful to prove Lemma 3 and also could be of independent interest.

LEMMA 4. *Let $v_j^\dagger$ and v_j^* be as defined in (3-4). For any $x \in S^d$ and $k = 1, \ldots, d-1$, the following are equivalent:*

(i) $x \in \mathfrak{A}_{d-k}$.

(ii) $v_k^T\{f_k \circ \cdots \circ f_1(x)\} = \cos(r_k)$.

(iii) For all $j = 1, \ldots, k$,

$$x^T v_j^\dagger = \prod_{i=1}^{j-1} \sin^2(r_i)\cos(r_j) + \prod_{i=1}^{j-2} \sin^2(r_i)\cos^2(r_{j-1}) + \cdots + \cos^2(r_1).$$

(iv) For all $j = 1, \ldots, k$,

$$x^T v_j^* = \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j).$$

Proof of Lemma 4. [*(i)* ⇔ *(ii)*] By Definition 3 and since each f_i is bijective, $x \in \mathfrak{A}_{d-k}$ is equivalent to $f_k \circ \cdots \circ f_1(x) \in A_{d-k}$. By Lemma 2(b), this is also equivalent to (ii).

[*(i)* ⇔ *(iii)*] First note that for any $k = 1, \ldots, d-1$, for $y \in S^{d-k}$,

$$f_1^{-1} \circ \cdots \circ f_k^{-1}(y) = R^T(v_1) \begin{bmatrix} \sin(r_1)\{f_2^{-1} \circ \cdots \circ f_k^{-1}(y)\} \\ \cos(r_1) \end{bmatrix}$$

$$= \{R(v_1, \ldots, v_k)\}^T \begin{bmatrix} \prod_{i=1}^{k} \sin(r_i)y \\ \prod_{i=1}^{k-1} \sin(r_i)\cos(r_{k-1}) \\ \vdots \\ \cos(r_1) \end{bmatrix},$$

where $R(v_1, \ldots, v_k)$ is a rotation matrix defined as

$$\{R(v_1, \ldots, v_k)\}^T = R^T(v_1) \begin{bmatrix} R^T(v_2) & 0_{d\times 1} \\ 0_{1\times d} & 1 \end{bmatrix} \cdots \begin{bmatrix} R^T(v_k) & 0_{(d+2-k)\times 1} \\ 0_{1\times(d+2-k)} & 1 \end{bmatrix}.$$

Then

$$\begin{aligned} x^T v_j^\dagger &= \left[f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}\{f_{j-1} \circ \cdots \circ f_1(x)\}\right]^T f_1^{-1} \circ \cdots \circ f_{j-1}^{-1}(v_j) \\ &= \prod_{i=1}^{j-1} \sin^2(r_i)\{f_{j-1} \circ \cdots \circ f_1(x)\}^T v_j + \prod_{i=1}^{j-2} \sin^2(r_i)\cos^2(r_{j-1}) + \cdots + \cos^2(r_1) \end{aligned}$$

and the result follows from (ii).

[*(i) ⇒ (iv)*] Since $x \in \mathfrak{A}_{d-k}$, we have $x^T v_1^* = \cos(r_1)$ by definition. Suppose $x^T v_\jmath^* = \prod_{i=1}^{\jmath-1} \sin(r_i)\cos(r_\jmath)$ for all $\jmath = 1, \ldots, j-1$, then (iii) and cancelling terms give

$$\begin{aligned} x^T v_j^* &= x^T \left\{ v_j^\dagger - \cos(r_1)v_1^* - \cdots - \prod_{i=1}^{j-2} \sin(r_1)\cos(r_{j-1})v_{j-1}^* \right\} \prod_{i=1}^{j-1} \sin^{-1}(r_i) \\ &= \prod_{i=1}^{j-1} \sin(r_1)\cos(r_j). \end{aligned}$$

Thus by induction, (iv) holds.

[*(iv) ⇒ (iii)*] Suppose (iv) holds, then for $j = 1, \ldots, k$,

$$\begin{aligned} &x^T v_j^* - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) \\ &= \prod_{i=1}^{j-1} \sin^{-1}(r_i) \left\{ x^T v_j^\dagger - \cos^2(r_1) - \cdots - \prod_{i=1}^{j-2} \sin^2(r_i)\cos^2(r_{j-1}) - \prod_{i=1}^{j-1} \sin^2(r_i)\cos(r_j) \right\}, \end{aligned}$$

which equals to zero if and only if (iii) holds. □

Proof of Lemma 3. We first show that $\{v_i^*; i = 1, \ldots, d-1\}$ is an orthonormal basis. Note that $v_1^* = v_1$, and $v_2^* = \sin^{-1}(r_1)\{v_2^\dagger - \cos(r_1)v_1^*\}$. Since $v_2^\dagger \in \mathfrak{A}_{d-1}$, by Lemma 4, we have

$$v_2^{*T} v_1 = \sin^{-1}(r_1)\{v_2^{\dagger T} v_1 - \cos(r_1)\} = 0,$$

and

$$v_2^{*T} v_2^* = \sin^{-1}(r_1) v_2^{*T} v_2^\dagger = \sin^{-1}(r_1)\{v_2^{\dagger T} v_2^\dagger - \cos(r_1) v_2^\dagger v_1^*\} = 1.$$

Suppose $v_i^{*T} v_j^* = 0$ and $\|v_i^*\| = \|v_i^*\| = 1$ for $1 \le i < j \le k-1$. Since $v_k^\dagger \in \mathfrak{A}_{d-k+1}$, by Lemma 4, we have

$$v_j^{*T} v_k^* = \prod_{i=1}^{k-1} \sin^{-1}(r_i) v_j^{*T} \left\{ v_k^\dagger - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) v_j^* \right\} = 0,$$

and

$$\|v_k^*\| = \prod_{i=1}^{k-1} \sin^{-1}(r_i) v_k^{*T} v_k^\dagger$$
$$= \prod_{i=1}^{k-1} \sin^{-2}(r_i) v_k^{\dagger T} \left\{ v_k^\dagger - \cos(r_1)v_1^* - \cdots - \prod_{i=1}^{k-2} \sin(r_i)\cos(r_{k-1}) v_{k-1}^* \right\} = 1.$$

Thus, by induction, v_i^* $(i = 1, \ldots, d-1)$ are orthonormal.

Now for (b), suppose first that $x \in \mathfrak{A}_{d-k}$. Then by Lemma 4, we get for $j = 1, \ldots, k$

$$x_{p,k}^T v_j^* = x^T v_j^\dagger - \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) = 0$$

and

$$\|x_{p,k}\|^2 = x^T x_{p,k} = x^T \{x - \cos(r_1)v_1^* - \cdots - \prod_{i=1}^{k-1} \sin(r_i)\cos(r_k) v_k^*\}$$
$$= 1 - \cos^2(r_1) - \cdots - \prod_{i=1}^{k-1} \sin^2(r_i)\cos^2(r_k).$$

Thus by rearranging terms, $\|x_{p,k}\| = \prod_{i=1}^k \sin(r_i)$.

Conversely, suppose that $x_{p,k}^T v_j^* = 0$ for all $j = 1, \ldots, k$ and $\|x_{p,k}\| = \prod_{i=1}^k \sin(r_i)$. Then since $x_{p,k}, v_1^*, \ldots, v_k^*$ are orthogonal to each other,

$$\|x\|^2 = \left\| x_{p,k} + \cos(r_1)v_1^* + \cdots + \prod_{i=1}^{k-1} \sin(r_i)\cos(r_k) v_k^* \right\|^2$$
$$= x_{p,k}^T x_{p,k} + \cos^2(r_1) + \cdots + \prod_{i=1}^{k-1} \sin^2(r_i)\cos^2(r_k) = 1.$$

One can check that for all $j = 1, \ldots, k$

$$x^T v_j^* = \left\{ x_{p,k} + \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j) v_j^* \right\}^T v_j^* = \prod_{i=1}^{j-1} \sin(r_i)\cos(r_j),$$

and again by Lemma 4, the result follows. (a) is directly obtained from (b). □

Proof of Proposition 2. (a) is readily derived by either Lemma 3 or the fact that $A_{m-1} \subsetneq S^m$ for all $m = 2, \ldots, d$.

For (b) and (c), it can be easily checked that $f_k \circ \cdots \circ f_1 : \mathfrak{A}_{d-k} \to S^{d-k}$ is a well defined bijective function. Since ρ_{d-k} is a metric and $\sin(r_i) > 0$, the metric ρ_{d-k} is nonnegative and symmetric, and the triangle inequality holds. In addition, since f is a bijection, $\rho_{d-k}^*(x, y) = 0$ if and only if $x = y$. This proves (b). Then by the definition of ρ_{d-k}^*, $f_k \circ \cdots \circ f_1$ is an isometry and (c) follows. □

Proof of Theorem 1 (Main article). Note that $w^T w^* = w^T M w = 0$, and for all $z \in S^d$ such that $w^T z \geq 0$ and $w^T M z = 0$, $z^T w^* = z^T M w = w^T M z = 0$. Thus $\rho_d(w, w^*) = \cos^{-1}(w^T w^*) = \pi/2$ and $\rho_d(z, w^*) = \pi/2$. Moreover, since $w^T z \geq 0$, we have $w, z \in hA_{d-1}$. □

4. TRANSFORMATIONS BETWEEN SPHERICAL VARIABLE AND THE SPACE OF PRINCIPAL SCORES

The material in this section supports Section 2.4 and 2.5 of the main article.

Suppose we have a sequence of nested spheres defined by the sequence of subspheres

$$\{A_{d-1}(v_1, r_1), \ldots, A_1(v_{d-1}, r_{d-1}), v_d\},$$

where $v_k \in S^{d-k+1}$ $(k = 1, \ldots, d-1)$ and $r_k \in (0, \pi/2]$.

A point $x \in S^d \subset \mathbb{R}^{d+1}$ is transformed into the subset E of a vector space that are coordinates of the principal scores matrix. A detailed description of the transformation is as follows.

The point x of interest is projected to the first subsphere $A_{d-1}(v_1, r_1)$, where the projection is

$$x^P = P\{x; A_{d-1}(v_1, r_1)\} = \frac{x \sin(r_1) + v_1 \sin(\xi_1)}{\sin(\xi_1 + r)}, \tag{5}$$

where ξ_1 is the signed deviance measured by the geodesic distance from x to $A_{d-1}(v_1, r_1)$ and is given by $\xi_1 = \arccos(v_1^T x) - r_1$. Define a transformation function $g_1 : S^d \to S^{d-1}$,

$$g_1(x) = \{1 - (v_1^T x)^2\}^{-\frac{1}{2}} R^-(v_1)x,$$

where $R(v)$ is the $(d+1) \times (d+1)$ rotation matrix that moves v to the north pole $e_{d+1} = (0, \ldots, 0, 1)^T$ and $R^-(v)$ is the $d \times (d+1)$ matrix consisting of the first d rows of $R(v)$. It is easy to see that $g_1(x) = f_1(x^P)$, for f_1 defined in (2), which means that g_1 is the transformed version of the projection.

The second signed deviance is obtained with $g_1(x) \in S^{d-1}$ in a recursive fashion. Define $g_k : S^{d-k+1} \to S^{d-k}$ for each $k = 1, \ldots, d$ as $g_k(x) = \{1 - (v_1^T x)^2\}^{-\frac{1}{2}} R^-(v_k)x$. Let $g_1^k : S^d \to S^{d-k}$ be $g_1^k(x) = g_k \circ \cdots \circ g_1(x)$. Then $\xi_k = \arccos\{v_k^T g_1^{k-1}(x)\} - r_k$ $(k = 1, \ldots, d-1)$. The last signed deviance is the signed deviation of $g_1^{d-1}(x) \in S^1$ from v_d and is

$$\xi_d = \text{atan2}\{(v_d^\perp)^T g_1^{d-1}(x), v_d^T g_1^{d-1}(x)\},$$

where

$$v_d^\perp = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} v_d$$

is the orthogonal axis of v_d and $\text{atan2}(y, x) = 2\arctan\{(\sqrt{x^2 + y^2} - x)/y\}$ if $y \neq 0$ and is 0 if $y = 0$. These residuals are scaled by multiplying by $\prod_{i=1}^{k-1} \sin(r_i)$ which makes the magnitude of the residuals commensurate. We then have

$$z = \begin{pmatrix} \xi_d \prod_{i=1}^{d-1} \sin(r_i) \\ \xi_{d-1} \prod_{i=1}^{d-2} \sin(r_i) \\ \vdots \\ \xi_2 \sin(r_1) \\ \xi_1 \end{pmatrix} \in E,$$

where $E = [-\pi, \pi) \times [-\pi/2, \pi/2)^{d-1}$. Let $h : S^d \to E \subset \mathbb{R}^d$ be $h : x \mapsto z$.

The inverse operation is recovering the original S^d from the deviances ξ_i. We begin with recovering S^1, then S^2, and recursively up to S^d. For $z = (z_1, \ldots, z_d)^T \in E$, let

$$\xi_j = z_{d+1-j} \left\{ \prod_{i=1}^{j-1} \sin(r_i) \right\}^{-1} \qquad (j = 1, \ldots, d) \tag{6}$$

Recall that ξ_j is the signed deviation from the jth nested sphere with dimension $d-j$ and a zero-dimensional sphere is understood as a point. Also note that a sequence of nested spheres defined by $\{(v_1, r_1), \ldots, (v_{d-1}, r_{r-1}), v_d\}$ is still used to recover S^d.

We begin with ξ_d and v_d to recover S^1 as follows.

$$\vec{\xi}_d = \cos(\xi_d) v_d + \sin(\xi_d) v_d^{\perp} \in S^1 \subset \mathbb{R}^2$$

The second step is to recover S^2 from $\vec{\xi}_d \in S^1$ and $\xi_{d-1} \in [-\pi/2, \pi/2]$. In general, all subsequent steps recursively recover S^k from some $\xi \in S^{k-1}$ and $\xi_k \in [-\pi/2, \pi/2]$. It is useful to define a transformation function $\tilde{g}_k : S^{d-k} \times [-\pi/2, \pi/2] \to S^{d-k+1}$ $(k = 1, \ldots, d-1)$ as

$$\tilde{g}_k(\xi, \xi_k) = R^T(v_k) \begin{bmatrix} \sin(r_k + \xi_k)\xi \\ \cos(r_k + \xi_k) \end{bmatrix} \in S^{d-k+1}$$

Also define that $\tilde{g}_{d-1}^k = \tilde{g}_k \circ \cdots \circ \tilde{g}_{d-1}$, aggregating recoveries from S^1 to S^{d-k+1} for some $k = 1, \ldots d-1$. The composition is understood as in the following example.

$$\begin{aligned} \tilde{g}_{d-1}^{d-2}(\{\xi_1, \ldots, \xi_d\}) &= \tilde{g}_{d-2} \circ \tilde{g}_{d-1}(\{\xi_1, \ldots, \xi_d\}) \\ &= \tilde{g}_{d-2}\left\{\tilde{g}_{d-1}(\vec{\xi}_d, \xi_{d-1}), \xi_{d-2}\right\}. \end{aligned}$$

For a given $z \in E$, a corresponding point x in S^d is found by

$$x = \tilde{g}_{d-1}^1(\{\xi_1, \ldots, \xi_d\})$$

with the deviations $\{\xi_1, \ldots, \xi_d\}$ obtained by (6). Let $\tilde{h} : E \to S^d$ be $\tilde{h} : z \to x$.

Finally we check that $\tilde{h}$ is the inverse function h. This ensures that h can be used to transform a spherical variable on S^d into the vector space E of principal scores and that $\tilde{h}$ is the inverse operation.

5. CONVERGENCE OF THE PROPOSED ALGORITHM

In this section, we show the convergence of the proposed solution to a local minimum empirically by showing the values of objective function at each iteration. Recall that the algorithm to minimize the objective function

$$F(v, r) = \sum_{i=1}^{n} \{\rho_d(x_i, v) - r\}^2 \tag{7}$$

consists of two layers: the inner and outer layers. The outer layer finds a tangential point for the inner layer. The inner layer fits the least squares sphere to the data projected onto the tangent space, which is solved by the well-known Levenberg–Marquardt algorithm (Scales, 1985, Ch. 4). In the outer iteration, the convergence to a local minimum is empirically shown by checking values of the objective function (7) at each iteration, until the difference becomes small enough to be ignored.

We have used two different initial cadidates for v. The first choice of the initial values is the last singular vector from the singular value decomposition of the original data matrix. Denote this choice of the initial values as c_1. The second choice (c_2) is the last singular vector from the singular value decomposition of the centreed data matrix.

We examined the convergence with four real data sets.

1. Elephant Seal data

The elephant seal data set consists of $n = 73$ observations on S^2 (Brillinger & Stewart, 1998; Rivest, 1999). The subsphere fitting occurs only once. Figure 6 shows the values of objective functions in each iteration with two different initial values c_1 and c_2. With both choices of initial values, the algorithm converges fast to the same v, and also r, in 3 or 6 iterations respectively.

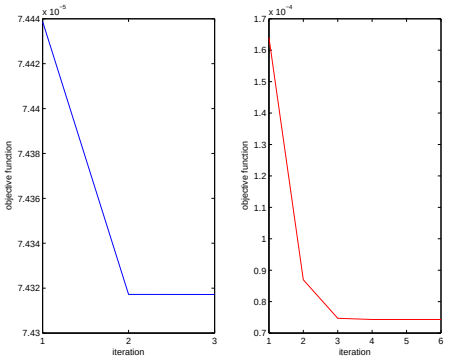

Fig. 6. Elephant seal data: Values of the objective functions with the initial value c_1 on the left, c_2 on the right panel. The values of the minimized objective function with different choices of initial values are the same as $7.4317e-005$.

2. Rat Skull Growth data

The rat skull growth data is a landmark shape data set with $k = 8$ landmarks. We process this data set as described in Section 5.2, then the algorithm is applied 11 times since the dimension of the shape space is 12. Figures 7 and 8 show the values of the objective functions.

The top left panel of Fig. 7 illustrates the evaluated objective function values with c_1 as the initial value when fitting A_{11} from points $x_1, \ldots, x_n \in S^{12}$. The algorithm converges to a local minimum at the 6th iteration. Then we move on to the fitting of A_{10} from S^{11}, shown in the second panel on the top row. Again, the algorithm converges to a local minimum in less than 20 iterations. We examine the convergence until finally fitting A_1 from S^2, shown at the rightmost panel at the bottom row. Throughout 11 small sphere fittings, the algorithm requires less than 20 iterations.

Figure 8 shows the convergence results with a different choice of the initial value as c_2. In the rat skull data application, different choices of initial values led to very similar results, which can be examined, for example, by checking that the minimized value of the objective function at the bottom left panel is about 8.75×10^{-4} at Fig. 8, which is the same at Fig. 7.

3. Digit 3 data

The digit 3 data set contained in Dryden & Mardia (1998) is also analyzed. The data set consists of $k = 13$ landmarks on digitized hand written figure 3s. When aligned appropriately, the preshapes are on S^{22}. We show the first 15 layers of the principal nested spheres fitting results.

The top left panel of Fig. 9 shows the values of objective functions with c_1 as the initial value when fitting A_{21} from points in S^{22}. The next, to the right, panel shows the values at the second fitting, and so on. In all the 21 sphere fittings, where the last six are not shown in the figures, the algorithm converges to a local minimum mostly within 20 iterations.

When choosing c_2 as the initial value, the convergence to a local minimum mostly within 20 iterations is also observed. See Fig. 10.

4. Human Movement data

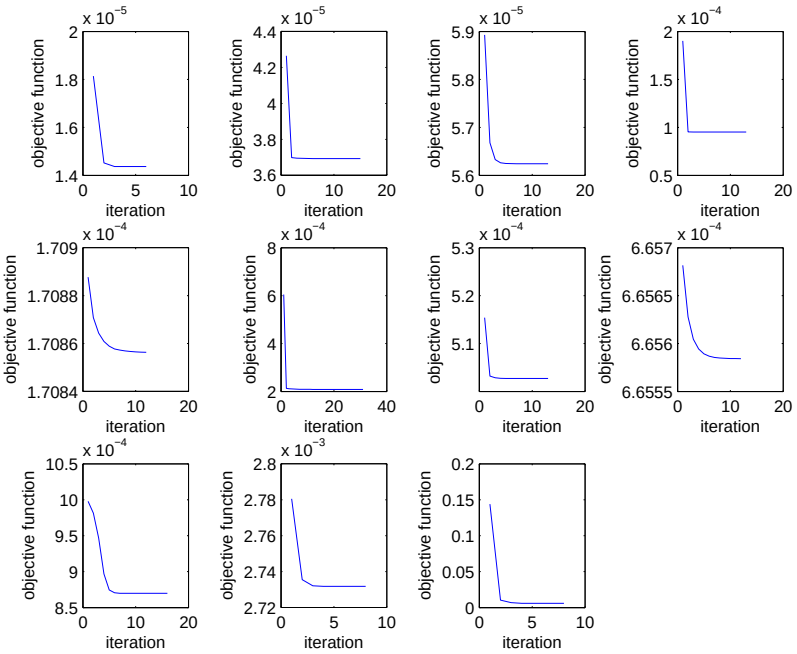

Fig. 7. Rat Skull data: values of the objective functions with the initial value c_1.

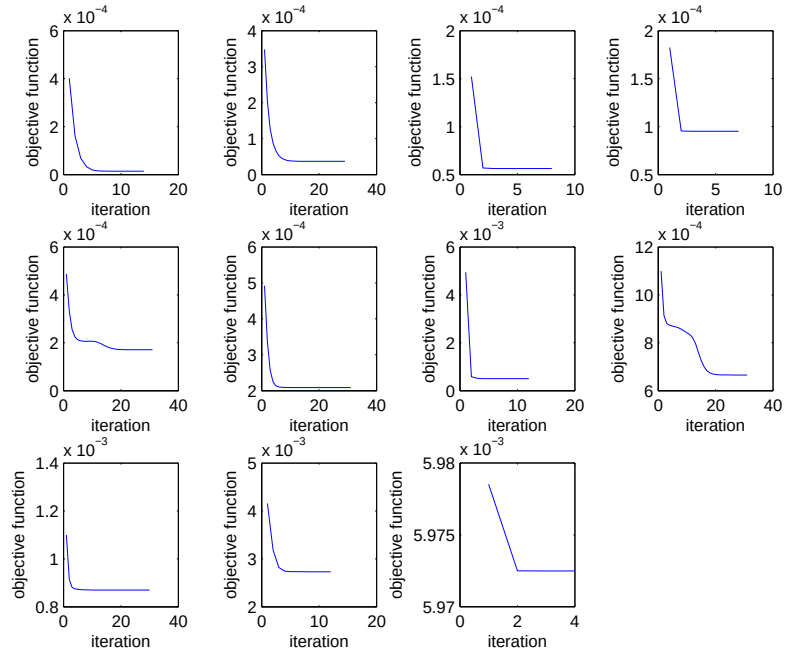

Fig. 8. Rat Skull data: values of the objective functions with the initial value c_2.

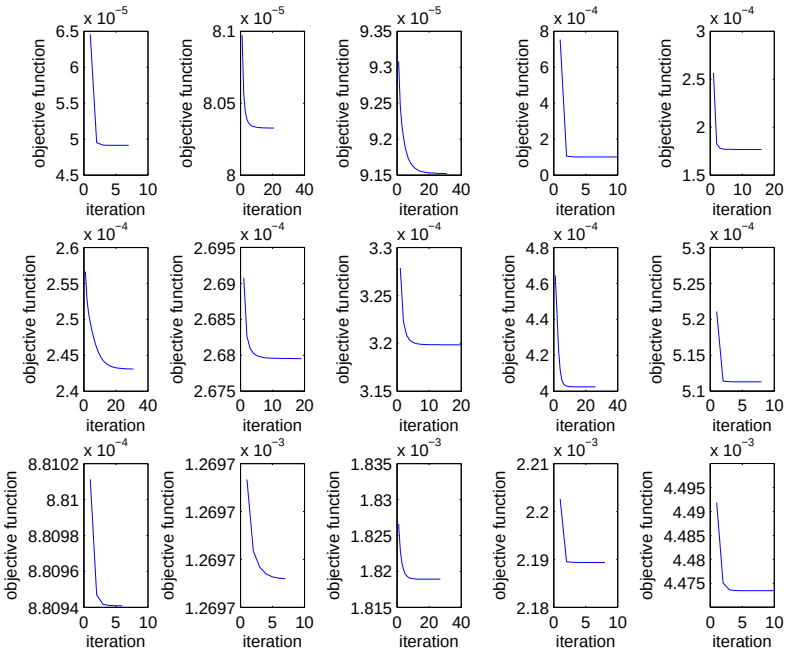

Fig. 9. Digit 3 data: values of the objective functions with the initial value c_1.

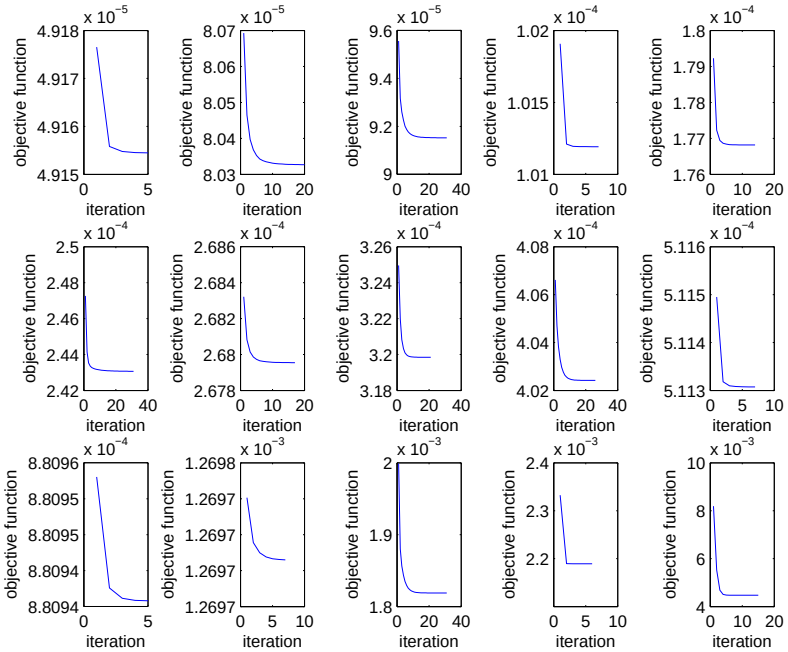

Fig. 10. Digit 3 data: values of the objective functions with the initial value c_2.

Table 4. *Human movement data: Minimized values of the objective function F.*

Initial value	$S^4 \to A_3$	$S^3 \to A_2$	$S^2 \to A_1$
c_1	1.3075×10^{-5}	1.4913×10^{-4}	0.0660
c_2	1.3075×10^{-5}	6.7660×10^{-5}	0.0016

The human movement data, analyzed in greater detail in Section 5.1, are on S^4, when aligned. The proposed algorithm does not guarantee a convergence to the global minimum. In an analysis of this data set, we have seen that different choices of initial values led to a convergence to a local, not global, minimum. While it can be checked that the algorithm finds a local minimum as in Fig. 11, the algorithm does not guarantee a convergence to the global minimum. In Table 4, we see that when c_1 is used for the initial value of the algorithm, the solution is indeed a local minimum, since the minimized values of the objective function are larger than those from choosing c_2.

In practice, we examine the optimization results from both initial values, and take the one with smaller value of the objective function.

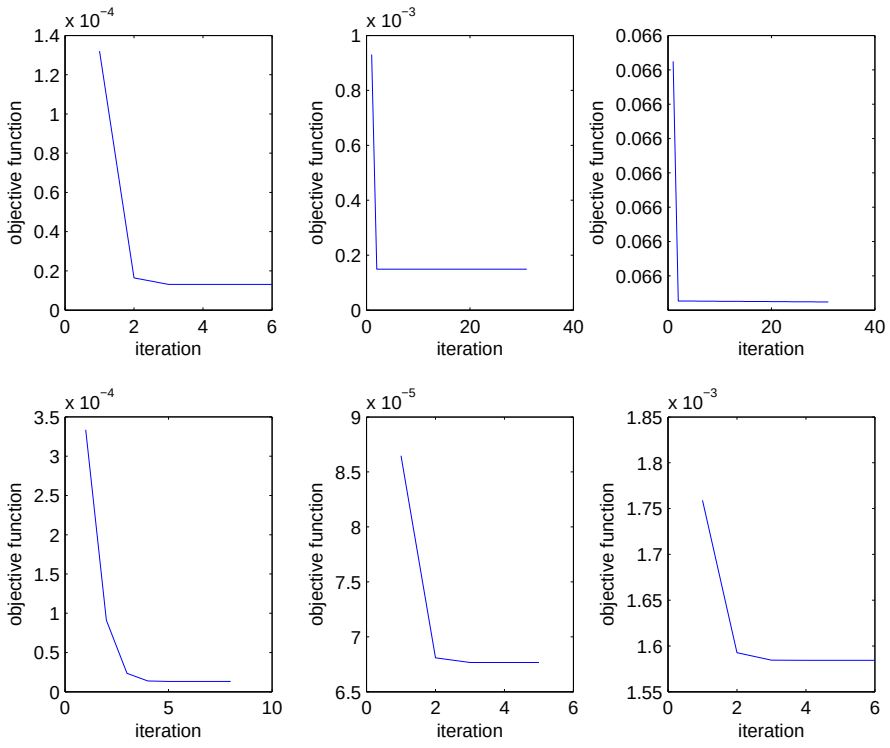

Fig. 11. Human data: values of the objective functions with the initial value c_1 in the top row and c_2 in the bottom row.

6. PRINCIPAL NESTED GREAT SPHERE MEAN AND MEAN OF HUCKEMANN & ZIEZOLD (2006)

We show that the principal nested great spheres mean, defined in Definition 3, and the mean of Huckemann & Ziezold (2006) do not coincide even on S^2, and also show an example where the two are the same.

Suppose $x_1, \ldots, x_n \in S^2$. Geodesic principal component analysis of Huckemann & Ziezold (2006) first finds a best fitting geodesic $\gamma_1 = \gamma_1(a, b)$, where $\gamma_1(a, b)$ ($a, b \in S^2$ such that $a^T b = 0$) can be parametrized as $\gamma_1(t) = \cos(t)a + \sin(t)b$ for $t \in \mathbb{R}$. The great circle γ_1 is found by minimizing the sum of squared deviances to the data and thus coincides with the fitted subsphere $\widehat{A}_1(v, \pi/2)$ of S^2 for $v = a \times b$. The great circle γ_1 is called the first geodesic principal component. The second geodesic principal component γ_2 minimizes

$$\sum_{i=1}^{n} \rho_2^2\{x_i, \gamma_2(v, c)\}, \qquad (c \in \gamma_1)$$

where $\rho_2\{x, \gamma_2(v, c)\}$ is the minimal great circle distance between x and γ_2. Huckemann & Ziezold (2006) defined a mean as one of the intersections between γ_1 and γ_2. Denote the mean by γ_0.

Clearly, the first geodesic principal component γ_1 is identical to the one dimensional principal nested great sphere $\mathfrak{A}_1$ for the special case of S^2.

However, γ_0 may not be the same as the principal nested great spheres mean $\mathfrak{A}_0$, because the objective function of γ_0 given γ_1 is

$$\sum_{i=1}^{n} \rho_2^2\{x_i, \gamma_2(v, c)\} = \sum_{i=1}^{n} \arccos^2[\{(v^T x_i)^2 + (\gamma_0^T x_i)^2\}^2], \tag{8}$$

which is different from the objective function of principal nested spheres mean, written in terms of γ_0

$$\begin{aligned} &\sum_{i=1}^{n} \rho_2^2[P\{x_i; A_1(v, \pi/2)\}, \gamma_0] \\ &= \sum_{i=1}^{n} \rho_2^2[\{x_i - (x_i^T v)v\}/\sin\{\rho_2(x_i, v)\}, \gamma_0] \\ &= \sum_{i=1}^{n} \arccos^2[\gamma_0^T x_i/\{1 - (x_i^T v)^2\}]. \end{aligned} \tag{9}$$

A case where γ_0 and $\mathfrak{A}_0$ are the same is when (8) and (9) are identical. A sufficient condition for the identity is that $v_i^T x_1 = 0$ for all i, which is the case where all data points are exactly contained in $\widehat{\mathfrak{A}}_1 = \gamma_1$.

REFERENCES

BANERJEE, A., DHILLON, I. S., GHOSH, J. & SRA, S. (2005). Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research* **6**, 1345–1382.

BAUER, P., RÖHMEL, J., MAURER, W. & HOTHORN, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

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

BRILLINGER, D. R. & STEWART, B. S. (1998). Elephant-seal movements: Modelling migration. *Canad. J. Statist.* **26**, 431–443.

DMITRIENKO, A., TAMHANE, A. C., LIU, L. & WIENS, B. L. (2008). A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine* **27**, 3446–3451.

DMITRIENKO, A., WIENS, B. L., TAMHANE, A. C. & WANG, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* **26**, 3465–2478.

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

FISHER, R. (1953). Dispersion on a sphere. *Proc. Roy. Soc. London. Ser. A.* **217**, 295–305.

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

KENT, J. T., DRYDEN, I. L. & ANDERSON, C. R. (2000). Using circulant symmetry to model featureless objects. *Biometrika* **87**, 527–544.

KUME, A., DRYDEN, I. L. & LE, H. (2007). Shape-space smoothing splines for planar landmark data. *Biometrika* **94**, 513–528.

MARDIA, K. V. & JUPP, P. E. (2000). *Directional Statistics*. Wiley.

MARRON, J. S., TODD, M. J. & AHN, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102**, 1267–1271.

PUTERMAN, M. L. (1994). *Markov Decision Processes*. Wiley.

RIVEST, L.-P. (1999). Some linear model techniques for analyzing small-circle spherical data. *Canad. J. Statist.* **27**, 623–638.

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

VAPNIK, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

WESTFALL, P. H. & KRISHEN, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

WICHERS, L., LEE, C., COSTA, D., WATKINSON, P. & MARRON, J. S. (2007). A functional data analysis approach for evaluating temporal physiologic responses to particulate matter. Tech. Rep. 5, University of North Carolina at Chapel Hill, Department of Statistics and Operations Research.