On Curvature and Limit Theorems

Stephan F. Huckemann[1]* and Thomas Hotz†

*Institute for Mathematical Stochastics, Georg-August Universität Göttingen
†Institute of Mathematics Ilmenau University of Technology

Abstract

For inference on non-Euclidean data, ideally the data space is at least locally approximated by a Euclidean space; for a manifold by a chart, say. If suitable data descriptors lie asymptotically in such a common Euclidean neighborhood, one would expect a normal central limit theorem to hold. Already for locally flat spaces like the circle there are exceptions causing smeariness. On stratified spaces, also singularities may prevent asymptotic normality. While for shape spaces, due to relative increased curvature, singularity sets are repulsive, for spaces with sudden drops in curvature, phenomena of stickiness arise.

1 Means on Stratified Spaces: Manifolds and Non-Manifolds

At least since the middle of the last century (cf. Fisher (1953)), statisticians have begun to investigate non-Euclidean data such as circular data (e.g. wind directions), directional data, orientational data and shape data to name only a few. On "nice" spaces such as the circle, spheres and even planar shape spaces (which are complex projective spaces) parametric statistical theory has provided for an ample collection of distributional models thus allowing for maximum likelihood estimation. In the sense that the parameters involved range over a subset of a Euclidean space, this methodology for non-Euclidean data can still be called Euclidean.

A different tack is persued by nonparametric statistical theory when it comes to estimating a non-Euclidean quantity such as a typical direction, orientation, shape, et cetera via Fréchet means (cf. Fréchet (1948)). Notably this methodology also carries over, when non-Euclidean spaces are no longer "nice", e.g. no longer symmetric spaces or even no longer manifolds as is the case for phylogenetic tree spaces (cf. Billera et al. (2001)) or Kendall's shape spaces of three- and higher-dimensional objects (cf. Kendall et al. (1999)). In principle, there are three types of Fréchet means: extrinsic, residual and intrinsic (for an overview cf. Huckemann (2011)). Extrinsic and residual means require an explicitly tractable embedding in a Euclidean space, which is often unavailable. By their very nature their asymptotics are assessed by a mixture of Euclidean and non-Euclidean methods. Intrinsic means, however, are available with every concept of a (geodesic) distance and the corresponding statistical methodology can be truly called non-Euclidean. In view of asymptotics of intrinsic Fréchet means, we focus here on

(a) the influence of mass near cut loci and

(b) the influence of singularities and discontinuities of the curvature.

In the following, let Q be a stratified space (e.g. Pflaum (2001)) with geodesic distance $d : Q \times Q \to [0, \infty)$ and top manifold stratum Q^* that is open and dense in Q. We assume that Q^* carries a Riemannian structure for which $d|_{Q^*}$ is the geodesic distance. By $\exp_q$ we denote the *Riemann exponential* centered at $q \in Q^*$ which is a diffeomorpism from an open set U_q

[1]supported by DFG HU 1575/2-1 and by the Felix Bernstein Institute for Mathematical Statistics in the Biosciences and the Niedersachsen Vorab of the Volkswagen Foundation

containing the origin in the tangent space T_qQ^* onto $\exp_q(U) \subset Q^*$. Then we have with the usual Euclidean norm $\|\cdot\|$ of T_qQ^* that $\|v\| = d\big(\exp_q(v), q\big)$ whenever $v \in U_q$. The *cut locus of* q is the image of the boundary of a maximal U_q if existing,

$$\mathcal{C}(q) \quad := \quad \Big\{\exp_q(t^*v) : v \in T_qQ^*, \|v\| = 1, t^* = \sup\{t > 0 : d(\exp_q(tv), q) = t\}\Big\}\,,$$

otherwise the cut locus is void, $\mathcal{C}(q) = \emptyset$. The latter is the case for non-positive curvature spaces. For compact manifolds cut loci are not void, e.g. for spheres the cut locus of a point is its antipodal.

Moreover, for random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ on Q ($n \in \mathbb{N}$) mapping from a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ define the *set of population Fréchet means of* X by

$$E(X) = \operatorname*{argmin}_{q \in Q} \mathbb{E}\big(d(X, q)^2\big)\,.$$

For $\omega \in \Omega$ denote the *set of sample Fréchet means* by

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

We say that

- X on Q is *regular* if $E(X)$ exists and $\mathbb{P}\{X \in Q^*\} > 0$,
- E is *manifold stable* if $E(X) \subset Q^*$ for all regular X, otherwise Q^0 is *hit* by E;
- Q^0 is *sticky* for E if there is a regular X such that for all compactly supported random variables Y on Q independent of X, there is $C_Y > 0$

$$E(Z) \cap Q^0 \neq \emptyset \quad \text{for all } \epsilon < C_Y \text{ and random variables } Z \sim \frac{\mathbb{P}^X + \epsilon\mathbb{P}^Y}{1+\epsilon} \text{ on } Q\,,$$

- X with unique mean $E(X) = \{\mu\} \in Q^*$ and non-void cut locus $\mathcal{C}(\mu)$ *omits the cut locus* if there is $\epsilon > 0$ such that

$$\mathbb{P}\Big\{d\big(X, \mathcal{C}(\mu)\big) > \epsilon\Big\} = 1\,.$$

2 Hitting and Sticking

Here we consider two scenarios. The first is that of a non-manifold *shape space* $Q = M/G$ that occurs from a complete finite dimensional Riemannian manifold M on which a Lie group G acts properly and isometrically where the shapes $\{[p] : gp, g \in G\}$ have varying dimension when considered as submanifolds of M. Here, we denote the action as from the left by $G \times M \to M, (g, p) \mapsto gp$. The manifold part Q^* is then a subset of all shapes of maximal dimension. For the detailed construction and how various Kendall's shape spaces such as Σ_m^k ($m > 2$) fit into this setup we refer to Bredon (1972); Kendall et al. (1999); Huckemann et al. (2010). As we have shown earlier (cf. Huckemann (2011, 2012)), on such spaces the set of Fréchet means is manifold stable and in consequence, there is no stickiness on shape spaces.

The second scenario is that of an *open book* which can be considered as a model space for the non-positive curvature tree space introduced by Billera et al. (2001). More precisely, near a co-dimension-1 singularity the tree space is locally an open book. The d-dimensional open book is defined as

$$Q = S \cup \bigcup_{j=1}^{k} H_j^+$$

with the *spine* $S = \mathbb{R}^{d-1} = Q^0$ and $k \in \mathbb{N}$ *leaves* $H_j^+ = \mathbb{R}^{d-1} \times (0, \infty)$, $(j = 1, \ldots, k)$ $d \in \mathbb{N}$ and $k > 2$. The topological identification is given by $S \sim R^{d-1} \times \{0\}$ and detailed in Hotz et al. (2012). For $j = 1, \ldots, k$ introduce the *folding maps*

$$F_j : Q \to \mathbb{R}^d, (x,t) \mapsto \begin{cases} (x,t) & \text{if } x \in H_j^+ \\ (x,-t) & \text{else} \end{cases}$$

and *folded moments*

$$m_j = \int_{\mathbb{R}^d} z \, d\mathbb{P}^{F_j \circ X}(z) \,.$$

Theorem 2.1 (Hotz et al. (2012))**.** *There is an index $j_0 \in \{1, \ldots, k\}$ such that $m_j < 0$ for all $j_0 \neq j \in \{1, \ldots, k\}$. Moreover, sample and population means are unique. If they are denoted by $\hat{\mu}_n$ and μ, respectively, with $Y_n = \sqrt{n}\big(F_{j_0}(\hat{\mu}_n) - F_{j_0}(\mu)\big)$, one of the following is true*

(i) $m_{j_0} > 0 \Leftrightarrow \mu \in H_{j_0}^+$ and the limiting distribution of Y_n is a Gaussian on $\mathbb{R}^{d-1} \times (0, \infty)$,

(ii) $m_{j_0} = 0 \Leftrightarrow$ the limiting distribution of Y_n is supported on $\mathbb{R}^{d-1} \times [0, \infty)$ assuming $\mathbb{R}^{d-1} \times \{0\}$ with positive probability,

(iii) $m_{j_0} < 0 \Leftrightarrow$ the limiting distribution for Y_n is a Gaussian on $\mathbb{R}^{d-1} \times \{0\}$.

In case (iii) above, X sticks to the spine S. In particular, for suitable random $N \in \mathbb{N}$ we have a.s. that $\hat{\mu}_n \in S$ for all $n \geq N$. In case (i), X is non-sticky as is always the case on Euclidean spaces.

3 Not Omitting the Cut Locus

As we have seen, on shape spaces (and of course on manifolds) there are no stickiness phenomena causing degeneracies of limiting distributions. Rather the opposite effect of *smeariness* can occur if there is mass near the cut locus of a mean. Let S^1 be the unit circle which we represent by $[-\pi, \pi)$ with the endpoints identified.

Theorem 3.1 (Hotz and Huckemann (2011))**.** *Let X be a random variable on the circle S^1 with unique intrinsic mean $\mu = 0$, $\mathbb{E}(X^2) = \sigma^2$ where X is viewed as taking values in $[-\pi, \pi)$ and suppose that $\hat{\mu}_n$ is an intrinsic sample mean. Then the following hold*

(i) $\mathbb{P}\{X = -\pi\} = 0$, i.e. there can be no point mass antipodal to an intrinsic mean,

(ii) if X restricted to some neighborhood of $-\pi$ features a continuous density f, then

(1) $f(-\pi) \leq \frac{1}{2\pi}$,

(2) if $f(-\pi) < \frac{1}{2\pi}$ then

$$\sqrt{n}\, \hat{\mu}_n \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{\big(1 - 2\pi f(-\pi)\big)^2}\right) ,$$

(3) if $f(-\pi) = \frac{1}{2\pi}$ suppose that there are $\delta > 0$ and $k \in \mathbb{N}$ such that f is $k-1$ times continuously differentiable in $(-\pi, \delta - \pi)$ and $(\pi - \delta, \pi)$ with $f^{(j)}(\pi-) = f^{(j)}(-\pi+) = 0$ for all $1 \leq j < k$, and that there are k-th order continuous directional derivatives with $0 \neq f^{(k)}(\pi-) = (-1)^k f^{(k)}(-\pi+) < \infty$; then

$$\sqrt{n}\,\operatorname{sign}(\hat{\mu}_n)\,|\hat{\mu}_n|^{k+1} \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2\big((k+1)!\big)^2}{\big(2\pi f^{(k)}(-\pi+)\big)^2}\right).$$

In case (ii), (3) above the limiting distribution of μ_n is k-th order *smeary* in the sense that

$$n^{\frac{1}{2(k+1)}}\,\hat{\mu}_n$$

has a non-trivial limiting distribution. In case (ii), (2), $\hat{\mu}_n$ is 0-th order smeary as is always the case on Euclidean spaces and in the non-sticky case on the open book.

References

Billera, L., Holmes, S., Vogtmann, K., 2001. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics 27 (4), 733-767.

Bredon, G. E., 1972. Introduction to Compact Transformation Groups. Vol. 46 of Pure and Applied Mathematics. Academic Press, New York.

Fisher, R., 1953. Dispersion on a sphere. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 217(1130), 295-305.

Fréchet, M., 1948. Les éléments aléatories de nature quelconque dans un espace distancié. Annales de l'Institue de Henri Poincaré 10 (4), 215-310.

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

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

Huckemann, S., 2011. Manifold stability and the central limit theorem for mean shape. In: Proceedings of the 30th LASR Workshop. Leeds University Press, pp. 99-103.

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

Kendall, D.G., Barden, D., Carne, T.K., Le, H., 1999. Shape and Shape Theory. Wiley, Chichester.

Pflaum, M., 2001. Analytic and Geometric Study of Stratified Spaces: Contributions to Analytic and Geometric Aspects. Vol. 1768. Springer Verlag.