(Semi-)Intrinsic Statistical Analysis on Stratified Spaces

Stephan F. Huckemann*
University of Göttingen, Germany
huckeman@math.uni-goettingen.de

Abstract

We present some applications from biology and medical imaging which lead to data on manifolds and stratified spaces. On such spaces the Euclidean concept of a mean forks into several canonical generalizations of non- Euclidean means. More involved data descriptors, for instance principal components generalize into even more complicated concepts. (Semi)-intrinsic statistical analysis allows to study inference on descriptors that can be represented as elements of another stratified space. We give examples for geodesic principal components on shape spaces and concentric small circles on spheres. In particular, with respect to the statistical inference via central limit theorems, due to the geometry of the spaces, we find curious non-Euclidean phenomena.

Key words and phrases: Shape spaces, Fréchet ρ-means, mean geodesics, limit theorems

1 Introduction

Many questions concerning dynamics of biological objects, due to natural constraining conditions, equivalence relations and/or identifications, lead to data of statistical interest which come to lie on spaces with no Euclidean structure. Often, these spaces can be equipped with structures of non-flat Riemannian manifolds or with structures of stratified spaces made up from manifolds of varying dimensions. Exemplary in this paper we consider dynamics caused by growth and by simple rotational deformations. In order to describe such dynamics, suitable descriptors are introduced which themselves live in stratified spaces.

(*Semi-*)*intrinsic statistical analysis* considers data on a stratified space Q, links these data via a *linking distance* ρ to Fréchet ρ-means in another stratified space P and conducts statistical analysis on P. This analysis is *semi-intrinsic* if for the final statistical analysis, e.g. for reasons of modeling and computational simplicity, extrinsic methods rather than intrinsic ones are used.

Example 1. Consider the problem of studying the leaf growth conducted in a joint research with the Department of Oecological Informatics, Biometry and Forest Growth at the University of Göttingen where the task is to describe growth and discriminate growth among leaves of the same tree of identical clones and over different genetic expressions. For this scenario it turns out that geodesics in underlying shape spaces qualify well as such descriptors, cf. Huckemann (2011b). Two such shape spaces come to mind. If only specific landmarks are taken into account, Kendall's planar shape spaces based on landmarks (cf. Dryden and Mardia (1998)) seem canonical.

Example 2. In a joint research with the Departments of Computer Science and Statistics at the Universities of North Carolina and Pittsburgh we consider the problem of estimating deformations of internal organs which is crucial for instance in radiation oncology, where one challenge consists in relocating an internal organ at therapy time while this organ has been carefully investigated at planning time. Frequently, this relocation involves not only the estimation of Euclidean motions

but also the estimation of deformations, e.g. a cancerous prostate might be bent around a filled bladder. For this task as an underlying model, *skeletal representations* introduced by Pizer et al. (2013) seem appropriate. The fundamental deformations due to rotations, bendings and twistings cause so called *spokes directions* to move on specific concentric small circles with a common axis for every fundamental rotational deformation. Thus the spaces of concentric small circles on the two-sphere qualify as descriptors of fundamental rotational deformations, cf. Schulz et al. (2012).

Example 3. The study of the third example originated from the AOOD Workshop 2010/11 at the Statistical and Applied Mathematical Sciences Institute (SAMSI) in North Carolina, cf. Skwerer et al. (2013). In their seminal paper Fitch and Margoliash (1967) introduced *phylogenetic trees* to assess genetic mutation distances. From the mutations of a specific gene over a fixed set of different species, an optimal descendant tree is estimated. Usually, different genes over the same set of species give different trees. Measuring tree distances by comparing interior edge lengths (corresponding to mutation distance) over equivalence classes of trees leads to a metric space which can be given the structure of a stratified space with flat manifold strata (orthants of varying dimensions) such that the entire space carries non-positive curvature. Similar models are appropriate to model biological tree-like structures such as the artery tree of the brain or of the respiratory tract.

2 (Semi)-Intrinsic Statistical Analysis

We begin with two topological spaces: Q is the *data* space and P is the *descriptor space*. Data and descriptors are linked via a continuous function $\rho : Q \times P \to [0, \infty)$ called a *linking function* which takes the role of a distance between a datum and a descriptor. Moreover, we assume that there is a continuous mapping $d : P \times P \to [0, \infty)$ vanishing on the diagonal $\{(p, p) : p \in P\}$.

Definition 2.1. *We call such a tuple* (ρ, d) *a* uniform link *if for every* $p \in P$ *and* $\epsilon > 0$ *there is a* $\delta = \delta(\epsilon, p) > 0$ *such that*

$$|\rho(x, p') - \rho(x, p)| < \epsilon \text{ for all } x \in Q, p' \in P \text{ with } d(p, p') < \delta.$$

Moreover, it is a coercive link *if for every* $p_0 \in P$, $C > 0$ *and sequence* $p_n \in P$ *with* $d(p_0, p_n) \to \infty$ *there is a sequence* $M_n \to \infty$ *with* $\rho(x, p_n) > M_n$ *for all* $x \in Q$ *with* $\rho(x, p_0) < C$*; and if* $p_n \in P$ *with* $d(p^*, p_n) \to \infty$ *for some* $p^* \in M$*, then* $d(p_0, p_n) \to \infty$.

In case of $P = Q$ and $\rho = d$ satisfying the triangle inequality, it is a uniform coercive link. More generally if P and Q are compact and d is a quasimetric we have a uniform coercive link.

Definition 2.2. *For random elements* $X, X_1, X_2, \ldots$ *on* Q *define the* set of population Fréchet ρ-means of X on P *by*

$$E^{(\rho)}(X) = \operatorname*{argmin}_{\mu \in P} \mathbb{E}\big(\rho(X, \mu)^2\big).$$

For $\omega \in \Omega$ *denote the* set of sample Fréchet ρ-means on P *by*

$$E_n^{(\rho)}(\omega) = \operatorname*{argmin}_{\mu \in P} \sum_{j=1}^{n} \rho\big(X_j(\omega), \mu\big)^2.$$

Means with a metric $\rho = d$ on $P = Q$ have been considered by Fréchet (1948) and generalized to quasimetrics by Ziezold (1977). Examples of Fréchet ρ-means are Procustes means on shape spaces (cf. Dryden and Mardia (1998)) and intrinsic and extrinsic means on manifolds (cf. Bhattacharya and Patrangenaru (2003, 2005)) or on stratified spaces (cf. Hotz et al. (2012)).

Geodesic principal components. On a metric space (Q, τ) a rectifiable curve $\gamma : I \to Q$, $I \subset \mathbb{R}$, is called a *geodesic* if (i): for all $s, t \in I$ with $s < t$, $p = \gamma(s)$ and $q = \gamma(t)$ we have that

$$\inf_{\substack{n\in\mathbb{N}\\ t=t_0<\ldots<t_n=s}} \sum_{j=1}^{n} \tau\big(\gamma(t_{j-1}),\gamma(t_j)\big) = \tau(p,q)\,,$$

(ii): if I has one or two endpoints, γ cannot be extended beyond such that it satisfies (i).

Denote by P the space of geodesics on Q. If P can be given a topological structure such that

$$\rho : Q\times P\to[0,\infty),\quad (q,\gamma)\mapsto \rho(q,\gamma) := \inf_{t\in I}\tau\big(p,\gamma(t)\big)$$

is a linking function, the corresponding Fréchet ρ-means are called *first geodesic principal components* (1stGPCs). If there is a concept of orthogonality of geodesics, higher order principal components can be defined as in Huckemann et al. (2010).

Concentric small circles. For spoke data $z=(z_1,\ldots,z_k)\in(S^2)^k$ as in Example 2, here $S^2:=\{x\in\mathbb{R}^3:\|x\|=1\}$ is the two-sphere and $k\in\mathbb{N}$, consider the space P of k concentric small circles with the straightforward quotient topology induced by the *Ziezold distance* (the quotient distance inherited from the extrinsic distance on $S^2\times[0,\pi]^k$, cf. Huckemann (2012)) as follows. With the usual Euclidean inner product $\langle\cdot,\cdot\rangle$, let

$$\begin{aligned}
\delta(c,r) &:= \big\{z\in(S^2)^k : \langle c,z_j\rangle = r_j, j=1,\ldots,k\big\} \text{ for } c\in S^2 \text{ and } r=(r_1,\ldots,r_k)\in[0,\pi]^k\,,\\
[\delta(c,r)] &:= \{\delta(c,r),\delta(-c,\pi-r)\} \text{ where } \pi-r:=(\pi-r_1,\ldots,\pi-r_k) \text{ and}\\
P &:= \big\{[\delta(c,r)] : c\in S^2, r\in[0,\pi]^k\big\}\,.
\end{aligned}$$

A linking function is given by the geodesic distance, cf. Schulz et al. (2012),

$$\rho\big(z,[\delta(c,r)]\big) := \sqrt{\sum_{j=1}^{k}\big(\arccos(\langle c,z_j\rangle)-r_j\big)^2}\,.$$

Definition 2.3. *Let $E_n^{(\rho)}(\omega), E^{(\rho)}\subset P$ be a random ($\omega\in\Omega$) and a deterministic closed set, resp., then,*

(ZC) $E_n^{(\rho)}(\omega)$ is a strongly consistent estimator *of $E^{(\rho)}$ in the sense of* Ziezold *if a.s. for $\omega\in\Omega$*

$$\bigcap_{n=1}^{\infty}\overline{\bigcup_{k=n}^{\infty}E_k^{(\rho)}(\omega)}\subset E^{(\rho)}\,,$$

(BPC) $E_n^{(\rho)}(\omega)$ is a strongly consistent estimator *of $E^{(\rho)}$ in the sense of* Bhattacharya-Patrangenaru *if $E^{(\rho)}\neq\emptyset$ and if for every $\epsilon>0$ and a.s. for $\omega\in\Omega$ there is $n=n(\epsilon,\omega)\in\mathbb{N}$ such that*

$$\bigcup_{k=n}^{\infty}E_k^{(\rho)}(\omega)\subset\{p\in P : d(E^{(\rho)},p)\le\epsilon\}\,.$$

Ziezold (1977) introduced (ZC) and proved it for quasi-metrical means on separable spaces. Bhattacharya and Patrangenaru (2003) introduced (BPC) and proved it for metrical means on spaces that enjoy the stronger d-Heine-Borel property, i.e. that every d-bounded (A is d-bounded if there is a point $p\in A$ such that $d(p,p_n)$ is bounded for every sequence $p_n\in A$) closed set is compact. Both properties, have been called 'strong consistency' by their respective authors. We have the following generalization, cf. (Huckemann, 2011b, Theorems A.3 and A.4).

Theorem 2.4. *Suppose that the data space Q is separable, ρ is a uniform link and that $\mathbb{E}\big(\rho(X,p)^2\big)<\infty$ for all $p\in P$. Then property (ZC) holds for the set of Fréchet ρ-means on P.*

If additionally $E^{(\rho)}\neq\emptyset$, P enjoys the d-Heine-Borel property and (ρ,d) is also a coercive link then property (PBC) holds for the set of Fréchet ρ-means on P.

In order to formulate a Gaussian central limit theorem, we require additional properties.

Assumption 2.5. *The population Fréchet-ρ mean μ is unique and there is an open neighborhood U of μ in P that is a D-dimensional twice differentiable manifold, $D \in \mathbb{N}$. Assume further that there is a local chart (ϕ, U) near $\mu = \phi^{-1}(0)$ such that*

$$\text{the mapping } x \mapsto \rho(X, \phi^{-1}(x))^2 \text{ is a.s. twice differentiable in } U \tag{1}$$

and with $\operatorname{grad}_2\rho(q,\nu)^2$ *and* $\operatorname{Hess}_2\rho(q,\nu)^2$ *denoting gradient and Hessian of the mapping that*

$$\left.\begin{array}{c}\mathbb{E}(\operatorname{grad}_2\rho(X,\mu)^2) \text{ and } \operatorname{Cov}(\operatorname{grad}_2\rho(X,\mu)^2), \text{ exist and} \\ \mathbb{E}(\operatorname{Hess}_2\rho(X,\nu)^2) \text{ exists for } \nu=\mu, \text{ is continuous at } \nu=\mu \text{ and of full rank there.}\end{array}\right\} \tag{2}$$

The following Theorem is a straightforward consequence of Huckemann (2011a, Theorem 6) which is a generalization of Bhattacharya and Patrangenaru (2005). On manifolds Kendall and Le (2011) have derived intrigued versions for independent but non-identically distributed samples.

Theorem 2.6. *Under Assumption 2.5 suppose that $E_n^{(\rho)}$ is a strongly consistent estimator of a Fréchet population ρ-mean set $E^{(\rho)} = \{\mu\}$ in the sense of Bhattacharya-Patrangenaru. Then for every measurable choice $\mu_n(\omega) \in E^{(\rho)}$*

$$\sqrt{n}\big(\phi(\mu_n) - \phi(\mu)\big) \stackrel{d}{\to} \mathcal{N}\big(0, A_\phi \operatorname{Cov}(\operatorname{grad}_2\rho(X,\mu)^2) A_\phi^{-1}\big) \text{ with } A_\phi = \mathbb{E}(\operatorname{Hess}_2\rho(X,\mu)^2)\,.$$

Remark 2.7. *The condition that A_ϕ be of full rank is not at all trivial. A consequence of A_ϕ failing to to so is discussed in Section 4.*

Two-sample tests for metrical means on shape spaces and on manifolds have been around for a while, e.g Dryden and Mardia (1998). In the following we extend these to our context of Fréchet ρ-means.

Suppose that $Y_1, \ldots, Y_n \stackrel{i.i.d}{\sim} Y$ and $Z_1, \ldots, Z_m \stackrel{i.i.d}{\sim} Z$ are independent samples on Q with unique Fréchet ρ-means μ_Y and μ_Z, respectively, on Q $(m, n > 0)$ and suppose X and Y are a.s. contained in $U \subset Q$, a twice differentiable Riemannian manifold with geodesic distance d. Under the null hypothesis we assume $\mu_Y = \mu_Z = \mu \in U$. For $p \in U$ let (ϕ_p, U) denote a chart with $\phi_p(p) = 0$. Moreover assume that ρ is uniform coercive and that Assumption 2.5 holds for a random variable X with $\mathbb{P}^X = \frac{n\mathbb{P}^Y + m\mathbb{P}^Z}{n+m}$ and that there is a constant $C > 0$ such that $\|\phi_p(X) - \phi_\mu(X)\|, \|\phi_p(Y) - \phi_\mu(Y)\| \le C d(p,\mu)$ a.s. for p near μ. Then the classical Hotelling T^2 statistic $T^2(n,m)$ of $\phi_{\mu_{n+m}}(Y_1), \ldots, \phi_{\mu_{n+m}}(Y_n)$ and $\phi_{\mu_{n+m}}(Z_1), \ldots, \phi_{\mu_{n+m}}(Z_m)$ is well defined for $n+m$ sufficiently large where μ_{n+m} denotes a measurable selection of a pooled Fréchet ρ-sample mean. In consequence of Theorem 2.6 we have the following, cf. Huckemann (2012, Theorem 10).

Theorem 2.8 (Two-Sample Test)**.** *Under the above hypotheses for $n, m \to \infty$, $T^2(n,m)$ is asymptotically Hotelling T^2-distributed if either $n/m \to 1$ or if $\operatorname{Cov}(\phi_\mu(X)) = \operatorname{Cov}(\phi_\mu(Y))$.*

3 Semi-Intrinsic Inference on the Mean Geodesic

The space of geodesics $\Gamma(\Sigma_m^k)$ of Kendall's shape space Σ_m^k of m-dimensional configurations with k landmarks $(m < k)$ can be given the following structure of a double quotient yielding a stratified space (cf. (Huckemann et al., 2010, Theorem 5.3))

$$\Gamma(\Sigma_m^k) = O(2)\backslash O^H(m, k-1)/SO(m)\,.$$

Here, the two dimensional orthogonal group $O(2)$ acts from the right and the m-dimensional special orthogonal group $SO(m)$ act from the left on the following submanifold of an orthogonal Stiefel manifold

$$O^H(m,k-1) \;:=\; \{(x,v) \in M(m,k-1)^2 : \langle x, v\rangle = 0, \langle x, x\rangle = 1 = \langle v, v\rangle, xv^T = vx^T\}$$

where $\langle x, y\rangle = \mathrm{trace}(xy^T)$ denotes the Euclidean inner product. The canonical embedding $O^H(m, k-1)$ in $\mathbb{R}^{m(k-1)} \times R^{m(k-1)}$ with the extrinsic metric leads to a Ziezold link ρ_Z on $\Gamma(\Sigma_m^k) \times \Gamma(\Sigma_m^k)$. We have the following extension of Huckemann (2011b, Theorem 3.1).

Theorem 3.1. *The following hold*

(i) The Ziezold link ρ_Z is a metric on $\Gamma(\Sigma_m^k)$.

(ii) There is an open and dense set $U \subset \Gamma(\Sigma_m^k)$ that carries the structure of a Riemannian manifold such that ρ_Z^2 is twice differentiable on $U \times U$ except at cut points.

(iii) $U = \Gamma(\Sigma_2^k)$ can be chosen in (ii) in case of $m = 2$.

Proof. Since the double action of the compact group $O(2) \times SO(m)$ can be extended isometrically onto $F_m^k \times F_m^k$ where $F_m^k := M(m, k-1) \setminus \{0\}$ with the extrinsic metric, the Principal Orbit Theorem (cf. Bredon (1972)) can be applied to

$$Q := O(2)\backslash\{(x, v) \in F_m^k \times F_m^k : xv^T = vx^T\}/SO(m)\,.$$

Thus ρ_Z extends to the geodesic metric on Q such that ρ_Z^2 is smooth except at cut points on an open and dense Riemannian manifold $\widetilde{U} \subset Q$. Since $\Gamma(\Sigma_m^k)$ is also a subspace of Q we have the assertions (i) and (ii). Assertion (iii) follows again from the Principal Orbit Theorem and the fact that the actions of $O(2)$ and $SO(m)$ commute and hence for $m = 2$ the isotropy groups are constant. □

In case of $m = 2$ the Ziezold metric can be computed explicitly. An example using the two sample test for inference on the mean geodesic of leaf growth as in Example 1 of the Introduction can be found in Huckemann (2011b).

4 Limit Theorems on Circles

Recall that the condition that $A_\phi = \mathbb{E}(\mathrm{Hess}_2\rho(X, \mu)^2)$ be of full rank is among the Assumptions 2.5 ensuring a Gaussian $\sqrt{n}$-CLT in Theorem 2.6. For the intrinsic metric on the circle it turns out that this condition is necessary. The other condition, local twice differentiability a.s. of $x \to \rho^2\big(X, \phi^{-1}(x)\big)$, is violated whenever X has a non-vanishing density near the cut locus of an intrinsic mean. On the circle this condition is not necessary, cf. (Hotz and Huckemann, 2011, Proposition 3.4 and Theorem 4.2). Here is a curious example for $A_\phi = 0$.

Example 4.1. *For a random variable X on the circle $[-\pi, \pi]$ with enpoints identified following a bimodal von Mises mixture density*

$$f(x) \;:=\; \frac{1}{I(a, b, \kappa, \tau)}\Big(a\, e^{\kappa\cos x} + b\, e^{-\tau\cos x}\Big) \text{ with } I(a, b, \kappa, \tau) \;:=\; \int_{-\pi}^{\pi} (a\, e^{\kappa\cos x} + b\, e^{-\tau\cos x})\, dx$$

and suitable $a, b, \kappa, \tau > 0$ such that there is a major mode at 0 of height $f(0) > (2\pi)^{-1}$ and minor mode at $-\pi$ of height $f(-\pi) = (2\pi)^{-1}$, we have that $\frac{1}{2}\mathbb{E}(\mathrm{Hess}_2\rho(X, 0)^2) = 1 - 2\pi f(-\pi) = 0$ while X features a unique intrinsic mean at 0 which is approached by sample means μ_n with a rate of

$$n^{1/6}\mu_n \xrightarrow{d} Y \text{ with } Y^3 \sim \mathcal{N}\left(0, \frac{9}{\pi^2}\,\frac{\int_{-\pi}^{\pi} x^2 f(x)\, dx}{(a\kappa e^{-\kappa} - b\tau e^{-\tau})^2}\right).$$

5 Outlook

In this survey we have briefly sketched and illustrated for some examples the concept of (semi)-intrinsic statistical analysis on stratified spaces. Although the theory is clear in its outline, many essential details still present challenging research topics, among those the effect of degeneracy of

$\mathbb{E}(\mathrm{Hess}_2\rho(X,0)^2)$ which may lead to arbitrary slow convergence rates on the circle. The opposite effect, namely that asymptotic rates may be much faster than $\sqrt{n}$ has been reported on open books being model spaces of tree spaces as described in the Introduction, cf. Hotz et al. (2012). A general theory for these non-Euclidean effects beyond simple model spaces is still open.

References

Bhattacharya, R. N. and V. Patrangenaru (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds I. *The Annals of Statistics 31*(1), 1–29.

Bhattacharya, R. N. and V. Patrangenaru (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds II. *The Annals of Statistics 33*(3), 1225–1259.

Bredon, G. E. (1972). *Introduction to Compact Transformation Groups*, Volume 46 of *Pure and Applied Mathematics*. New York: Academic Press.

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

Fitch, W. and E. Margoliash (1967). Construction of phylogenetic trees. *Science 155*(760), 279–284.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut de Henri Poincaré 10*(4), 215–310.

Hotz, T. and S. Huckemann (2011). Intrinsic means on the circle: Uniqueness, locus and asymptotics. *arXiv.org*, 1108.2141.

Hotz, T., S. Huckemann, H. Le, J. S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer (2012). Sticky central limit theorems on open books. *Annals of Applied Probability*. accepted.

Huckemann, S. (2011a). Inference on 3D Procrustes means: Tree boles growth, rank-deficient diffusion tensors and perturbation models. *Scandinavian Journal of Statistics 38*(3), 424–446.

Huckemann, S. (2011b). Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics 39*(2), 1098–1124.

Huckemann, S. (2012). On the meaning of mean shape: Manifold stability, locus and the two sample test. *Annals of the Institute of Mathematical Statistics 64*(6), 1227–1259.

Huckemann, S., T. Hotz, and A. Munk (2010). Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). *Statistica Sinica 20*(1), 1–100.

Kendall, W. S. and H. Le (2011). Limit theorems for empirical fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics 25*(3), 323–352.

Pizer, S. M., S. Jung, D. Goswami, X. Zhao, R. Chaudhuri, J. N. Damon, S. Huckemann, and J. Marron (2013). Nested sphere statistics of skeletal models. In *Proc. Dagstuhl Workshop on Innovations for Shape Analysis: Models and Algorithms*. to appear.

Schulz, J., S. Jung, S. Huckemann, J. Marron, and S. Pizer (2012). Analysis of rotational deformations from directional data. preprint.

Skwerer, S., E. Bullitt, S. Huckemann, E. Miller, I. Oguz, M. Owen, V. Patrangenaru, S. Provan, and J. Marron (2013). Tree-oriented analysis of brain artery structure. preprint.

Ziezold, H. (1977). Expected figures and a strong law of large numbers for random elements in quasi-metric spaces. *Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes A*, 591–602.