Chapter 10
Statistical Methods Generalizing Principal Component Analysis to Non-Euclidean Spaces

Stephan Huckemann and Benjamin Eltzner

Contents

10.1 Introduction 318
10.2 Some Euclidean Statistics Building on Mean and Covariance 318
10.3 Fréchet ρ-Means and Their Strong Laws 320
10.4 Procrustes Analysis Viewed Through Fréchet Means 323
10.5 A CLT for Fréchet ρ-Means 325
10.6 Geodesic Principal Component Analysis 328
10.7 Backward Nested Descriptors Analysis (BNDA) 330
10.8 Two Bootstrap Two-Sample Tests 332
10.9 Examples of BNDA 334
10.10 Outlook 336
References 336

Abstract Very generally speaking, statistical data analysis builds on descriptors reflecting data distributions. In a linear context, well studied nonparametric descriptors are means and PCs (principal components, the eigenorientations of covariance matrices). In 1963, T.W. Anderson derived his celebrated result of joint asymptotic normality of PCs under very general conditions. As means and PCs can also be defined geometrically, there have been various generalizations of PC analysis (PCA) proposed for manifolds and manifold stratified spaces. These generalizations play an increasingly important role in statistical dimension reduction of non-Euclidean data. We review their beginnings from Procrustes analysis (GPA), over principal geodesic analysis (PGA) and geodesic PCA (GPCA) to principal nested spheres (PNS), horizontal PCA, barycentric subspace analysis (BSA) and backward nested descriptors analysis (BNDA). Along with this, we review the current state of the art of their asymptotic statistical theory and applications for statistical testing, including open challenges, e.g. new insights into scenarios of nonstandard rates and asymptotic nonnormality.

S. Huckemann (✉) · B. Eltzner
Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Göttingen, Germany
e-mail: huckeman@math.uni-goettingen.de; beltzne@math.uni-goettingen.de

P. Grohs et al. (eds.), *Handbook of Variational Methods for Nonlinear Geometric Data*, https://doi.org/10.1007/978-3-030-31351-7_10

10.1 Introduction

The *mean* and the *covariance* are among the most elementary statistical descriptors describing a distribution in a nonparametric way, i.e. in the absence of a distributional model. They can be used for dimension reduction and for statistical testing based on their asymptotics. Extending these two quantities to non-Euclidean random deviates and designing statistical methods for these has been the subject of intense research in the last 50 years, beginning with *Procrustes analysis* introduced by Gower [23] and the *strong law of large numbers* for Fréchet means by Ziezold [51]. This chapter intends to provide a brief review of the development of this research until now and to put it into context.

We begin with the Euclidean version including classical PCA, introduce the more general concept of generalized Fréchet ρ-means, their strong laws and recover *general Procrustes analysis* (GPA) as a special case. Continuing with *principal geodesic analysis* we derive a rather general central limit theorem for generalized Fréchet ρ-means and illustrate how to recover from this Anderson's asymptotic theorem for the classical first PC and the CLT for Procrustes means. Next, as another application of our CLT we introduce *geodesic principal component analysis* (GPCA), which, upon closer inspection, turns out to be a nested descriptor. The corresponding *backward nested descriptor analysis* (BNDA) requires a far more complicated CLT, which we state. We put the rather recently developed methods of *principal nested spheres* (PNS), *horizontal PCA* and *barycentric subspace analysis* (BSA) into context and conclude with a list of open problems in the field.

10.2 Some Euclidean Statistics Building on Mean and Covariance

Asymptotics and the Two-Sample Test

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ be random vectors in $\mathbb{R}^D$, $D \in \mathbb{N}$, with existing *population mean* $\mathbb{E}[X]$. Denoting the *sample mean* by

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j \,,$$

the *strong law of large numbers* (SLLN) asserts that (e.g. [8, Chapter 22])

$$\bar{X}_n \overset{\text{a.s.}}{\to} \mathbb{E}[X] \,.$$

Upon existence of the second moment $\mathbb{E}[\|X\|^2]$, the *covariance* $\operatorname{cov}[X]$ exists and the *central limit theorem* (CLT) asserts that the fluctuation between sample and population mean is asymptotically normal (e.g. [14, Section 9.5]), namely that

$$\sqrt{n}\big(\bar{X}_n - \mathbb{E}[X]\big) \overset{\mathcal{D}}{\to} \mathcal{N}\big(0, \operatorname{cov}[X]\big)\,. \tag{10.1}$$

Using the *sample covariance*

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \mathbb{E}[X])(X_j - \mathbb{E}[X])^T$$

as a *plugin estimate* for $\operatorname{cov}[X]$ in (10.1), asymptotic confidence bands for $\mathbb{E}[X]$ can be obtained as well as corresponding tests.

A particularly useful test is the *two-sample* test, namely that for random vectors $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ in $\mathbb{R}^D$ and independent random vectors $Y_1, \dots, Y_m \overset{i.i.d.}{\sim} Y$ in $\mathbb{R}^D$ with full rank population and sample covariance matrices, $\operatorname{cov}[X]$ and $\operatorname{cov}[Y]$, $\hat{\Sigma}_n^X$ and $\hat{\Sigma}_m^Y$, respectively,

$$T^2 = \frac{n+m-2}{\frac{1}{n}+\frac{1}{m}}(\bar{X}_n - \bar{Y}_m)^T\left((n-1)\hat{\Sigma}_n^X + (m-1)\hat{\Sigma}_m^Y\right)^{-1}(\bar{X}_n - \bar{Y}_m) \tag{10.2}$$

follows a *Hotelling* distribution if X and Y are multivariate normal, cf. [40, Section 3.6.1]. More precisely, $T^2\,\frac{nm(n+m-D-1)}{(n+m)(n+m-2)D}$ follows a $F_{D,n+m-D-1}$-distribution. Remarkably, this holds also asymptotically under nonnormality of X and Y, if $\operatorname{cov}[X] = \operatorname{cov}[Y]$ or $n/m \to 1$, cf. [45, Section 11.3].

Principal Component Analysis (PCA)

Consider again random vectors $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ in $\mathbb{R}^D$, $D \in \mathbb{N}$, with sample covariance matrix $\hat{\Sigma}$ and existing population covariance $\Sigma = \operatorname{cov}[X]$. Further let $\Sigma = \Gamma\Lambda\Gamma^T$ and $\hat{\Sigma} = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^T$ be spectral decompositions, i.e. $\Gamma = (\gamma_1, \dots, \gamma_D), \hat{\Gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_D) \in SO(D)$ and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_D)$, $\hat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_D)$ with $\lambda_1 \geq \dots \lambda_D \geq 0$ and $\hat{\lambda}_1 \geq \dots \hat{\lambda}_D \geq 0$, respectively. Then the vectors γ_j $(j = 1, \dots, D)$ are called *population principal components* and $\hat{\gamma}_j$ $(j = 1, \dots, D)$ are called *sample principal components*, abbreviated as PCs. These PCs can be used for *dimension reduction*, namely considering instead of $X_1, \dots, X_n \in \mathbb{R}^D$ their projections, also called *scores*,

$$\left(X_1^T\hat{\gamma}_j\right)_{j=1}^{J}, \dots, \left(X_n^T\hat{\gamma}_j\right)_{j=1}^{J} \in \mathbb{R}^J\,,$$

to the first $1 \leq J \leq D$ PCs. The *variance explained* by the first J PCs is

$$\hat{\lambda}_1 + \dots + \hat{\lambda}_J\,.$$

Due to the seminal result by Anderson [1], among others, there is a CLT for $\hat{\gamma}_j$ $(1 \leq j \leq D)$, stating that if X is multivariate normal, $\Sigma > 0$ and λ_j simple,

$$\sqrt{n}(\hat{\gamma}_j - \gamma_j) \overset{\mathcal{D}}{\to} \mathcal{N}\left(0, \sum_{j \neq k=1}^{D} \frac{\lambda_j \lambda_k}{\lambda_j - \lambda_k} \gamma_k \gamma_k^T\right). \qquad (10.3)$$

Here, we have assumed, w.l.o.g., that $\gamma_j^T \hat{\gamma}_j \geq 0$.

This CLT has been extended to nonnormal X with existing fourth moment $\mathbb{E}[\|X\|^4]$ by Davis [11] with a more complicated covariance matrix in (10.3). With little effort we reproduce the above result in Corollary 10.4 for $j = 1$ in the context of generalized Fréchet ρ-means.

10.3 Fréchet ρ-Means and Their Strong Laws

What is a good analog to $\mathbb{E}[X]$ when data are no longer vectors but points on a sphere, as are principal components, say? More generally, one may want to statistically assess points on manifolds or even on stratified spaces. For example, data on stratified spaces are encountered in modeling three-dimensional landmark-based shapes by Kendall [36] (cf. Sect. 10.4) or in modeling *phylogenetic descendants trees* in the space introduced by Billera et al. [7].

For a vector-valued random variable X in $\mathbb{R}^D$, upon existence of second moments $\mathbb{E}[\|X\|^2]$ note that,

$$\mathbb{E}[X] = \operatorname*{argmin}_{x \in \mathbb{R}^D} \mathbb{E}\left[\|X - x\|^2\right].$$

For this reason, [21] generalized the classical Euclidean expectation to random deviates X taking values in a metric space (Q, d) via

$$E(X) = \operatorname*{argmin}_{q \in Q} \mathbb{E}\left[d(X, q)^2\right]. \qquad (10.4)$$

In contrast to the Euclidean expectation, $E(X)$ can be set-valued, as is easily seen by a symmetry argument for $Q = \mathbb{S}^{D-1}$ equipped with the spherical metric d and X uniform on $\mathbb{S}^{D-1}$. Then $E(X) = \mathbb{S}^{D-1}$.

Revisiting PCA, note that PCs are not elements of the data space $Q = \mathbb{R}^D$ but elements of $\mathbb{S}^{D-1}$, or more precisely, elements of real projective space of dimension $D - 1$

$$\mathbb{R}P^{D-1} = \{[x] : x \in \mathbb{S}^{D-1}\} = \mathbb{S}^{D-1}/\mathbb{S}^0 \text{ with } [x] = \{-x, x\}.$$

Moreover, the PCs (as elements in $\mathbb{S}^{D-1}$) are also solutions to a minimization problem, e.g. for the first PC we have

$$\gamma_1 \in \underset{\gamma \in \mathbb{S}^{D-1}}{\operatorname{argmin}} \left(\operatorname{var}[X] - \gamma^T \operatorname{cov}[X]\gamma\right). \tag{10.5}$$

Since $\operatorname{var}[X] - \gamma^T \operatorname{cov}[X]\gamma = \mathbb{E}\Big[\|X - \gamma^T X \gamma\|^2\Big]$, in case of $\mathbb{E}[X] = 0$, this motivates the following distinction between *data space* and *descriptor space* leading to *Fréchet* ρ*-means*.

Definition 10.1 (Generalized Fréchet Means) Let Q, P be topological spaces and let $\rho : Q \times P \to \mathbb{R}$ be continuous. We call Q the *data space*, P the *descriptor space* and ρ the *link function*. Suppose that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ are random elements on Q with the property that

$$F(p) = \mathbb{E}[\rho(X, p)^2], \quad F_n(p) = \frac{1}{n}\sum_{j=1}^{n} \rho(X_j, p)^2\,,$$

called the *population* and *sample Fréchet functions*, are finite for all $p \in P$. Every minimizer of the population Fréchet function is called a *population Fréchet mean* and every minimizer of the sample Fréchet function is called a *sample Fréchet mean*. The corresponding sets are denoted by

$$E = \underset{p \in P}{\operatorname{argmin}}\, F(p), \quad E_n = \underset{p \in P}{\operatorname{argmin}}\, F_n(p)\,. \qquad \square$$

Remark By construction, E_n and E are closed sets, but they may be empty without additional assumptions. □

For the following we require that the topological space P is equipped with a *loss function* d, i.e.

1. $d : P \times P \to [0, \infty)$, is a continuous function
2. that vanishes only on the diagonal, that is $d(p, p') = 0$ if and only if $p = p'$.

We now consider the following two versions of a set valued strong law,

$$\bigcap_{n=1}^{\infty} \overline{\bigcup_{k=n} E_k} \subseteq E \text{ a.s.} \tag{10.6}$$

$$\forall \epsilon > 0\, \exists N \in \mathbb{N} \text{ such that } E_n \subseteq \{p \in P : d(E, p) < \epsilon\}\, \forall n \geq N \text{ a.s.} \tag{10.7}$$

In (10.7), N is random as well.

Ziezold [51] established (10.6) for separable $P = Q$ and $\rho = d$ a quasi-metric. Notably, this also holds in case of void E. Bhattacharya and Patrangenaru [5] proved (10.7) under the additional assumptions that $E \neq \emptyset$, $\rho = d$ is a metric and $P = Q$ satisfies the Heine–Borel property (stating that every closed bounded subset is compact). Remarkably, (10.6) implies (10.7) for compact spaces P; this has been observed by Bhattacharya and Patrangenaru [5, Remark 2.5] for $P = Q$ and $\rho = d$ a metric and their argument carries over at once to the general case.

For generalized Fréchet ρ-means we assume the following strongly relaxed analogs of the triangle inequality for (quasi-)metrics.

Definition 10.2 Let Q, P be topological spaces with link function ρ and let d be a loss function on P. We say that (ρ, d) is *uniform* if

$$\begin{aligned} &\forall p \in P, \epsilon > 0\ \exists \delta = \delta(\epsilon, p) > 0 \text{ such that} \\ &|\rho(x, p') - \rho(x, p)| < \epsilon\ \forall x \in Q, p' \in P \text{ with } d(p, p') < \delta. \end{aligned}$$

Further, we say that (ρ, d) is *coercive*, if $\forall p_0, p^* \in P$ and $p_n \in P$ with $d(p^*, p_n) \to \infty$,

$$\begin{aligned} &d(p_0, p_n) \to \infty \ \text{ and } \exists C > 0 \text{ such that} \\ &\rho(x, p_n) \to \infty\ \forall x \in Q \text{ with } \rho(x, p_0) < C \end{aligned}$$

Theorem ([27]) *With the notation of Definition 10.1 we have (10.6) if (ρ, d) is uniform and P is separable. If additionally (ρ, d) is coercive, $E \neq \emptyset$ and $\cap_{n=1}^{\infty}\overline{\cup_{k=n} E_k}$ satisfies the Heine Borel property with respect to d then (10.7) holds true.* □

Let us conclude this section with another example. In biomechanics, e.g. traversing skin markers placed around the knee joint (e.g. [49]), or in medical imaging, modeling deformation of internal organs via skeletal representations (cf. [47]), typical motion of markers occurs naturally along small circles in $\mathbb{S}^2$, c.f. [46]. For a fixed number $k \in \mathbb{N}$, considering k markers as one point $q = (q_1, \ldots, q_k) \in (\mathbb{S}^2)^k$, define the descriptor space P of k concentric small circles $p = (p_1, \ldots, p_k)$ defined by a common axis $w \in \mathbb{S}^2$ and respective latitudes $0 < \theta_1 < \ldots < \theta_k < \pi$. Setting

$$\rho(q, p) = \sqrt{\sum_{j=1}^{k} \min_{y \in p_j} \arccos^2 y^T q_j}$$

and

$$d(p, p') = \sqrt{\arccos^2(w^T w') + \sum_{j=1}^{k} (\theta_j - \theta'_j)^2}$$

we obtain a link ρ and a loss d which form a uniform and coercive pair. Moreover, even P satisfies the Heine–Borel property.

10.4 Procrustes Analysis Viewed Through Fréchet Means

Long before the notion of Fréchet means entered the statistics of shape, *Procrustes analysis* became a tool of choice and has been ever after for the statistical analysis of shape.

Kendall's Shape Spaces
Consider n geometric objects in $\mathbb{R}^m$, each described by k landmarks ($n, k, m \in \mathbb{N}$), i.e. every object is described by a matrix $X_j \in \mathbb{R}^{m\times k}$ ($1 \leq j \leq n$), the columns of which are the k landmark vectors in $\mathbb{R}^m$ of the j-th object. When only the *shape* of the objects is of concern, consider every

$$\lambda_j R_j (X_j - a_j \, 1_k^T/n)$$

equivalent with X_j, where $\lambda_j \in (0,\infty)$ reflects size, $R_j \in SO(m)$ rotation and $a_j \in \mathbb{R}^m$ translation. Here, 1_k is the k-dimensional column vector with all entries equal to 1. Note that the canonical quotient topology of $\mathbb{R}^{m\times k}/(0,\infty)$ gives a non-Hausdorff space which is a dead end for statistics, because all points have zero distance from one another. For this reason, one projects instead to the unit sphere $\mathbb{S}^{m\times k-1}$ and the canonical quotient

$$\Sigma_m^k = \mathbb{S}^{m\times k-1}/SO(m)/\mathbb{R}^m \cong \mathbb{S}^{m\times(k-1)-1}/SO(m)$$

is called *Kendall's shape space*, for details see [13].

Procrustes Analysis
Before the introduction of Kendall's shape spaces, well aware that the canonical quotient is statistically meaningless, [23] suggested to minimize the *Procrustes sum of squares*

$$\sum_{j=1}^{n} \left\| \lambda_i R_i (X_i - a_i \, 1_k^T/n) - \lambda_j R_j (X_j - a_j \, 1_k^T/n) \right\|^2$$

over $\lambda_j, R_j, a_j \in (0,\infty) \times SO(m) \times \mathbb{R}^m$ ($1 \leq j \leq n$) under the constraining condition

$$\left\| \sum_{j=1}^{n} \lambda_j R_j (X_j - a_j \, 1_k^T/n) \right\| = 1\,.$$

It turns out that the minimizing a_j are the mean landmarks, so for the following, we may well assume that every X_j is centered, i.e. $X_j 1_k = 0$ and dropping one landmark, e.g. via Helmertizing, i.e. by multiplying each X_j with a *sub-Helmert* matrix $\mathcal{H}$

$$\mathcal{H} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{k(k-1)}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{k(k-1)}} \\ 0 & -\frac{2}{\sqrt{6}} & \cdots & \frac{1}{\sqrt{k(k-1)}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -\frac{k-1}{\sqrt{k(k-1)}} \end{pmatrix} \in M(k, k-1)$$

from the right, see [13], we may even more assume that $X_j \in \mathbb{R}^{m\times(k-1)}$ $(j = 1, \ldots, n)$. Further, with minimizing λ_j, R_j, every *Procrustes mean*

$$\mu = \frac{1}{n}\sum_{j=1}^{n} \lambda_j R_j X_j$$

is also a representative of a Fréchet mean on $Q = P = \Sigma_m^k$ using the canonical quotient $\rho = d$ of the residual quasi-metric

$$\widetilde{\rho}(X, Y) = \|X - (X^T Y)Y\| \tag{10.8}$$

on $\mathbb{S}^{m\times(k-1)-1}$, in this context, called the *pre-shape space*, see [26] for a detailed discussion.

If μ is a Procrustes mean with minimizing λ_j, R_j $(j = 1, \ldots, n)$, notably, this implies $\operatorname{trace}(R_j X_j^T \mu) = \lambda_j$, then

$$\lambda_j R_j X_j - \mu$$

are called the *Procrustes residuals*. By construction they live in the tangent space $T_\mu \mathbb{S}^{m\times(k-1)-1}$ of $\mathbb{S}^{m\times(k-1)-1}$ at μ. In particular, this is a linear space and hence, the residuals can be subjected to PCA. Computing the Procrustes mean and performing PCA for the Procrustes residuals is *full Procrustes analysis* as proposed by Gower [23].

Note that at this point, we have neither a CLT for Procrustes means nor can we apply the CLT (10.3) because the tangent space is random.

This nested randomness can be attacked directly by nested subspace analysis in Sect. 10.7 or circumvented by the approach detailed in the Sect. 10.6. Let us conclude the present section by briefly mentioning an approach for Riemannian manifolds similar to Procrustes analysis.

Principal Geodesic Analysis

Suppose that $Q = P$ is a Riemannian manifold with intrinsic geodesic distance $\rho = d$. Fréchet means with respect to ρ are called *intrinsic means* and [20] compute an intrinsic mean μ and perform PCA with the data mapped under the inverse exponential at μ to the tangent space $T_\mu Q$ of Q at μ. Again, the base point of the tangent space is random, prohibiting the application of the CLT (10.3).

10.5 A CLT for Fréchet ρ-Means

For this section we require the following assumptions.

(A1) $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ are random elements in a topological data space Q, which is linked to a topological descriptor space P via a continuous function $\rho : Q \times P \to \mathbb{R}$, featuring a unique Fréchet ρ-mean $\mu \in P$.

(A2) There is a loss function $d : P \times P \to [0, \infty)$ and P has locally the structure of a D-dimensional manifold near μ, i.e. there is an open set $U \subset P$, $\mu \in U$ and a homeomorphism $\phi : U \to V$ onto an open set $V \subset \mathbb{R}^D$. W.l.o.g. assume that $\phi(\mu) = 0 \in V$.

(A3) In local coordinates the population Fréchet function is twice differentiable at μ with non-singular Hessian there, i.e. for $p \in U$, $x = \phi^{-1}(p)$,

$$F(p) = F(\mu) + x^T \frac{1}{2} \operatorname{Hess}\left(F \circ \phi^{-1}\right)(0)\, x + o(\|x\|^2)\,,$$

$$H := \operatorname{Hess}\left(F \circ \phi^{-1}\right)(0) > 0\,.$$

(A4) The gradient $\dot\rho_0(X) := \operatorname{grad}_x \rho(X, \phi^{-1}(x))^2|_{x=0}$ exists almost surely and there is a measurable function $\dot\rho : Q \to \mathbb{R}$, satisfying $\mathbb{E}[\dot\rho(X)^2] < \infty$, such that the following Lipschitz condition

$$|\rho(X, \phi^{-1}(x_1))^2 - \rho(X, \phi^{-1}(x_2))^2| \le \dot\rho(X)\|x_1 - x_2\| \text{ a.s.}$$

holds for all $x_1, x_2 \in U$.

Theorem 10.3 *Under the above Assumptions (A1)–(A4), if $\mu_n \in E_n$ is a measurable selection of sample Fréchet ρ-means with $\mu_n \overset{\mathbb{P}}{\to} \mu$, then*

$$\sqrt{n}\phi^{-1}(\mu_n) \overset{\mathcal{D}}{\to} \mathcal{N}\left(0, H^{-1}\operatorname{cov}[\dot\rho_0(X)]H^{-1}\right).$$

Proof We use [17, Theorem 2.11] for $r = 2$. While this theorem has been formulated for intrinsic means on manifolds, upon close inspection, the proof utilizing empirical process theory from [50], rests only on the above assumptions, so that it can be transferred word by word to the situation of Fréchet ρ-means. □

Remark Since the seminal formulation of the first version of the CLT for intrinsic means on manifolds by Bhattacharya and Patrangenaru [6] there has been a vivid discussion on extensions and necessary assumptions (e.g. [2–4, 19, 24, 25, 33, 37, 39, 41]). Recently it has been shown that the rather complicated assumptions originally required by Bhattacharya and Patrangenaru [6] could be relaxed to the above. Further relaxing the assumption $H > 0$ yields so-called *smeary* CLTs, cf. [17]. □

Classical PCA as a Special Case of Fréchet ρ-Means

As an illustration how asymptotic normality of PCs shown by Anderson [1] in an elaborate proof follows simply from Theorem 10.3 we give the simple argument for the first PC.

Corollary 10.4 *Suppose that* $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ *are random vectors in* $\mathbb{R}^D$ *with* $\mathbb{E}[X] = 0$, *finite fourth moment and orthogonal PCs* $\gamma_1, \dots, \gamma_D \in S^{D-1}$ *to descending eigenvalues* $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_D > 0$. *Further let* $\hat{\gamma}_1$ *be a first sample PC with* $\hat{\gamma}_1^T \gamma_1 \geq 0$ *and local coordinates* $\hat{x}_n = \hat{\gamma}_1 - \gamma_1^T \hat{\gamma}_1 \, \gamma_1$. *Then, with* $H^{-1} = \sum_{k=2}^D \frac{1}{\lambda_1 - \lambda_k} \gamma_k \gamma_k^T$,

$$\sqrt{n}\hat{x}_n \overset{\mathcal{D}}{\to} \mathcal{N}(0, H^{-1} \mathrm{cov}[XX^T \gamma_1] H^{-1}) \,,$$

If X is multivariate normal then the covariance of the above r.h.s. is given by the r.h.s. of (10.3) for $j = 1$. □

Proof With the representation $\gamma = x + \sqrt{1 - \|x\|^2} \, \gamma_1 \in \mathbb{S}^{D-1}$, $\gamma_1 \perp x \in U \subset T_{\gamma_1} S^{D-1} \subset \mathbb{R}^D$, we have that the link function underlying (10.5) is given by

$$\begin{aligned} \rho(X, x)^2 &= \|X - \gamma^T X \, \gamma\|^2 \ = \ \|X\|^2 - (\gamma^T X)^2 \\ &= \|X\|^2 - (x^T X + \sqrt{1 - \|x\|^2} \, \gamma_1^T X)^2 \\ &= \|X\|^2 - x^T X X^T x - (1 - \|x\|^2)(\gamma_1^T X)^2 - 2x^T X \sqrt{1 - \|x\|^2} \, \gamma_1^T X \,. \end{aligned}$$

From

$$\mathrm{grad}_x \rho(X, x)^2 = -2XX^T x + 2x(\gamma_1^T X)^2 - 2\left(\sqrt{1 - \|x\|^2} X - \frac{x^T X x}{\sqrt{1 - \|x\|^2}}\right) \gamma_1^T X$$

and, with the unit matrix I,

$$\mathrm{Hess}_x \rho(X, x)^2 = -2XX^T + 2I(\gamma_1^T X)^2 + 2\left(\frac{Xx^T + xX^T - Xx^T}{\sqrt{1 - \|x\|^2}} + \frac{x^T X x x^T}{(1 - \|x\|^2)^{3/2}}\right) \gamma_1^T X,$$

verify that it satisfies Assumption (A4) with $\dot{\rho}_0(X) = -2XX^T \gamma_1$ and $\dot{\rho}(X) = 4\|XX^T \gamma_1\|$ for U sufficiently small, which is square integrable by hypothesis. Since $\mathrm{Hess}_x \rho(X, x)^2|_{x=0} = 2(\gamma_1^T XX^T \gamma_1 I - XX^T)$, with

$$H = 2\mathbb{E}[\gamma_1^T XX^T \gamma_1 I - XX^T] = 2 \sum_{k=2}^{D} (\lambda_1 - \lambda_k) \gamma_k \gamma_k^T \,,$$

which is, by hypothesis, positive definite in $T_{\gamma_1} \mathbb{S}^{D-1}$, we obtain the first assertion of Theorem 10.3. Since in case of multivariate normality $X = \sum_{k=1}^D c_k \gamma_k$ with independent real random variables $c_1, \dots, c_D$, the second assertion follows at once. □

The CLT for Procrustes Means

For $m \geq 3$, Kendall's shape spaces are stratified as follows. There is an open and dense manifold part $(\Sigma_m^k)^*$ and a lower dimensional rest $(\Sigma_m^k)^0$ that is similarly stratified (comprising a dense manifold part and a lower dimensional rest, and so on), e.g. [9, 32, 38]. For a precise definition of stratified spaces, see the following Sect. 10.6.

As a toy example one may think of the unit two-sphere $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$ on which $SO(2) \subset SO(3)$ acts via

$$\begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 \cos\phi + x_2 \sin\phi \\ -x_1 \sin\phi + x_2 \cos\phi \\ x_3 \end{pmatrix}.$$

The canonical quotient space has the structure of the closed interval $\mathbb{S}^2/SO(2) \cong [-1, 1]$ in which $(-1, 1)$ is an open dense one-dimensional manifold and $\{1, -1\}$ is the rest, a zero-dimensional manifold.

Let $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} Z$ be random configurations of m-dimensional objects with k landmarks, with pre-shapes $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ in $\mathbb{S}^{m\times(k-1)-1}$ and shapes $\xi_1, \dots, \xi_n \overset{i.i.d.}{\sim} \xi$ in Σ_m^k with the link function ρ given by the Procrustes metric from the pre-shape (i.e. residual) quasi-metric (10.8).

Theorem (Manifold Stability, cf. [28, 29]) *If, with the above setup, $\mathbb{P}\{\xi \in (\Sigma_m^k)^*\} > 0$ and if the probability that two shapes are maximally remote is zero then every Procrustes mean μ is assumed on the manifold part.* □

In consequence, for $Q = P = \Sigma_m^k$, if the manifold part is assumed at all, Assumption (A1) implies Assumption (A2). With the same reasoning as in the proof of Corollary 10.4, Assumption (A4) is verified. This yields the following.

Corollary *Let $Z_1, \dots, Z_n \overset{i.i.d.}{\sim} Z$ be random configurations of m-dimensional objects with k landmarks, with pre-shapes $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ in $\mathbb{S}^{D\times(k-1)-1}$ and shapes $\xi_1, \dots, \xi_n \overset{i.i.d.}{\sim} \xi$ in Σ_m^k such that*

- $\mathbb{P}\{\xi \in (\Sigma_m^k)^*\} > 0$, *the probability that two shapes are maximally remote is zero and*
- *Assumptions (A1) and (A3) are satisfied.*

Then, every measurable selection μ_n of Procrustes sample means satisfies a CLT as in Theorem 10.3. □

10.6 Geodesic Principal Component Analysis

In this section we assume that random deviates $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ take values in a Riemann stratified space Q.

Definition (Stratified Space) A stratified space Q of dimension m embedded in a Euclidean space can be defined as a direct sum

$$Q = \bigcup_{j=1}^{k} Q_{d_j}$$

such that $0 \le d_1 < \dots < d_k = m$, each Q_{d_j} is a d_j-dimensional manifold and $Q_{d_j} \cap Q_{d_l} = \emptyset$ for $j \neq l$.

A stratified space $Q \subset \mathbb{R}^s$ is called *Whitney stratified*, if for every $j < l$

(i) If $Q_{d_j} \cap \overline{Q_{d_l}} \neq \emptyset$ then $Q_{d_j} \subset \overline{Q_{d_l}}$.

(ii) For sequences $x_1, x_2, \dots \in Q_{d_j}$ and $y_1, y_2, \dots \in Q_{d_l}$ which converge to the same point $x \in Q_{d_j}$ such that the sequence of secant lines c_i between x_i and y_i converges to a line c as $i \to \infty$, and such that the sequence of tangent planes $T_{y_i} Q_{d_l}$ converges to a d_l dimensional plane T as $i \to \infty$, the line c is contained in T.

We call a Whitney stratified space *Riemann stratified*, if

(iii) for every $j < l$ and sequence $y_1, y_2, \dots \in Q_{d_l}$ which converges to the point $x \in Q_{d_j}$ the Riemannian metric tensors $g_{l,y_i} \in T^2_{y_i} Q_{d_l}$ converge to a rank two tensor $g_{l,x} \in T \otimes T$ and the Riemannian metric tensor $g_{j,x} \in T^2_x Q_{d_j}$ is given by the restriction $g_{j,x} = g_{l,x}|_{T^2_x Q_{d_j}}$.

□

Geodesics, i.e. curves of locally minimal length, exist locally in every stratum Q_{d_j}. Due to the Whitney condition, a geodesic can also pass through strata of different dimensions if these strata are connected. Property (ii) is called Whitney condition B and it follows from this condition that $T_x Q_{d_j} \subset T$, which is called Whitney condition A, e.g. [22].

Of course, all Riemannian manifolds are stratified spaces. Typical examples for stratified spaces that are not Riemannian manifolds are Kendall's shape spaces Σ^k_m for $m \ge 3$ dimensional objects with $k \ge 4$ landmarks or the BHV space of phylogenetic descendants trees $\mathcal{T}_n$ with $n \ge 3$ leaves.

Let $\Gamma(Q)$ be the space of point sets of maximal geodesics in Q. With the intrinsic geodesic metric d_Q on Q we have the link function

$$\rho : Q \times \Gamma(Q) \to [0, \infty), \quad (q, \gamma) \mapsto \inf_{q' \in \gamma} d_Q(q, q') .$$

Further, we assume that $\Gamma(Q)$ also carries a metric d_Γ. This can be either Hausdorff distance based on d_Q, or a quotient metric, e.g. induced from $\Gamma(Q) = (Q \times Q)/\sim$ with a suitable equivalence relation. An example for the latter is the identification of $\Gamma(\mathbb{S}^{D-1})$ with $G(D,2)$, the Grassmannian structure of the space of two-dimensional linear subspaces in $\mathbb{R}^D$ (every geodesic on $\mathbb{S}^{D-1}$ is a great circle which is the intersection with $\mathbb{S}^{D-1} \subset \mathbb{R}^D$ of a plane through the origin).

Definition 10.5 (cf. [32]) With the above assumptions, setting $P_0 = \Gamma(Q)$, every population Fréchet ρ-mean on $P = P_0$ is a first population *geodesic principal component* (GPC) and every such sample mean is a first sample GPC.

Given a unique first population GPC γ_1, setting $P_1 = \{\gamma \in \Gamma(Q) : \gamma \cap \gamma_1 \neq \emptyset$ and $\gamma \perp \gamma_1$ there$\}$, every population Fréchet ρ-mean on $P = P_1$ is a second population GPC.

Higher order population GPCs are defined by requiring them to pass through a common point $p \in \gamma_1 \cap \gamma_2$ and being orthogonal there to all previous unique population GPCs.

Similarly, for the second sample GPC, for a given unique first GPC $\hat{\gamma}_1$ use $P = \hat{P}_1 = \{\gamma \in \Gamma(Q) : \gamma \cap \gamma_1 \neq \emptyset$ and $\gamma \perp \hat{\gamma}_1$ there$\}$ and higher order sample GPCs are defined by requiring them to pass through a common point $\hat{p} \in \hat{\gamma}_1 \cap \hat{\gamma}_2$ and being orthogonal there to all previous unique sample GPCs.

The GPC scores are the orthogonal projections of X, or of the data, respectively, to the respective GPCs. □

Remark In case of valid assumptions (A1) – (A4) the CLT from Theorem 10.3 yields asymptotic $\sqrt{n}$-normality for the first PC in a local chart. An example and an application to $Q = \Sigma_2^k$ $(k \geq 3)$ can be found in [27]. □

Obviously, there are many other canonical intrinsic generalizations of PCA to non-Euclidean spaces, e.g. in his *horizontal PCA* [48] defines the second PC by a parallel translation of a suitable tangent space vector, orthogonally along the first PC. One difficulty is that GPCs usually do not define subspaces, as classical PCs do, which define affine subspaces. However, there are stratified spaces which have rich sets of preferred subspaces.

Definition (Totally Geodesic Subspace) A Riemann stratified space $S \subset Q$ with Riemannian metric induced by the Riemannian metric of a Riemann stratified space Q is called *totally geodesic* if every geodesic of S is a geodesic of Q. □

The totally geodesic property is transitive in the following sense. Consider a sequence of Riemann stratified subspaces $Q_1 \subset Q_2 \subset Q_3$ where Q_1 is totally geodesic with respect to Q_2 and Q_2 is totally geodesic with respect to Q_3. Then Q_1 is also totally geodesic with respect to Q_3.

In the following, we will use the term *rich space of subspaces* for a space of k-dimensional Riemann stratified subspaces of an m-dimensional Riemann stratified space, if it has dimension at least $(m-k)(k+1)$. This means that the space of k-dimensional subspaces has at least the same dimension as the space of affine k-dimensional subspaces in $\mathbb{R}^m$. If a Riemann stratified space has a rich space of sequences of totally geodesic subspaces $Q_0 \subset Q_1 \subset \cdots \subset Q_{m-1} \subset Q_m = Q$ where every Q_j is a Riemann stratified space of dimension j, a generalization which is very close in spirit to PCA can be defined. This is especially the case, if Q has a rich space of $(m-1)$-dimensional subspaces which are of the same class as Q. For example, the sphere S^m has a rich space of great subspheres S^{m-1}, which are totally geodesic submanifolds. Therefore, spheres are well suited to introduce an analog of PCA, and [34, 35] have defined *principal nested spheres* (PNS) which even exist as *principal nested small spheres*, which are not totally geodesic, however. In the latter case the dimension of the space of k-dimensional submanifolds is even $(m-k)(k+2)$, cf. [30].

Generalizing this concept [43], has introduced *barycentric subspaces*, cf. Chapter 18 of this book.

In the following penultimate section we develop an inferential framework for such nested approaches.

10.7 Backward Nested Descriptors Analysis (BNDA)

As seen in Definition 10.5, higher order GPCs depend on lower order GPCs and are hence defined in a nested way. More generally, one can consider sequences of subspaces of descending dimension, where every subspace is also contained in all higher dimensional subspaces. Here we introduce the framework of *backward nested families of descriptors* to treat such constructions in a very general way.

In Sect. 10.9 we introduce several examples of such backward nested families of descriptors.

Definitions and Assumptions 10.6 Let Q be a separable topological space, called the *data space* and let $\{P_j\}_{j=0}^m$ be a family of separable topological spaces called *descriptor spaces*, each equipped with a loss function $d_j : P_j \times P_j \to [0,\infty)$ (i.e. it is continuous and vanishes exactly on the diagonal) with $P_m = \{Q\}$ $(j = 1,\ldots,m)$.

Next, assume that every $p \in P_j$ $(j = 1,\ldots,m)$ is itself a topological space giving rise to a topological space $\emptyset \neq S_p \subseteq P_{j-1}$ with a continuous function $\rho_p : p \times S_p \to [0,\infty)$ called a *link function*.

Further, assume that for all $p \in P_j$ $(j = 1,\ldots,m)$ and $s \in S_p$ there exists a measurable mapping $\pi_{p,s} : p \to s$ called *projection*.

Then for $j \in \{1,\ldots,m\}$ every

$$f = \{p^m,\ldots,p^j\}, \text{ with } p^{l-1} \in S_{p^l}, l = j+1,\ldots,m$$

is a *backward nested family of descriptors* (BNFD) from P_m to P_j which lives in the space

$$P_{m,j} = \left\{ f = \{p^l\}_{l=m}^{j} : p^{l-1} \in S_{p^l}, l = j+1,\ldots,m \right\},$$

with *projection* along each descriptor

$$\pi_f = \pi_{p^{j+1},p^j} \circ \ldots \circ \pi_{p^m,p^{m-1}} : p^m \to p^j$$

For another BNFD $f' = \{p'^l\}_{l=m}^j \in T_{m,j}$ set

$$d^j(f, f') = \sqrt{\sum_{l=m}^{j} d_j(p^l, p'^l)^2}\,.$$

Definition With the above Definitions and Assumptions 10.6, random elements $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ on the data space Q admitting BNFDs give rise to *backward nested population* and *sample means* (BN-means) $f = (p^m, \ldots, p^j)$ and $f_n = (p_n^m, \ldots, p_n^j)$, respectively, recursively defined via $f^m = \{Q\} = f_n^m$, i.e. $p^m = Q = p_n^m$ and for $j = m, \ldots, 2$,

$$p^{j-1} \in \underset{s \in S_{p^j}}{\operatorname{argmin}}\, \mathbb{E}[\rho_{p^j}(\pi_{f^j} \circ X, s)^2], \qquad f^{j-1} = \{p^l\}_{l=m}^{j-1}$$

$$p_n^{j-1} \in \underset{s \in S_{p_n^j}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_{p_n^j}(\pi_{f_n^j} \circ X_i, s)^2, \qquad f_n^{j-1} = \{p_n^l\}_{l=m}^{j-1}\,.$$

If all of the population minimizers are unique, we speak of *unique BN-means*.

Remark A nested sequence of subspaces is desirable for various scenarios. Firstly, it can serve as a basis for dimension reduction as is also often done using PCA in Euclidean space. Secondly, the residuals of projections along the BNFD can be used as residuals of orthogonal directions in Euclidean space in order to achieve a "Euclideanization" of the data (e.g. [44]). Thirdly, lower dimensional representations of the data or scatter plots of residuals can be used for more comprehensible visualization.

Backward approaches empirically achieve better results than forward approaches, starting from a point and building up spaces of increasing dimension, in terms of data fidelity. The simplest example, determining the intrinsic mean first and then requiring the geodesic representing the one-dimensional subspace to pass through it, usually leads to higher residual variance than fitting the principal geodesic without reference to the mean. □

For a strong law and a CLT for BN-means we require assumptions corresponding to Definition 10.2 and corresponding to assumptions in [4]. Both sets of assumptions are rather complicated, so that they are only referenced here.

(B1) Assumptions 3.1–3.6 from [31]

(B2) Assumption 3.10 from [31]

To the best knowledge of the authors, instead of (B2), more simple assumptions corresponding to (A1)–(A4) from Sect. 10.5 have not been derived for the backward nested descriptors scenario.

Theorem ([31]) *If the BN population mean $f = (p^m, \ldots, p^j)$ is unique and if $f_n = (p_n^m, \ldots, p_n^j)$ is a measurable selection of BN sample means then under (B1),*

$$f_n \to f \text{ a.s.}$$

i.e. there is $\Omega' \subseteq \Omega$ measurable with $\mathbb{P}(\Omega') = 1$ such that for all $\epsilon > 0$ and $\omega \in \Omega'$, there is $N(\epsilon, \omega) \in \mathbb{N}$ with

$$d(f_n, f) < \epsilon \quad \text{for all } n \geq N(\epsilon, \omega)\,.$$

Theorem 10.7 ([31]) *Under Assumptions (B2), with unique BN population mean $f \in P = P_{m,j}$ and local chart ϕ with $\phi^{-1}(0) = f$, for every measurable selection f_n of BN sample means $f_n \overset{\mathbb{P}}{\to} f$, there is a symmetric positive definite matrix B_ϕ such that*

$$\sqrt{n}\,\phi^{-1}(f_n) \overset{\mathcal{D}}{\to} \mathcal{N}(0, B_\phi)\,.$$

Remark 10.8 Under *factoring charts* as detailed in [31], asymptotic normality also holds for the last descriptor $p_n^{j-1} \overset{\text{a.s.}}{\to} p^{j-1}$,

$$\sqrt{n}\phi^{-1}(p_n^{j-1}) \overset{\mathcal{D}}{\to} \mathcal{N}(0, C_\phi)$$

with a suitable local chart ϕ such that $\phi^{-1}(0) = p^{j-1}$ and a symmetric positive definite matrix C_ϕ. □

10.8 Two Bootstrap Two-Sample Tests

Exploiting the CLT for ρ-means, BN-means or the last descriptor of BN-means (cf. Remark 10.8) in order to obtain an analog of the two-sample test (10.2), we inspect its ingredients. Suppose that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ and $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Y$ are independent random elements on Q. In case of ρ-means, we assume that Assumptions (A1)–(A4) from Sect. 10.5 are valid and in case of BN-means (or a last descriptor thereof) assume that Assumption (B2) from Sect. 10.7 is valid for X and Y, in particular that unique means μ^X and μ^Y lie within one single open set $U \subset P$ that homeomorphically maps to an open set $V \subset \mathbb{R}^D$ under ϕ. With measurable selections $\hat{\mu}_n^X$ and $\hat{\mu}_m^Y$ of sample means, respectively, replace $\bar{X}_n - \bar{Y}_m$ with $\phi^{-1}\big(\hat{\mu}_n^X\big) - \phi^{-1}\big(\hat{\mu}_m^Y\big) \in \mathbb{R}^D$.

Obviously, $\hat{\Sigma}_n^X$ and $\hat{\Sigma}_m^Y$ are not directly assessable, however. If one had a large number B of samples $\{X_{1,1}, \dots, X_{1,n}\}, \dots, \{X_{B,1}, \dots, X_{B,n}\}$, one could calculate the descriptors $\phi^{-1}\big(\hat{\mu}_{n,1}^X\big), \dots, \phi^{-1}\big(\hat{\mu}_{n,B}^X\big)$ and estimate the covariance of these. But since we only have one sample, we use the bootstrap instead. The idea of the n out-of n non-parametric bootstrap (e.g. [12, 15]) is to generate a large number B of *bootstrap samples* $\{X_1^{*1}, \dots, X_n^{*1}\}, \dots, \{X_1^{*B}, \dots, X_n^{*B}\}$ of the same size n by drawing with replacement from the sample $X_1, \dots, X_n$. From each of these bootstrap samples one can calculate estimators $\phi^{-1}\big(\mu_n^{X,*1}\big), \dots, \phi^{-1}\big(\mu_n^{X,*B}\big)$, which serve as so-called *bootstrap estimators* of μ. From these, one can now calculate the estimator for the covariance of $\hat{\mu}_n^X$

$$\Sigma_n^{X,*} := \frac{1}{B}\sum_{j=1}^{B}\Big(\phi^{-1}(\mu_n^{X,*j}) - \phi^{-1}(\hat{\mu}_n^X)\Big)\Big(\phi^{-1}(\mu_n^{X,*j}) - \phi^{-1}(\hat{\mu}_n^X)\Big)^T \tag{10.9}$$

For the First Test

Perform B_X times n out-of n bootstrap from $X_1, \dots, X_n$ to obtain Fŕechet ρ-means $\mu_n^{X,*1}, \dots, \mu_n^{X,*B_X}$ and replace $\hat{\Sigma}_n^X$ with the n-fold of the bootstrap covariance $\Sigma_n^{X,*}$ as defined in Eq. (10.9). With the analog m out-of m bootstrap, replace $\hat{\Sigma}_m^Y$ with $m\Sigma_m^{Y,*}$.

Then, under $H_0 : \mu^X = \mu^Y$, if $m/n \to 1$ or $n\,\mathrm{cov}\big[\phi^{-1}(\mu_n^X)\big] = m\,\mathrm{cov}\big[\phi^{-1}(\mu_m^Y)\big]$, under typical regularity conditions, e.g. [10], the statistic

$$T^2 = (\phi^{-1}(\hat{\mu}_n^X) - \phi^{-1}(\hat{\mu}_m^Y))^T\left(\left(\frac{1}{n} + \frac{1}{m}\right)\Sigma_p^*\right)^{-1}(\phi^{-1}(\hat{\mu}_n^X) - \phi^{-1}(\hat{\mu}_m^Y)) \tag{10.10}$$

$$\Sigma_p^* := \frac{1}{n+m-2}\Big(n(n-1)\Sigma_n^{X,*} + m(m-1)\Sigma_m^{Y,*}\Big)$$

adapted from Eq. (10.2), is asymptotically Hotelling distributed as discussed in Sect. 10.2.

For the Second Test

Observe that, alternatively, the test statistic

$$T^2 = \Big(\phi^{-1}(\mu_n^{\widetilde{X}}) - \phi^{-1}(\mu_m^{\widetilde{Y}})\Big)^T\Big(\Sigma_n^{X,*} + \Sigma_m^{Y,*}\Big)^{-1}\Big(\phi^{-1}(\mu_n^{\widetilde{X}}) - \phi^{-1}(\mu_m^{\widetilde{Y}})\Big),$$

can be used. Notably, this second test for $H_0 : \mu_X = \mu_Y$ does not rely on $n/m \to 1$ or equal covariances, as does the first. However, the test statistic is only approximately F distributed, even for normally distributed data and the parameters of the F distribution have to be determined by an approximation procedure.

To enhance the power of either test, quantiles can be determined using the bootstrap. A naive approach would be to pool samples and use $\widetilde{X}$ for the first n

data points, $\widetilde{Y}$ for the last m data points of bootstrapped samples from the pooled data $X_1, \dots, X_n, Y_1, \dots, Y_m$. However, it turns out that this approach suffers from significantly diminished power.

Instead, we generate the same number B of bootstrap samples from $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ separately, thus getting $\mu_n^{X,*1}, \dots, \mu_n^{X,*B}$ and $\mu_m^{Y,*1}, \dots, \mu_m^{Y,*B}$. Due to the CLT 10.7 and Remark 10.8, $\phi^{-1}(\mu_n^{X,*1}), \dots, \phi^{-1}(\mu_n^{X,*B})$ are samples from a distribution which is close to normal with mean $\phi^{-1}(\hat{\mu}_n^X)$. The analog holds for Y. As a consequence, the residuals $d_n^{X,*j} = \phi^{-1}(\mu_n^{X,*j}) - \phi^{-1}(\hat{\mu}_n^X)$ are close to normally distributed with mean 0. To simulate quantiles from the null hypothesis $\mu_n^X = \mu_m^Y$, we therefore only use the residuals $d_{n,*j}^X$ and $d_{m,*j}^Y$ and calculate

$$T_j^2 = \left(d_n^{X,*j} - d_m^{Y,*j})\right)^T \left(\Sigma_n^{X,*} + \Sigma_m^{Y,*}\right)^{-1} \left(d_n^{X,*j} - d_m^{Y,*j}\right) . \tag{10.11}$$

Then we order these values ascendingly and use them as $(j - 1/2)/B$-quantiles as usual for empirical quantiles. Tuning the corresponding test to the right level, its power is usually larger than using the F-quantiles corresponding to the Hotelling distribution.

For a detailed discussion and justification see [16, 31].

10.9 Examples of BNDA

Scenarios of BNDA are given by *flags*, namely, by nested subspaces,

$$Q \supseteq p^m \supseteq p^{m-1} \supseteq \ldots \supseteq p^0 \in Q .$$

We give three examples.

The Intrinsic Mean on the First GPC

It is well known, that the intrinsic mean usually comes not to lie on the first GPC. For example a distribution on $\mathbb{S}^2$ that is uniform on a great circle has this great circle as its first GPC with respect to the spherical metric. The Fréchet mean with respect to this metric is given, as is easily verified, by the two poles having this great circle as the equator. In order to enforce nestedness, we consider the first GPC on a Riemannian manifold and the intrinsic mean on it. The corresponding descriptor spaces are

$$P_2 = \{Q\},\ S_Q = P_1 = \left\{\gamma_{q,v} : (q, v) \in TQ, v \neq 0\right\} / \sim,\ P_0 = Q .$$

with the *tangent bundle* TQ over Q, the maximal geodesic $\gamma_{q,v}$ through q with initial velocity $v \in T_q Q$ and $\gamma_{q,v} \sim \gamma_{q',v'}$ if the two geodesics agree as point sets. Denoting the class of $\gamma_{q,v}$ by $[\gamma_{q,v}]$, it turns out that

$$T_{2,0} = \{(p, s) : p = [\gamma_{q,v}] \in P_1,\ s \in p\} \cong PQ,$$
$$([\gamma_{q,v}], s) \cong \left(s, \left\{\frac{w}{\|w\|}, -\frac{w}{\|w\|}\right\}\right),$$

where $[\gamma_{q,v}] = [\gamma_{s,w}]$ and PQ denotes the *projective bundle* over Q. With the local trivialization of the tangent bundle one obtains a local trivialization of the projective bundle and thus factoring charts, so that, under suitable conditions, Theorem 10.7 and Remark 10.8 are valid. In fact, this construction also works for suitable Riemann stratified spaces, e.g. also for $Q = \Sigma_m^k$ with $m \geq 3$, cf. [31].

Principal Nested Spheres (PNS)
For the special case of $Q = \mathbb{S}^{D-1} \subset \mathbb{R}^D$ let P_j be the space of all j-dimensional unit-spheres ($j = 1, \ldots, D-1$) and $P_{D-1} = \mathbb{S}^{D-1}$. Note that P_j can be given the manifold structure of the Grassmannian $G(D, j+1)$ of $(j+1)$-dimensional linear subspaces in $\mathbb{R}^D$. The corresponding BNDA has been introduced by Jung et al. [34, 35] as *principal nested great spheres analysis* (PNGSA) in contrast to *principal nested small sphere analysis* (PNSSA), when allowing also small subspheres in every step. Notably, estimation of small spheres involves a test for great spheres to avoid overfitting, cf. [18, 34].

Furthermore, PNSSA offers more flexibility than PNGSA because the family of all j-dimensional small subspheres in $\mathbb{S}^{D-1}$ has dimension $\dim\big(G(D, j+1)\big) + D - j$, cf. [18].

As shown in [31], under suitable conditions, Theorem 10.7 and Remark 10.8 are valid for both versions of PNS.

Extensions of PNS to general Riemannian manifolds can be sought by considering flags of totally geodesic subspaces. While there are always geodesics, which are one-dimensional geodesic subspaces, there may be none for a given dimension j. And even, if there are, for instance on a torus, totally geodesic subspaces winding around infinitely are statistically meaningless because they approximate any given data set arbitrarily well. As a workaround, tori can be topologically and geometrically deformed into stratified spheres and on these PNS with all of its flexibility, described above, can be performed, as in [18].

Barycentric Subspace Analysis (BSA) by Pennec [43] constitutes another extension circumventing the above difficulties. Here P_j is the space of *exponential spans* of any $j+1$ points in general position. More precisely, with the geodesic distance d on Q, for $q_1, \ldots, q_{j+1} \in Q$, define their exponential span by

$$\mathfrak{M}(q_1, \ldots, q_{j+1}) = \left\{ \operatorname*{argmin}_{q \in Q} \sum_{k=1}^{j+1} a_k d(q_k, q)^2 : a_1, \ldots, a_{j+1} \in \mathbb{R},\ \sum_{k=1}^{j+1} a_k = 1 \right\}.$$

For an m-dimensional manifold Q a suitable choice of $m+1$ points $q_1, \ldots, q_{m+1} \in Q$ thus yields the flag

$$Q = \mathfrak{M}(q_1, \ldots, q_{m+1}) \supset \mathfrak{M}(q_1, \ldots, q_m) \supset \ldots \supset \mathfrak{M}(q_1, q_2) \supset \{q_1\}\,.$$

For the space of phylogenetic descendants tree by Billera et al. [7] in a similar approach by Nye et al. [42] the *locus of the Fréchet mean* of a given point set has been introduced along with corresponding optimization algorithms.

Barycentric subspaces and similar constructions are the subject of Chapter 11.

To the knowledge of the authors, there have been no attempts, to date, to investigate applicability of Theorem 10.7 and Remark 10.8 to BSA.

10.10 Outlook

Beginning with Anderson's CLT for PCA, we have sketched some extensions of PCA to non-Euclidean spaces and have come up with a rather general CLT, the assumptions of which are more general than those of [4, 6]. Let us conclude with listing a number of open tasks, which we deem of high importance for the development of suitable statistical non-Euclidean tools.

1. Formulate the CLT for BNFDs in terms of assumptions corresponding to Assumptions (A1)–(A4).
2. Apply the CLT for BNFDs to BSA if possible.
3. Formulate BNFD not as a sequential but as a simultaneous optimization problem, derive corresponding CLTs and apply them to BSA with simultaneous estimation of the entire flag.
4. In some cases we have no longer $\sqrt{n}$-Gaussian CLTs but so-called *smeary* CLTs which feature a lower rate, cf. [17]. Extend the CLTs presented here to the general smeary scenario.
5. Further reduce and generalize Assumptions (A1)–(A4), especially identify necessary and sufficient conditions for (A3).

Acknowledgements The authors acknowledge support from the Niedersachen Vorab of the Volkswagen Foundation and support from DFG HU 1575-7.

References

1. Anderson, T.: Asymptotic theory for principal component analysis. Ann. Math. Statist. **34**(1), 122–148 (1963)
2. Barden, D., Le, H., Owen, M.: Central limit theorems for Fréchet means in the space of phylogenetic trees. Electron. J. Probab. **18**(25), 1–25 (2013)
3. Barden, D., Le, H., Owen, M.: Limiting behaviour of fréchet means in the space of phylogenetic trees. Ann. Inst. Stat. Math. **70**(1), 99–129 (2018)
4. Bhattacharya, R., Lin, L.: Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. Proc. Am. Math. Soc. **145**(1), 413–428 (2017)

5. Bhattacharya, R.N., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds I. Ann. Stat. **31**(1), 1–29 (2003)
6. Bhattacharya, R.N., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds II. Ann. Stat. **33**(3), 1225–1259 (2005)
7. Billera, L., Holmes, S., Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27**(4), 733–767 (2001)
8. Billingsley, P.: Probability and Measure, vol. 939. Wiley, London (2012)
9. Bredon, G.E.: Introduction to Compact Transformation Groups. Pure and Applied Mathematics, vol. 46. Academic Press, New York (1972)
10. Cheng, G.: Moment consistency of the exchangeably weighted bootstrap for semiparametric m-estimation. Scand. J. Stat. **42**(3), 665–684 (2015)
11. Davis, A.W.: Asymptotic theory for principal component analysis: non-normal case. Aust. J. Stat. **19**, 206–212 (1977)
12. Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application, vol. 1. Cambridge University Press, Cambridge (1997)
13. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, Chichester (2014)
14. Durrett, R.: Probability: Theory and Examples. Cambridge University Press, Cambridge (2010)
15. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. CRC Press, Boca Raton (1994)
16. Eltzner, B., Huckemann, S.: Bootstrapping descriptors for non-Euclidean data. In: Geometric Science of Information 2017 Proceedings, pp. 12–19. Springer, Berlin (2017)
17. Eltzner, B., Huckemann, S.F.: A smeary central limit theorem for manifolds with application to high dimensional spheres. Ann. Stat. **47**, 3360–3381 (2019)
18. Eltzner, B., Huckemann, S., Mardia, K.V.: Torus principal component analysis with applications to RNA structure. Ann. Appl. Statist. **12**(2), 1332–1359 (2018)
19. Eltzner, B., Galaz-García, F., Huckemann, S.F., Tuschmann, W.: Stability of the cut locus and a central limit theorem for Fréchet means of Riemannian manifolds (2019). arXiv: 1909.00410
20. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.C.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imag. **23**(8), 995–1005 (2004)
21. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Ann. Inst. Henri Poincare **10**(4), 215–310 (1948)
22. Goresky, M., MacPherson, R.: Stratified Morse Theory. Springer, Berlin (1988)
23. Gower, J.C.: Generalized Procrustes analysis. Psychometrika **40**, 33–51 (1975)
24. Hotz, T., Huckemann, S.: Intrinsic means on the circle: uniqueness, locus and asymptotics. Ann. Inst. Stat. Math. **67**(1), 177–193 (2015)
25. Hotz, T., Huckemann, S., Le, H., Marron, J.S., Mattingly, J., Miller, E., Nolen, J., Owen, M., Patrangenaru, V., Skwerer, S.: Sticky central limit theorems on open books. Ann. Appl. Probab. **23**(6), 2238–2258 (2013)
26. Huckemann, S.: Inference on 3D Procrustes means: Tree boles growth, rank-deficient diffusion tensors and perturbation models. Scand. J. Stat. **38**(3), 424–446 (2011)
27. Huckemann, S.: Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Ann. Stat. **39**(2), 1098–1124 (2011)
28. Huckemann, S.: Manifold stability and the central limit theorem for mean shape. In: Gusnanto, A., Mardia, K.V., Fallaize, C.J. (eds.) Proceedings of the 30th LASR Workshop, pp. 99–103. Leeds University Press, Leeds (2011)
29. Huckemann, S.: On the meaning of mean shape: Manifold stability, locus and the two sample test. Ann. Inst. Stat. Math. **64**(6), 1227–1259 (2012)
30. Huckemann, S., Eltzner, B.: Polysphere PCA with applications. In: Proceedings of the Leeds Annual Statistical Research (LASR) Workshop, pp. 51–55. Leeds University Press, Leeds (2015)
31. Huckemann, S.F., Eltzner, B.: Backward nested descriptors asymptotics with inference on stem cell differentiation. Ann. Stat. **46**(5), 1994–2019 (2018)
32. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). Stat. Sin. **20**(1), 1–100 (2010)

33. Huckemann, S., Mattingly, J.C., Miller, E., Nolen, J.: Sticky central limit theorems at isolated hyperbolic planar singularities. Electron. J. Probab. **20**(78), 1–34 (2015)
34. Jung, S., Foskey, M., Marron, J.S.: Principal arc analysis on direct product manifolds. Ann. Appl. Stat. **5**, 578–603 (2011)
35. Jung, S., Dryden, I.L., Marron, J.S.: Analysis of principal nested spheres. Biometrika **99**(3), 551–568 (2012)
36. Kendall, D.G.: The diffusion of shape. Adv. Appl. Probab. **9**, 428–430 (1977)
37. Kendall, W.S., Le, H.: Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. Braz. J. Probab. Stat. **25**(3), 323–352 (2011)
38. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley, Chichester (1999)
39. Le, H., Barden, D.: On the measure of the cut locus of a Fréchet mean. Bull. Lond. Math. Soc. **46**(4), 698–708 (2014)
40. Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis. Academic Press, New York (1980)
41. McKilliam, R.G., Quinn, B.G., Clarkson, I.V.L.: Direction estimation by minimum squared arc length. IEEE Trans. Signal Process. **60**(5), 2115–2124 (2012)
42. Nye, T.M., Tang, X., Weyenberg, G., Yoshida, R.: Principal component analysis and the locus of the fréchet mean in the space of phylogenetic trees. Biometrika **104**(4), 901–922 (2017)
43. Pennec, X.: Barycentric subspace analysis on manifolds. Ann. Stat. **46**(6A), 2711–2746 (2018)
44. Pizer, S.M., Jung, S., Goswami, D., Vicory, J., Zhao, X., Chaudhuri, R., Damon, J.N., Huckemann, S., Marron, J.: Nested sphere statistics of skeletal models. In: Innovations for Shape Analysis, pp. 93–115. Springer, Berlin (2013)
45. Romano, J.P., Lehmann, E.L.: Testing Statistical Hypotheses. Springer, Berlin (2005)
46. Schulz, J.S., Jung, S., Huckemann, S., Pierrynowski, M., Marron, J., Pizer, S.: Analysis of rotational deformations from directional data. J. Comput. Graph. Stat. **24**(2), 539–560 (2015)
47. Siddiqi, K., Pizer, S.: Medial Representations: Mathematics, Algorithms and Applications. Springer, Berlin (2008)
48. Sommer, S.: Horizontal dimensionality reduction and iterated frame bundle development. In: Geometric Science of Information, pp. 76–83. Springer, Berlin (2013)
49. Telschow, F.J., Huckemann, S.F., Pierrynowski, M.R.: Functional inference on rotational curves and identification of human gait at the knee joint (2016). arXiv preprint arXiv:1611.03665
50. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)
51. Ziezold, H.: Expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In: Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes, pp. 591–602. Springer, Berlin (1977)