Bootstrapping Descriptors for Non-Euclidean Data

Benjamin Eltzner(✉) and Stephan Huckemann

Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences,
University of Goettingen, Goettingen, Germany
beltzne@uni-goettingen.de

Abstract. For data carrying a non-Euclidean geometric structure it is natural to perform statistics via geometric descriptors. Typical candidates are means, geodesics, or more generally, lower dimensional subspaces, which carry specific structure. Asymptotic theory for such descriptors is slowly unfolding and its application to statistical testing usually requires one more step: Assessing the distribution of such descriptors. To this end, one may use the bootstrap that has proven to be a very successful tool to extract inferential information from small samples. In this communication we review asymptotics for descriptors of manifold valued data and study a non-parametric bootstrap test that aims at a high power, also under the alternative.

1 Introduction

In recent years, the study of data on non-Euclidean spaces has found increasing attention in statistics. Non-Euclidean data spaces have lead to a surge of specialized fields: directional statistics is concerned with data on spheres of different dimensions (e.g. [15]); shape analysis studies lead to data on quotient spaces (e.g. [6]), some of which are manifolds and some of which are non-manifold stratified spaces; and applications in population genetics have lead to increasing interest in data on non-manifold phylogenetic tree spaces (e.g. [4]) and to graph data in general.

As a basis for statistics on these spaces, it is important to investigate asymptotic consistency of estimators, as has been done for intrinsic and extrinsic Fréchet means on manifolds by [3,8], and more generally for a class of descriptors called *generalized Fréchet means* by [11,12]. Examples of such generalized Fréchet means are not only Procrustes means on non-manifold shape spaces ([6,11]) but also geodesic principal components on such spaces (cf. [10]), or more generally, barycentric subspaces by [17], see also [16] for a similar approach on phylogenetic tree spaces, or more specifically, small and great subspheres for spherical data by [14,18].

B. Eltzner and S. Huckemann—Acknowledging the Niedersachsen Vorab of the Volkswagen Foundation.

© Springer International Publishing AG 2017
F. Nielsen and F. Barbaresco (Eds.): GSI 2017, LNCS 10589, pp. 12–19, 2017.
https://doi.org/10.1007/978-3-319-68445-1_2

In particular, the question of asymptotic consistency and normality of principal nested spheres analysis [14], say, goes beyond generalized Fréchet means analysis. In all *nested* schemes, several estimators are determined sequentially, where each estimation depends on all previous ones. Recently, asymptotic consistency of *nested generalized Fréchet means* was introduced in [13], as a generalization of classical PCA's asymptotics, e.g. by [1], where nestedness of approximating subspaces is not an issue because it is trivially given.

Based on asymptotic consistency of nested and non-nested descriptors, hypothesis tests, like the two-sample test can be considered. Since by construction, every sample determines only one single descriptor and not its distribution, resampling techniques like the bootstrap are necessary to produce confidence sets. Notably, this is a very generic technique independent of specific sample spaces and descriptors. In the following, after introducing non-nested and nested generalized Fréchet means, we will elaborate on bootstrapping quantiles for a two-sample test. We will show that a *separated* approach in general leads to greatly increased power of the test in comparison to a *pooled* approach, both with correct asymptotic size. Also, we illustrate the benefit of *nested* over non-nested descriptors.

2 Descriptors for Manifold Valued Data

2.1 Single Descriptors

With a silently underlying probability space $(\Omega, \mathfrak{A}, \mathbb{P})$, *random elements* on a topological space Q are mappings $X : \Omega \to Q$ that are measurable with respect to the Borel σ-algebra of Q.

For a topological space Q we say that a continuous function $d : Q \times Q \to [0, \infty)$ is a *loss function* if $d(q, q') = 0$ if and only if $q = q'$.

Definition 1 (Generalized Fréchet Means [11]). *Let Q be a separable topological space, called the* data space*, and P a separable topological space, called the* descriptor space*, with loss function $d : P \times P \to [0, \infty)$ and a continuous map $\rho : Q \times P \to [0, \infty)$. Random elements $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$ on Q give rise to* population *and* sample descriptors

$$\mu \in \operatorname*{argmin}_{p \in P} \mathbb{E}[\rho(X, p)^2], \qquad \mu_n \in \operatorname*{argmin}_{p \in P} \sum_{j=1}^{n} \rho(X_j, p)^2.$$

The descriptors are also called generalized ρ-Fréchet means*. The sample descriptor is a* least squares M-estimator*.*

Asymptotic theory for generalized ρ-Fréchet means under additional assumptions, among them that the means be unique and attained on a twice differentiable manifold part of P has been established by [11,12].

2.2 Nested Descriptors

For nested descriptors, we need to establish a notion of nestedness and the relations between the successive descriptor spaces.

Definition 2 ([13]). *A separable topological data space Q admits* backward nested families of descriptors *(BNFDs) if*

(i) there is a collection P_j $(j = 0, \ldots, m)$ of topological separable spaces with loss functions $d_j : P_j \times P_j \to [0, \infty)$;
(ii) $P_m = \{Q\}$;
(iii) every $p \in P_j$ $(j = 1, \ldots, m)$ is itself a topological space and gives rise to a topological space $\emptyset \neq S_p \subset P_{j-1}$ which comes with a continuous map

$$\rho_p : p \times S_p \to [0, \infty);$$

(iv) for every pair $p \in P_j$ $(j = 1, \ldots, m)$ and $s \in S_p$ there is a measurable projection *map*

$$\pi_{p,s} : p \to s.$$

For $j \in \{1, \ldots, m-2\}$ call a family

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

a backward nested family of descriptors (BNFD) ending in P_j*, where we ignore the unique $p^m = Q \in P_m$. The space of all BNFDs ending in P_j is given by*

$$T_j = \left\{ f = \{p^k\}_{k=j}^{m-1} : p^{k-1} \in S_{p^k}, k = j+1, \ldots, m \right\} \subseteq \prod_{k=j}^{m-1} P_k.$$

For $j \in \{1, \ldots, m\}$, given a BNFD $f = \{p^k\}_{k=j}^{m-1}$ set

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

which projects along each descriptor. For another BNFD $f' = \{p'^k\}_{k=j}^{m-1} \in T_j$ set

$$d^j(f, f') = \sqrt{\sum_{k=j}^{m-1} d_k(p^k, p'^k)^2}.$$

Building on this notion, we can now define nested population and sample descriptors similar to Definition 1.

Definition 3 (Nested Generalized Fréchet Means [13]). *Random elements $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$ on a data space Q admitting BNFDs give rise to* backward nested population *and* sample descriptors *(abbreviated as BN descriptors)*

$$\{E^{f^j} : j = m-1, \ldots, 0\}, \quad \{E_n^{f_n^j} : j = m-1, \ldots, 0\}$$

recursively defined using $p^m = Q = p_n^m$ *via*

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

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

where $p^j \in E^{f^j}$ *and* $p_n^j \in E^{f_n^j}$ *is a measurable choice for* $j = 1, \ldots, m-1$.

We say that a BNFD $f = \{p^k\}_{k=0}^{m-1}$ *gives* unique *BN population descriptors if* $E^{f^j} = \{p^j\}$ *with* $f^j = \{p^k\}_{k=j}^{m-1}$ *for all* $j = 0, \ldots, m-1$.

Each of the E^{f^j} *and* $E_n^{f_n^j}$ *is called a* nested generalized Fréchet mean *and* $E_n^{f_n^j}$ *can be viewed as* nested least squares M-estimator.

Asymptotic theory for such backward nested families of descriptors, again under additional assumptions, among them being assumed on twice-differentiable manifold parts, has been established in [13].

In order to asses asymptotics of single elements in a family of nested generalized ρ-Fréchet means, the last element, say, a key ingredient is the following definition from [13].

Definition 4 (Factoring Charts [13]). *Let* $W \subset T_j$, $U \subset P^j$ *open subsets with* C^2 *manifold structure,* $f' = (p'^{m-1}, \ldots, p'^j) \in W$ *and* $p'^j \in U$, *and with local chart*

$$\psi : W \to \psi(W) \subset \mathbb{R}^{\dim(W)}, \qquad f = (p^{m-1}, \ldots, p^j) \mapsto \eta = (\theta, \xi)$$

the chart ψ factors, *if there is a chart* ϕ *and projections* π^U, $\pi^{\phi(U)}$

$$\phi : U \to \phi(U) \subset \mathbb{R}^{\dim(U)},\ p^j \mapsto \theta$$

$$\pi^U : W \to U,\ f \mapsto p^j, \qquad \pi^{\phi(U)} : \psi(W) \to \phi(U),\ (\theta, \xi) \mapsto \theta$$

such that the following diagram commutes

$$\begin{array}{ccc} W & \xrightarrow{\psi} & \psi(W) \\ \downarrow{\scriptstyle \pi^U} & & \downarrow{\scriptstyle \pi^{\phi(U)}} \\ U & \xrightarrow{\phi} & \phi(U) \end{array} \tag{1}$$

In case that factoring charts exist, from the asymptotics of an entire backward nested descriptor family it is possible to project to a chart, describing the last element descriptor only, and such a projection preserves asymptotic Gaussianity, cf. [13].

3 Bootstrap Testing

Based on the central limit theorems proved in [11,13], it is possible to introduce a T^2-like two-sample test for non-nested descriptors, BNFDs and single nested descriptors.

3.1 The Test Statistic

Suppose that we have two independent i.i.d. samples $X_1, \ldots, X_n \sim X \in Q$, $Y_1, \ldots, Y_m \sim Y \in Q$ in a data space Q admitting non-nested descriptors, BNFDs and single nested descriptors in P and we want to test

$$H_0 : X \sim Y \quad \text{versus} \quad H_1 : X \not\sim Y$$

using descriptors in $p \in P$. Here, $p \in P$ stands either for a single $p_k \in P_k$ or for a suitable sequence $f \in T_j$. We assume that the first sample gives rise to $\hat{p}_n^X \in P$, the second to $\hat{p}_m^Y \in P$, and that these are unique. We introduce shorthand notation to simplify the following complex expressions

$$d_{n,b}^{X,*} = \phi(\hat{p}_{n,b}^{X,*}) - \phi(\hat{p}_n^X) \qquad d_{m,b}^{Y,*} = \phi(\hat{p}_{m,b}^{Y,*}) - \phi(\hat{p}_m^Y)$$
$$\Sigma_{\phi,n}^{X,*} := \frac{1}{B}\sum_{b=1}^{B} d_{n,b}^{X,*} d_{n,b}^{X,*T} \qquad \Sigma_{\phi,m}^{Y,*} := \frac{1}{B}\sum_{b=1}^{B} d_{m,b}^{Y,*} d_{m,b}^{Y,*\,T}.$$

Define the statistic

$$T^2 := \left(\phi(\hat{p}_n^X) - \phi(\hat{p}_m^Y)\right)^T \left(\Sigma_{\phi,n}^{X,*} + \Sigma_{\phi,m}^{Y,*}\right)^{-1} \left(\phi(\hat{p}_n^X) - \phi(\hat{p}_m^Y)\right). \tag{2}$$

Under H_0 and the assumptions of the CLTs shown in [11,13], this is asymptotically Hotelling T^2 distributed if the corresponding bootstrapped covariance matrices exist. Notably, under slightly stronger regularity assumptions, which are needed for the bootstrap, this estimator is asymptotically consistent, cf. [5, Corollary 1].

3.2 Pooled Bootstrapped Quantiles

Since the test statistic (2) is only asymptotically T^2 distributed and especially deeply nested estimators may have sizable bias for finite sample size, it can be advantageous to use the bootstrap to simulate quantiles, whose covering rate usually has better convergence properties, cf. [7]. A pooled approach to simulated quantiles runs as follows. From $X_1, \ldots, X_n, Y_1 \ldots, Y_m$, sample $Z_{1,b}, \ldots, Z_{n+m,b}$ and compute the corresponding T_b^{*2} $(b = 1, \ldots, B)$ following (2) from $X_{i,b}^* = Z_{i,b}, Y_{j,b}^* = Z_{n+j,b}$ $(i = 1, \ldots, n,\ j = 1, \ldots, m)$. From these, for a given level $\alpha \in (0, 1)$ we compute the empirical quantile $c_{1-\alpha}^*$ such that

$$\mathbb{P}\left\{T^{*2} \leq c_{1-\alpha}^* | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\} = 1 - \alpha.$$

We have then under H_0 that $c_{1-\alpha}^*$ gives an asymptotic coverage of $1-\alpha$ for T^2, i. e. $\mathbb{P}\{T^2 \leq c_{1-\alpha}^*\} \to 1-\alpha$ as $n, m \to \infty$ if $n/m \to c$ with a fixed $c \in (0, \infty)$. Under H_1, however, the bootstrap samples $X_{i,b}^*$ and $Y_{j,b}^*$ have substantially higher variance than both the original X_i and Y_j. This leads to a large spread between the values of the quantiles and thus to diminished power of the test. This will be exemplified in the simulations below.

3.3 Separated Bootstrapped Quantiles

To improve the power of the test while still achieving the asymptotic size, we simulate a slightly changed statistic under H_0, by again bootstrapping, but now separately, from $X_1, \ldots, X_n$ and $Y_1 \ldots, Y_m$ (for $b = 1, \ldots, B$),

$$T^{*2} = \left(d_{n,b}^{X,*} - d_{m,b}^{Y,*}\right)^T \left(\Sigma_{\phi,n}^{X,*} + \Sigma_{\phi,m}^{Y,*}\right)^{-1} \left(d_{n,b}^{X,*} - d_{m,b}^{Y,*}\right). \tag{3}$$

From these values, for a given level $\alpha \in (0,1)$ we compute the empirical quantile $c_{1-\alpha}^*$ such that

$$\mathbb{P}\left\{T^{*2}(A) \leq c_{1-\alpha}^* | X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\} = 1 - \alpha.$$

Then, in consequence of [2, Theorems 3.2 and 3.5], asymptotic normality of $\sqrt{n}\big(\big(\phi(\hat{p}_n^X) - \phi(\hat{p}^X)\big)\big)$, and $\sqrt{m}\big(\big(\phi(\hat{p}_m^Y) - \phi(\hat{p}^Y)\big)\big)$, guaranteed by the CLT in [13], extends to the same asymptotic normality for $\sqrt{n}\, d_{n\,b}^{X\,*}$, and $\sqrt{m}\, d_{m\,b}^{Y\,*}$, respectively. We have then under H_0 that $c_{1-\alpha}^*$ gives an asymptotic coverage of $1-\alpha$ for T^2 from Eq. (2), i. e. $\mathbb{P}\{T^{*2} \leq c_{1-\alpha}^*\} \to 1-\alpha$ as $B, n, m \to \infty$ if $n/m \to c$ with a fixed $c \in (0, \infty)$.

We note that also the argument from [3, Corollary 2.3 and Remark 2.6] extends at once to our setup, as we assume that the corresponding population covariance matrix Σ_ψ or Σ_ϕ, respectively, is invertible.

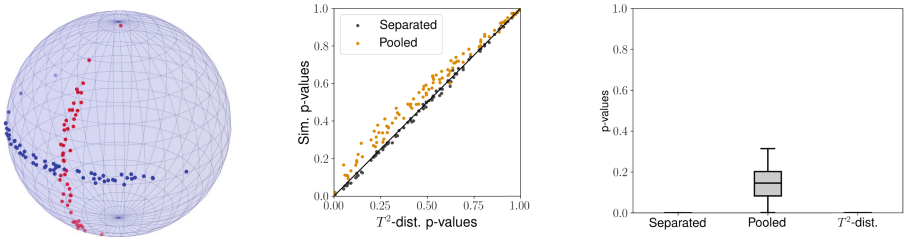
4 Simulations

We perform simulations to illustrate two important points. For our simulations we use the nested descriptors of Principal Nested Great Spheres (PNGS) analysis [14] and the intrinsic Fréchet mean [3]. In all tests and simulated quantiles we use $B = 1000$ bootstrap samples for each data set.

4.1 Differences Between Pooled and Separated Bootstrap

The first simulated example uses the nested mean and first geodesic principal component (GPC) to compare the two different bootstrapped quantiles with T^2-distribution quantiles in order to illustrate the benefits provided by separated quantiles. The two data sets we use are concentrated along two great circle arcs on an $\mathbb{S}^2$ which are perpendicular to each other. The data sets are normally distributed along these clearly different great circles with common nested mean and have sample size of 60 and 50 points, respectively, cf. Fig. 1a.

We simulate 100 samples from the two distributions and compare the p-values for the different quantiles. By design, we expect a roughly uniform distribution of p-values for the nested mean, indicating correct size of the test, and a clear rejection of the null for the first GPC, showing the power of the test. Both is satisfied for the separated quantiles and T^2-quantiles but not for the pooled quantiles, leading to diminished power under the alternative, cf. Fig. 1c. Under closer inspection, Fig. 1b shows that separated quantile p-values are closer to T^2-quantile p-values than pooled quantile p-values, which are systematically higher due to the different covariance structures rendering the test too conservative.



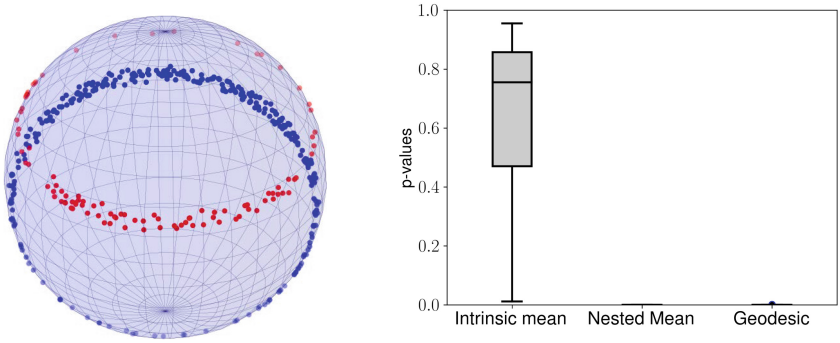
(a) Data set I (b) p-values for nested mean (c) p-values for first GPC

Fig. 1. Simulated data set I on $\mathbb{S}^2$ (a) with correct size under the null hypothesis of equal nested means (b) and power under the alternative of different first GPCs (c). The red sample has 50 points, the blue 60 points; we use p-values for 100 simulations each. (Color figure online)

4.2 Nested Descriptors May Outperform Non-nested Descriptors

The second point we highlight is that the nested mean of PNGS analysis is generically much closer to the data than the ordinary intrinsic mean and can thus, in specific situations, be more suitable to distinguish two populations. The same may also hold true for other nested estimators in comparison with their non-nested kin. The data set II considered here provides an example for such a situation. It consists of two samples of 300 and 100 points, respectively, on an $\mathbb{S}^2$ with coinciding intrinsic mean but different nested mean.

Here we only consider separated simulated quantiles, for both nested and intrinsic means. For the intrinsic mean two-sample test, we also use the bootstrap to estimate covariances for simplicity as outlined by [3], although closed forms for variance estimates exist, cf. [9]. Data set II and the distribution of resulting



(a) Data set II (b) p-values for data set II

Fig. 2. Simulated data set II (red: 100 points, blue: 300 points) on $\mathbb{S}^2$ (left), and box plots displaying the distribution of 100 p-values for PNGS nested mean and intrinsic mean (right) from the two-sample test. (Color figure online)

p-values are displayed in Fig. 2. These values are in perfect agreement with the intuition guiding the design of the data showing that the nested mean is suited to distinguish the data sets where the intrinsic mean fails to do so.

References

1. Anderson, T.: Asymptotic theory for principal component analysis. Ann. Math. Stat. **34**(1), 122–148 (1963)
2. Arcones, M.A., Giné, E.: On the bootstrap of m-estimators and other statistical functionals. In: LePage, R., Billard, L. (eds.) Exploring the Limits of Bootstrap, pp. 13–47. Wiley (1992)
3. Bhattacharya, R.N., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds II. Ann. Stat. **33**(3), 1225–1259 (2005)
4. Billera, L., Holmes, S., Vogtmann, K.: Geometry of the space of phylogenetic trees. Adv. Appl. Math. **27**(4), 733–767 (2001)
5. Cheng, G.: Moment consistency of the exchangeably weighted bootstrap for semiparametric m-estimation. Scand. J. Stat. **42**(3), 665–684 (2015)
6. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, Chichester (1998)
7. Fisher, N.I., Hall, P., Jing, B.Y., Wood, A.T.: Improved pivotal methods for constructing confidence regions with directional data. J. Am. Stat. Assoc. **91**(435), 1062–1070 (1996)
8. Hendriks, H., Landsman, Z.: Asymptotic behaviour of sample mean location for manifolds. Stat. Probab. Lett. **26**, 169–178 (1996)
9. Huckemann, S., Hotz, T., Munk, A.: Intrinsic MANOVA for Riemannian manifolds with an application to Kendall's space of planar shapes. IEEE Trans. Pattern Anal. Mach. Intell. **32**(4), 593–603 (2010)
10. 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)
11. 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)
12. Huckemann, S.: Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. Ann. Stat. **39**(2), 1098–1124 (2011)
13. Huckemann, S.F., Eltzner, B.: Backward nested descriptors asymptotics with inference on stem cell differentiation (2017). arXiv:1609.00814
14. Jung, S., Dryden, I.L., Marron, J.S.: Analysis of principal nested spheres. Biometrika **99**(3), 551–568 (2012)
15. Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley, New York (2000)
16. Nye, T., Tang, X., d Weyenberg, G., Yoshida, R.: Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. arXiv:1609.03045 (2016)
17. Pennec, X.: Barycentric subspace analysis on manifolds. arXiv preprint arXiv:1607.02833 (2016)
18. Schulz, J., Jung, S., Huckemann, S., Pierrynowski, M., Marron, J., Pizer, S.M.: Analysis of rotational deformations from directional data. J. Comput. Graph. Stat. **24**(2), 539–560 (2015)