

Data Analysis on Manifolds

Vic Patrangenu

Geometric Topological and Graphical Model Methods in Statistics
Fields Institute, Toronto, Canada, May 22-23, 2014

Thanks

Thanks Rudy Beran for his invaluable suggestions on Data Analysis on Manifolds (DAM)!

Thanks Peter Kim for his invitation and for his TDA and DAM leadership!

Thanks Ezra Miller for inviting me at SAMSI-LDHD and our inspirational joint work at SAMSI-AOD!

Thanks NSA-MSP-H98230-14-1-0135 and NSF-DMS-1106935!

Thanks to coauthors, WG colleagues and students for their dedication to advancing DAM to a new level, including: B. Afsari, V. Balan, A. Bandulasiri, R. N. Bhattacharya, M. Buibas, T. Chang, M. Crane, G. Derado, I. L. Dryden, L. A. Ellingson, D. Groisser, R. Guo, H. Hendricks, G. Heo, S. Huckemann, T. Hotz, P. T. Kim, H. Le, X. Liu, L. Lin, K. V. Mardia, J. S. Marron, E. Miller, W. Mio, H. G. Mueller, A. Munk, J. Nolen, D. E. Osborne, M. Owen, N. Oza, R. Paige, V. P. Patrangenaru, M. Qiu, F. H. Ruymgaart, P. San Valentin, A. Schwartzman, S. Skwerer, S. Sughatadasa, H. W. Thompson, A.T.A. Wood, K. D. Yao, H. T. Zhu

Historical Note

- ▶ My original sources for DAM were a combo between applied directional data analysis (Fisher, Hall, Jing and Wood (1996)), shape data analysis (Dryden and Mardia (1992)) and theoretical Statistical Manifolds ideas (Efron(1975, 1978))
- ▶ DAM started in the nineties (*extrinsic sample means and function estimation on submanifolds*)
- ▶ continued with the *Analysis of Object Data* program at SAMSI in 2010 - 2011, in WG's such as "Geometric Correspondence", "Trees" and "Data Analysis on Sample Spaces with a Manifold Stratification"
- ▶ advanced in 2012 - MBI Workshop of *Statistics, Geometry, and Combinatorics on Stratified Spaces Arising from Biological Problems*, and a meeting in Denmark on *Geometry and Statistics*
- ▶ added methodologies due to progress in WG's of the 2013-2014 LDHD - SAMSI program including "Data Analysis on Hilbert Manifolds and Their Applications" and "Nonlinear Low-dimensional Structures in High-dimensions for Biological Data"

Why Should Statisticians Care about Manifolds?

See <http://www.fields.utoronto.ca/programs/scientific/13-14/modelmethods/abstracts.html>

- ▶ numerical data
- ▶ astronomy and cosmology data: galaxies, stars and planetary orbits data
- ▶ Spatial Statistics : temperatures, snow and other functions measured across the planet
- ▶ vector fields of wind velocities on the Earth surface
- ▶ Geology : paleomagnetic data, plate tectonics, volcanos
- ▶ Morphometric data
- ▶ protein and DNA structures
- ▶ medical imaging outputs, including : angiography data (such as brain vessels structure), CT, MRI
- ▶ satellite or aerial imaging
- ▶ digital camera imaging data

Data Collection and Sample Space Assumptions

We experience life, through senses, reflections and thoughts about Mother Nature.

Most of the time Statisticians, under the influence of Mathematical Analysis, may get fairly narrow or even incorrect scientific views through simulations with nonexistent data. To get closer to the “truth”, we have to **collect and analyze data** by

- ▶ repeatedly, and, hopefully randomly, perform an experiment
- ▶ extract from an individual *outcome* information in the form of a point on a *sample space*, which is assumed to be a complete metric space, thus leading to **consistent estimates**
- ▶ bringing in a **common sense assumption** that the sample space has a certain **local dimension**, making it locally homeomorphic to a numerical space, or, worst case scenario, to an **open cone or open book** like structure.
- ▶ accept the idea of infinite dimensionality (Hilbert).

Analyzing Linearized and Studentized Cartan Means

Compute the **Cartan sample mean**, when unique, and use its asymptotic distribution around the **population Fréchet mean**. Using dimensionality, estimate sample mean second (or higher) moments, yielding a geometric sample covariance matrix, in the tangent space. In the finite D numeric case this generalizes the sample covariance $S_n = \frac{1}{n}((x_1 - \bar{x})(x_1 - \bar{x})^T + \cdots + (x_n - \bar{x})(x_n - \bar{x})^T)$, to Fréchet sample covariance.

Use the linearity in the *tangent cone*, and via $S_n^{-\frac{1}{2}}$ (or a regularized version of it) **studentize** the Cartan sample mean, bringing in the game asymptotic pivots that are tabulated, for improved estimation.

Large Sample Size Numerical Data Analysis

Assuming the *population mean vector* is μ and the sample size n is large.

- ▶ Studentize the sample mean vector to get the *vectorial z-scores* vector $z_n = \sqrt{n}S_n^{-\frac{1}{2}}(\bar{x} - \mu)$
- ▶ Compute the square distance to the zero vector $T^2 = z_n^T z_n$, also known as the **Hotelling T-square** statistic
- ▶ Classical results in **Large Sample Theory** due to H. Cramer, Slutsky, Feller, Billingsley a.o. show that T^2 has asymptotically a χ_p^2 distribution.
- ▶ given $\alpha \in (0, 1)$, construct a $100(1 - \alpha)\%$ **large sample confidence region** C_α for μ :

$$C_\alpha = \{\mu \in \mathbb{R}^p \mid T^2 \leq \chi_p^2(\alpha)\}, \quad (1)$$

where the probability for a χ_p^2 distributed random variable being greater than $\chi_p^2(\alpha)$ is α .

- ▶ Such confidence regions are bounded by $p - 1$ dimensional **ellipsoids** .

Nonparametric Bootstrap based Numerical Data Analysis

Assuming the sample size n is **small, say $n < 30 + p$** , repeatedly sample with replacement from the original sample, thus obtaining **bootstrap resamples** ${}_r X_1^*, \dots, {}_r X_n^*$, for $r = 1, \dots, N$. Typically $N \geq 5000$. or each resample ${}_r X^*$:

- ▶ compute the corresponding *bootstrap statistics*: the bootstrap sample mean vector ${}_r \bar{X}^*$, the bootstrap sample covariance matrix ${}_r S_n^* = \frac{1}{n}(({}_r X_1^* - {}_r \bar{X}^*)({}_r X_1^* - {}_r \bar{X}^*)^T + \dots + ({}_r X_n^* - {}_r \bar{X}^*)({}_r X_n^* - {}_r \bar{X}^*))$
- ▶ if ${}_r S_n^*$ has an inverse, compute the *bootstrap z-scores* vector ${}_r Z_n^* = \sqrt{n}({}_r S_n^*)^{-\frac{1}{2}}({}_r \bar{X}^* - \bar{X})$
- ▶ compute bootstrap T -square ${}_r T^{2*} = ({}_r Z_n^*)^T {}_r Z_n^*$

Given the **bootstrap distribution** ${}_r T^{2*}$, $r = 1, N$:

- ▶ given $\alpha \in (0, 1)$, compute the bootstrap $(1 - \alpha)\%$ percentile $\chi^{2*}(\alpha)$
- ▶ use **Efron's nonparametric bootstrap** to construct a $100(1 - \alpha)\%$ **bootstrap confidence region** C_α for μ :

$$C_\alpha^* = \{\mu \in \mathbb{R}^p \mid T^2 \leq \chi^{2*}(\alpha)\}. \quad (2)$$

Solar System Data Example

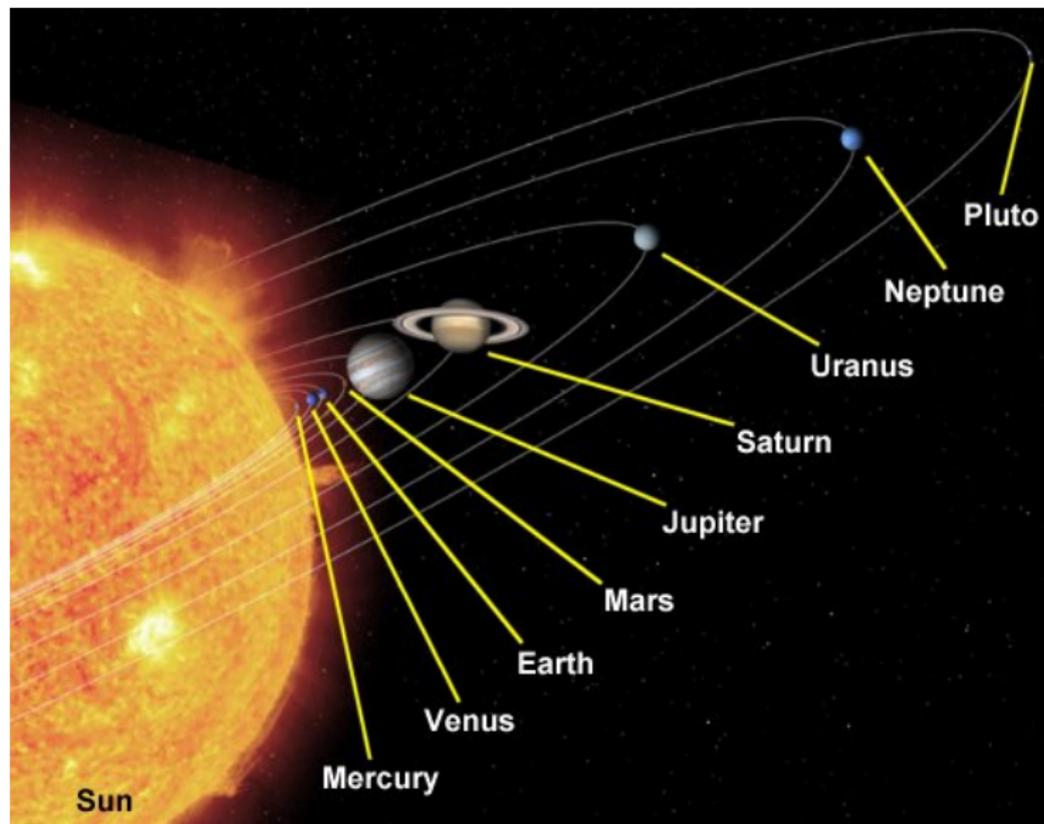


Figure: A pictorial representation of the Solar System.

Solar System - Data on a Sphere Example

For planets data, let i be the inclination of the orbital plane of a planet to the ecliptic and Ω be the angle between a fixed line in the ecliptic (the line joining the Sun and the Earth at the time of the vernal equinox), and the line joining the ascending node of the planet (the point where the orbit of the planet rises to the positive side of the ecliptic). Then each orbit determines one directed unit vector n , perpendicular to the orbital plane of the planet:

$$n = (\sin \Omega \sin i, -\cos \Omega \sin i, \cos i)$$

Table: The normals to the orbital planes of the nine planets

Planet	n_x	n_y	n_z
Mercury	0.001151	0.121864	0.99255
Venus	0.022170	-0.054694	0.99826
Earth	0.000000	0.000000	1.00000
Mars	0.032156	-0.002858	0.99948
Jupiter	0.020454	-0.010471	0.99974
Saturn	0.013473	-0.041487	0.99905
Uranus	0.012596	0.004514	0.99991
Neptune	0.029663	-0.009412	0.99952
Pluto	0.241063	0.170303	0.95545

Solar System Data Example - tangent plane representation

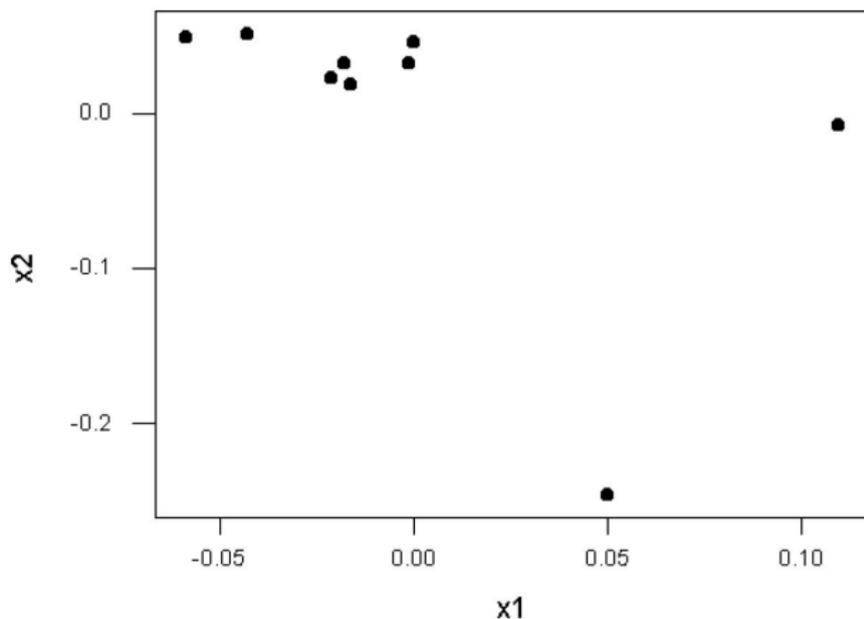


Figure: Unit normals to the orbital planes of the nine planets projected on a plane tangent at the sphere.

Solar System Data Analysis - Expected Results

Given a sample x_1, \dots, x_n of points on a sphere, their sample mean \bar{x} is located inside the sphere. The **extrinsic sample mean** is

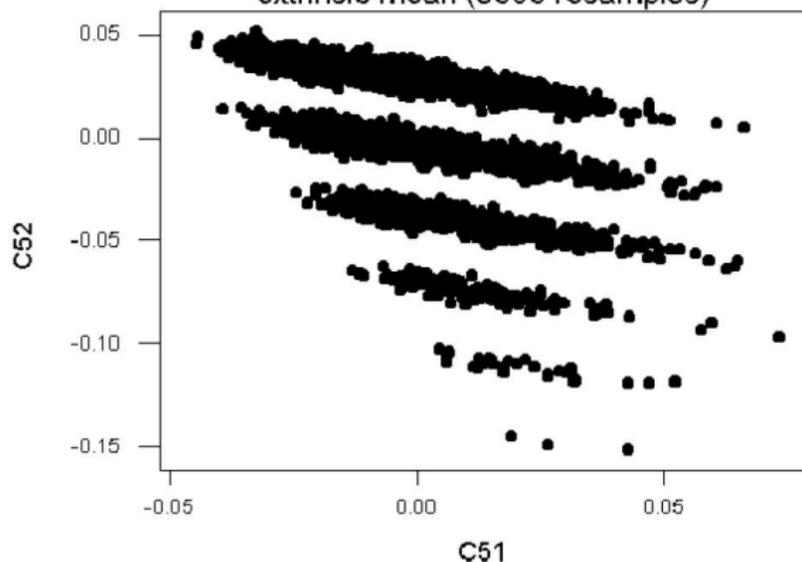
$$\bar{x}_E = \frac{\bar{x}}{\|\bar{x}\|} \quad (3)$$

We compute the **bootstrap distribution of the extrinsic sample means**. From the general theory presented above, the bootstrap distribution of the sample mean should look like a football shaped point cloud inside a sphere, that is concentrated around the population mean vector. If we are radially projecting this distribution on the sphere, we obtain the distribution of the extrinsic sample mean in the shape of a small elliptic cloud of points on the sphere around the extrinsic sample mean. Further if we project the bootstrap distribution of the extrinsic sample means on a **plane tangent** to the sphere at the extrinsic sample mean, the resulting distribution should appear in the shape of an ellipse centered around the null vector.

Solar System Data Analysis - Results

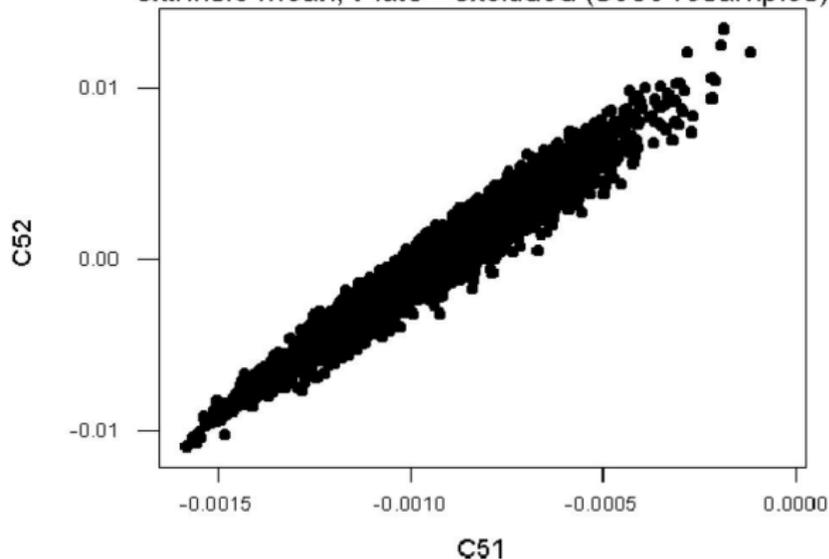
Surprisingly, what we get looks a bit different. We see more ellipses.

Boostrapped distribution of projections of normals to
orbital planes on the tangent plane at their
extrinsic mean (5000 resamples)



Solar System Data Analysis- Pluto Removed from the sample

Bootstrapped distribution of projections of normals to orbital planes on the tangent plane at their extrinsic mean, Pluto - excluded (5000 resamples)



Take home message: Pluto has a different origin than the other planets in the Solar system

Object Data Types and Data Analysis Methods

In general, besides classical multivariate analysis, so far statisticians have considered three types of noncategorical data:

- ▶ directional data, shape and image data
- ▶ trees and graphs data
- ▶ functional data and shape of contours data

All these were studied using certain methodologies, that are specific for each of these types, and there was little interaction between them.

These are

- ▶ Statistics on manifolds, or on orbifolds = spaces of orbits of group actions
- ▶ Statistics of folded Euclidean spaces
- ▶ Statistics on Hilbert manifolds

A unified methodology for Object Data Analysis

We suggest a unified methodology, by extending the type sample spaces considered to a generality that captures all the types of object data known so far under one umbrella:

Data Analysis on Sample Spaces that admit a (possibly infinitely dimensional) Manifold Stratification

Our approach is **nonparametric**, given the fairly complex structure of the sample space, it is still early to consider parametric families. The logic is to follow that classical ideas of first developing basic concepts of means, medians, variances and function estimation and understand the asymptotic behavior of the sample counterparts of indices for location and spread, aiming at nonparametric inference.

Data on Stratified Spaces

Beside multivariate data, the most elementary example of data on manifolds, other, less elementary example include directional data, shape data, medical imaging data including DTI, protein data, visual data for control of manufacturing processes, astronomy data, geology data, spatial and temporal data, pattern recognition data, DNA data, pretty much almost anything can be reduced to data analysis on spaces with a manifold stratification.

Example of Digital Camera Image Data



Figure: Pictures of Boston Marathon bombings suspect taken prior to April 2013



Figure: Picture of Boston Marathon bombings suspect taken on site on April 13, 2013

Basic Example of Nonparametric Image Data Analysis

The x-coordinates of four linearly associated landmarks were collected from each of the images in figures 3 and 4:

- ▶ x^1 -right corner of the right eye
- ▶ x^2 -left corner of the left eye
- ▶ x^3 -right corner of the left eye
- ▶ x^4 -left corner of the left eye

The **cross ratio** $k = \frac{x^1 - x^3}{x^2 - x^3} : \frac{x^1 - x^4}{x^2 - x^4}$, a projective invariant, was computed for each of these images. A bootstrap 90% c. i. of the mean of the cross-ratios of the test images in figure 3 was found (1.466714, 1.588492). This includes $k_0 = 1.544601$, the null hypothesis for the crossratio being from figure 4, thus we do not reject at .1 level the possibility that the suspect in figure 1. is one and the same with the individual in figure 2.

Acrosterigma Magnum Clamshell

See poster presented by Ruite Guo.

- ▶ One of the largest species from the Cardiidae bivalve family
- ▶ Large numbers washed ashore St George Island during Deepwater Horizon oil spill in May 2010
- ▶ Bilateral Symmetry implies that a shape data analysis can be performed based on any one of the two shells found
- ▶ one such live specimen is pictured in figure 21

Acrosterigma Magnum Clamshell



Compare Mean Reflection Shape Change

- ▶ Two samples: large shells and small shells
- ▶ Select landmarks consistently throughout the two samples
- ▶ Obtain Euclidean 3D similarity reconstructions using
- ▶ Compute Schoenberg (extrinsic) sample means
- ▶ Compare Schoenberg (extrinsic) population means
- ▶ The methodology is nonparametric nonpivotal bootstrap

Motivation - Binocular Vision

Without exception animals: insects, fish, mammals, reptiles, birds etc. have two eyes on their heads, to understand the surrounding 3D world.



Figure: Predatory animals capture a wide 3D scene in front of their eyes using stereopsis.

Emulating animal vision, in absence of occlusions, the 3D projective shape of a spatial scene can be recovered from a stereo pair of its images.

Basic Projective Geometry

- ▶ A point in the outer space and its central projection via the camera pinhole, determine a unique line in space, leading to the definition of the real projective plane $\mathbb{R}P^2$ as space of all straight lines going through the origin of \mathbb{R}^3 .
- ▶ Consider a real vector space V , and let 0_V be the zero of this vector space. Two vectors $x, y \in V \setminus \{0_V\}$ are equivalent if they differ by a scalar multiple. The equivalence class of $x \in V \setminus \{0_V\}$ is labeled $[x]$, and the set of all such equivalence classes is the *projective space* $P(V)$ associated with V , $P(V) = \{[x], x \in V \setminus \{0_V\}\}$. The real projective space in m dimensions, $\mathbb{R}P^m$, is $P(\mathbb{R}^{m+1})$. Another notation for a *projective point* $p = [x] \in \mathbb{R}P^m$, equivalence class of $x = (x^1, \dots, x^{m+1}) \in \mathbb{R}^{m+1}$, is $p = [x^1 : x^2 : \dots : x^{m+1}]$ featuring the *homogeneous coordinates* of p .
- ▶ A **projective transformation** is a map $\pi : P(V) \rightarrow P(V)$, given by $\pi([x]) = [Ax]$, $A \in GL(V)$.

Image acquisition in ideal digital cameras.

- ▶ *Ideal pinhole camera* image acquisition can be thought of in terms of a central projection β from $\mathbb{R}P^3 \setminus \mathbb{R}P^2$ to $\mathbb{R}P^2$. In homogeneous coordinates $[x : y : z : w]$, $[u : v : t]$ the *perspective map* β is given by the matrix $B \in M(3, 4; \mathbb{R})$ given by:

$$B = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \quad (4)$$

- ▶ In addition to the projective map (4), image formation in digital cameras assumes a composition with matrices accounting for camera internal calibration parameters, such as *pixel aspect ratio*, *skew parameter*, *origin of image coordinates in the principal plane (principal point)* and for a change of coordinates between two camera positions involving a roto-translation $(R, T) \in SO(3) \times \mathbb{R}^3$. The projective map of pinhole camera image acquisition $\tilde{\pi}$, in homogeneous coordinates has the matrix:

$$\tilde{B} = K_{\text{int}} B C = \begin{pmatrix} k_U & k_C & u_0 \\ 0 & k_V & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} R & T \\ 0_3^T & 1 \end{pmatrix} = K C, \quad (5)$$

where k_U and k_V are scale factors of the image plane in units of the focal length f , and $\theta = \cot^{-1} k_C$ is the skew, and (u_0, v_0) is the *principal point*. K for internal parameters and perspective (4), while E for external parameters. The matrix \tilde{B} can be decomposed into a 3×3 matrix P and a 3×1 vector p $\tilde{B} = (P p)$, so that $P = AR$ and $p = AT$.

Epipoles. Epipolar lines

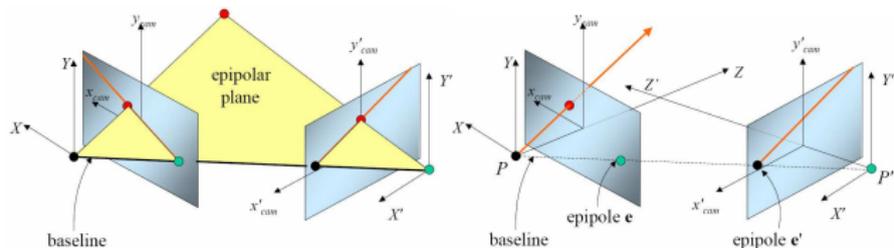


Figure: Virtual point correspondences in two camera views. Epipoles.

The epipolar constraint For every point observed in the left image we know that its correspondence must lie along the corresponding epipolar line in the right image For every epipolar line in the left image there is a corresponding epipolar line in the right image This observation can substantially simplify the search for correspondences

Epipolar Geometry

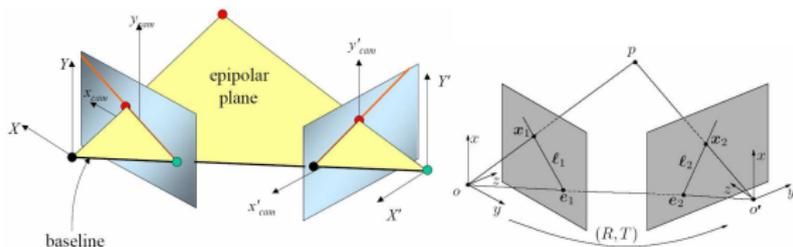


Figure: $T = \overline{OO'}$ and R is the camera rotation.

$T \times R x_1 \perp$ epipolar plane.

x_2 is on the epipolar plane, therefore $x_2^T T \times R x_1 = 0$.

Essential and fundamental matrices

- ▶ The application $x \rightarrow T \times x$ is linear. The matrix associated with this linear function is labeled T_{\times} . The **essential matrix (Longuet-Higgins (1981))** is $E = T_{\times} R$. It turns out that we have

$$x_2^T E x_1 = 0. \quad (6)$$

- ▶ In addition, due to internal camera parameter matrices K_1, K_2 , $x_a = K_a u_a$, $a = 1, 2$, and in pixel coordinates u_1, u_2 we have

$$u_2^T F u_1 = 0, F = K_2^T E K_1. \quad (7)$$

- ▶ F is the fundamental matrix. If only one camera is used, $F = K^T E K$. The camera is **uncalibrated** if K is unknown.

3D Reconstruction Problem.

The problem of the reconstruction of a configuration of points in 3D from two ideal noncalibrated camera images with unknown camera parameters, is equivalent to the following: given two camera images regarded as subsets in $\mathbb{R}P_1^2, \mathbb{R}P_2^2$ of unknown relative position and internal camera parameters and two matched sets of labeled points $\{p_{a,1}, \dots, p_{a,k}\} \subset \mathbb{R}P_a^2, a = 1, 2$, find all the sets of points in space p_1, \dots, p_k in such that there exist two positions of the planes $\mathbb{R}P_1^2, \mathbb{R}P_2^2$ and internal parameters of the two cameras $c_a, a = 1, 2$ with the property that the c_a -image of p_j is $p_{a,j}, \forall a = 1, 2, j = 1, \dots, k$.

Projective Ambiguity of the Reconstruction from pairs of Uncalibrated Images.

The projective ambiguity of the reconstruction from uncalibrated cameras is due to **Faugeras(1992) and Hartley et. al. (1992)** leading to:

THEOREM(2006-Sughatadasa-Patragenaru). In absence of occlusions, the projective shape of a 3D configuration \mathcal{R} reconstructed from a pair of matched configurations in noncalibrated cameras images of a 3D configuration \mathcal{C} , and the projective shape of \mathcal{C} are the same.

Note that for any pair of 2D images, a **landmark correspondence is needed** to obtain a 3D reconstruction

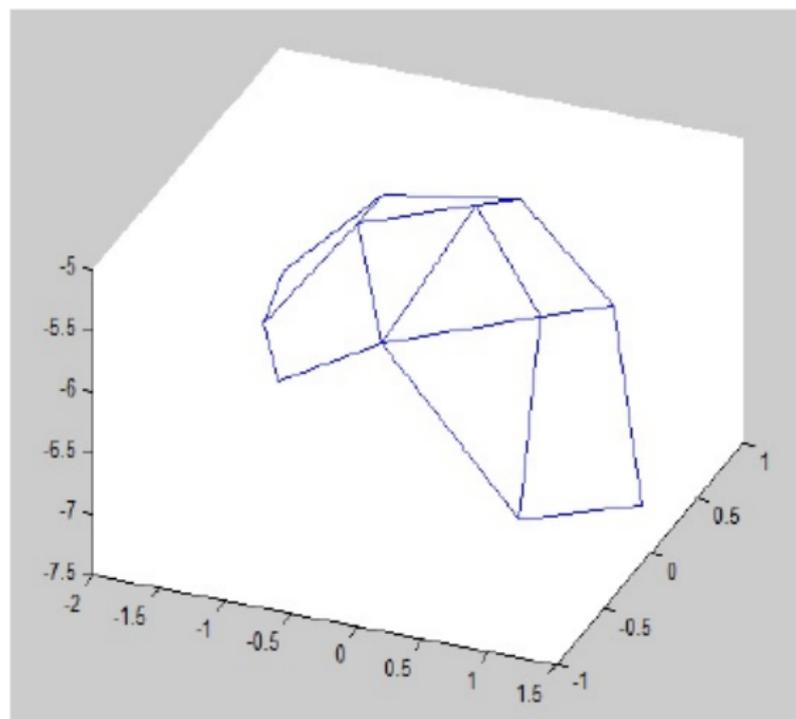
Planar pattern used for camera calibration



Two pictures of shell with landmarks



Reconstruction obtained from landmark correspondences



Asymptotic nonparametric inference based on Schoenberg means on $\sum_{p,0}^k$

The manifold approach to reflection-shape analysis, including Schoenberg embeddings and connections to multidimensional scaling (MDS), was initiated by Bandulasiri and Patrangenaru (2005). Two configurations of points have the same *reflection shape* if they differ by a similarity of the ambient Euclidean space. The set of all configurations having the same reflection shape as a given configuration is said to be a *reflection shape* (see Dryden and Mardia (1998)).

Asymptotic nonparametric inference based on Schoenberg means on $\sum_{p,0}^k$

Given a sample of reflection shapes $x_i = [u_i], i = 1, \dots, n$, their sample *Schoenberg mean reflection shape* \hat{X}_R considered here is based on the Schoenberg embedding in Bandulasiri et al. (2009a), where the distances between configurations are chord distances in the ambient space of matrices. Here $\hat{X}_R = [\hat{u}]_R$, where $\hat{u} \in M(p, k; \mathbb{R})$ is the unique minimizer of

$$\mathbf{u} \rightarrow \text{Tr}(\hat{B} - \mathbf{u}^T \mathbf{u})^2$$

which is determined up to an orthogonal transformation, satisfying the constraints

$$\mathbf{u} \mathbf{1}_k = 0$$

and

$$\text{Tr}(\mathbf{u}^T \mathbf{u}) = 1.$$

Asymptotic nonparametric inference based on Schoenberg means on $\sum_{p,0}^k$

In general, given a random reflection shape $X = [U]$, and the spectral decomposition $B = \sum_{i=1}^k \lambda_i e_i e_i^T$ of

$$B = E(U^T U),$$

the Schoenberg mean reflection shape exists if and only if $\lambda_p > \lambda_{p+1}$. If this is the case, then \mathbf{u}^T can be taken as the matrix $V = (v_1 v_2 \dots v_p)$, whose columns are orthogonal eigenvectors of B corresponding to the largest eigenvalues $\lambda_1 \geq \dots \geq \lambda_p$ of B , with

$$(a) v_j^T \mathbf{1}_k = 0$$

and

$$(b) v_j^T v_j = \lambda_j + \frac{1}{p}(\lambda_{p+1} + \dots + \lambda_k), \quad \forall j = 1, \dots, p.$$

Schoenberg sample mean computations for the 3D data

We then have $\hat{C}_{large} = \frac{1}{8} \sum_{j=\{1,2,3,4,5,6,8,9\}} U_j^T U_j$

The eigenvalues of \hat{C}_{large} satisfies $\lambda_p > \lambda_{p+1}$. So the extrinsic mean μ_E reflection shape exists and $\mu_E = [u]_R$ where u^T can be taken as

$$V = (v_1 v_2 v_3)$$

whose columns are orthogonal eigenvectors of \hat{C}_{large} corresponding to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ of \hat{C}_{large} , with

$$v_j^T v_j = \lambda_j + \frac{1}{3}(\lambda_4 + \dots + \lambda_k), \quad \forall j = 1, 2, 3.$$

The Schoenberg sample mean $[\hat{\xi}_{large}]_R$ reflection shape of the large group is

$$[\hat{\xi}_{large}]_R = \{A\hat{u}_{large} : A \in O(3)\}$$

Schoenberg sample mean computations for the 3D data

Similarly, $\hat{C}_{small} = \frac{1}{7} \sum_{j=\{10,11,12,14,17,18,21\}} U_j^T U_j$.

The eigenvalues of \hat{C}_{small} satisfy $\lambda_p > \lambda_{p+1}$. So the Schoenberg mean reflection shape $\hat{\mu}_E$ exists and $\mu_E = [u]_R$ where u^T can be taken as

$$V = (v_1 v_2 v_3)$$

whose columns are orthogonal eigenvectors of \hat{C}_{small} corresponding to the largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ of \hat{C}_{small} , with

$$v_j^T v_j = \lambda_j + \frac{1}{3}(\lambda_4 + \dots + \lambda_k), \quad \forall j = 1, 2, 3.$$

The Schoenberg sample mean $[\hat{\xi}_{large}]_R$ reflection shape of the small group is

$$[\hat{\xi}_{small}]_R = \{A\hat{u}_{small} : A \in O(3)\}$$

Schoenberg sample mean computations for the 3D data

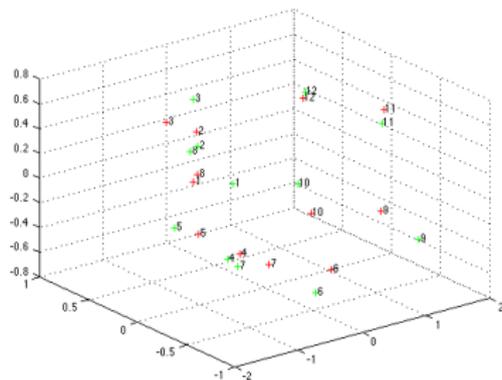


Figure: Icons of extrinsic mean shape for large shells sample(red) and small shells sample(blue)

Schoenberg sample mean computations for the 3D data

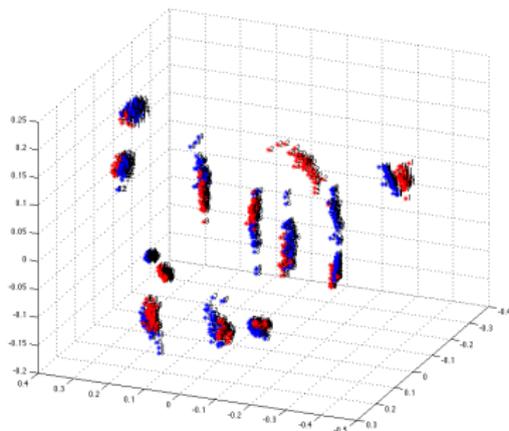


Figure: Distributions of bootstrapped extrinsic mean shape for large shells sample(red) and small shells sample(blue).

Diffusion Tensor Imaging

DTI data is an MRI byproduct signaling the presence and direction of water flow in the brain. At each voxel, the axis of an ellipsoid pointing the directions and amounts of water flows are given. The ellipsoid is described by the coefficients of the equation

$$\sum_{i,j=1}^3 d_{ij}x^i x^j = 1$$

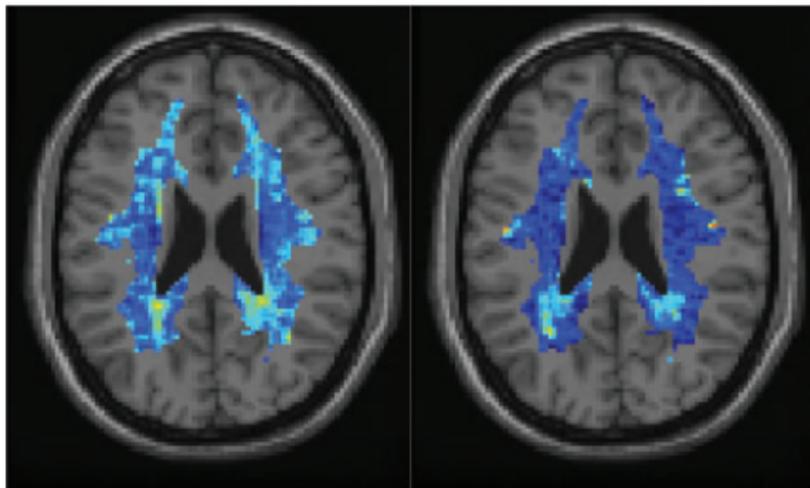


Figure: DTI slice intensity level of the d_{11} entry a control subject (left) and of a dyslexia subject (right).

Diffusion Tensor Imaging Analysis

We compare the average DTI changes between clinically normal and dyslexic kids (Osborne et al. (2013)) from two small samples (sample size 6) of DTI outputs. Dyslexia is curable (the cure seldom consists in studying Math and Stat:)). Hint : two sample tests on $Sym_+(3)$ via Cholesky decompositions.

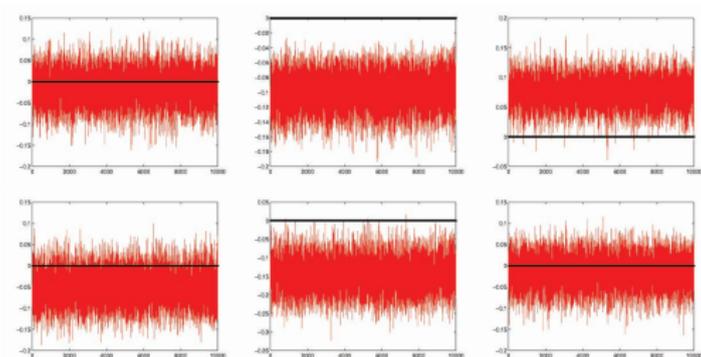


Figure: Bootstrap distribution of our test statistics V : The images (1 - 3) in the first row corresponds to the diagonal entries of the matrices V^* : v_{11} , v_{22} , v_{33} and images (4 - 6) in the second row corresponds to the lower off-diagonal entries of the matrices V^* : v_{21} , v_{31} , v_{32}

Why Object Data Analysis on Lie Groups?



(a) Statue A.

(b) Statue B.

Figure: Statue Data.

Question: Could one tell from random views of two busts from Musei Capitolini, Rome, Italy, if they might represent the same person?

Projective Shape Manifolds.

- ▶ In general, if a configuration of points contains a projective frame, then the projective shape of that configuration, is determined by the projective coordinates of the other points of this configuration w.r.t. this projective frame.
- ▶ Let $P\Sigma_m^k$ be the subset of $P_0\Sigma_m^k$, of *generic* orbits of k-tuples in $\mathbb{R}P^m$, such that the first $m + 2$ points in the k-tuple form a projective frame. $P\Sigma_m^k$ is a manifold diffeomorphic with $(\mathbb{R}P^m)^{k-m-2}$, leading to a **multivariate axial data analysis** (Mardia and P., Ann. Statist., 2005)

Veronese-Whitney Embedding of $P\Sigma_m^k$.

Assume $q = k - m - 2$. Mardia and Patrangenaru (2005) considered the diagonal equivariant embedding

$$J = j_k : P\Sigma_m^k = (\mathbb{R}P^m)^q \rightarrow (S(m+1))^q$$

defined by

$$j_k([x_1], \dots, [x_q]) = (j([x_1]), \dots, j([x_q])), \quad (8)$$

where $x_s \in \mathbb{R}^{m+1}$, $x_s^T x_s = 1$, $\forall s = 1, \dots, q$ and j is the Veronese-Whitney embedding

$$j([x]) = xx^T, x^T x = 1. \quad (9)$$

Mean projective shapes.

A random projective shape Y of a k -ad in $\mathbb{R}P^m$ is given in axial representation by the multivariate random axes

$$(Y^1, \dots, Y^q), Y^s = [X^s], (X^s)^T X^s = 1, \forall s = 1, \dots, q = k - m - 2. \quad (10)$$

The extrinsic mean projective shape of (Y^1, \dots, Y^q) exists if $\forall s = 1, \dots, q$, the largest eigenvalue of $E(X^s(X^s)^T)$ is simple. In this case μ_{j_k} is given by $\mu_{j_k} = ([\gamma_1(m+1)], \dots, [\gamma_q(m+1)])$, where $\lambda_s(a)$ and $\gamma_s(a)$, $a = 1, \dots, m+1$ are the eigenvalues in increasing order and the corresponding unit eigenvector of $E(X^s(X^s)^T)$.

If Y_r , $r = 1, \dots, n$ are i.i.d.r.o.'s from a population of projective shapes, for which the mean shape μ_{j_k} exists, the extrinsic sample mean $[\bar{Y}]_{j_k, n}$ can be obtained from $Y_r = ([X_r^1], \dots, [X_r^q])$, $(X_r^s)^T X_r^s = 1$; $s = 1, \dots, q$. Let J_s as follows: $J_s = n^{-1} \sum_{r=1}^n X_r^s (X_r^s)^T$, $s = 1, \dots, q$, $d_s(a)$ and $g_s(a)$ are the eigenvalues in increasing order and the corresponding unit eigenvectors of J_s , $a = 1, \dots, m+1$, then *sample mean projective shape* is given by

$$\bar{Y}_{j_k, n} = ([g_1(m+1)], \dots, [g_q(m+1)]). \quad (11)$$

Large Sample Theory for Projective Shapes-1

Let Σ be the covariance matrix of $j_k(Y^1, \dots, Y^q)$ regarded as a random vector in $(S(m+1))^q$, with respect to this standard basis, and let $P =: P_{j_k} : (S(m+1))^q \rightarrow j_k((\mathbb{R}P^m)^q)$ be the projection on $j_k((\mathbb{R}P^m)^q)$. Patrangenaru et. al. (2009) showed that with respect to a conveniently selected orthonormal basis of $T_{\mu_{j_k}}(\mathbb{R}P^m)^q$, the extrinsic covariance matrix of (Y^1, \dots, Y^q) is given by

$$\begin{aligned} \Sigma_{j_k} &= \left[e_{(s,a)}(P_{j_k}(\mu)) \cdot D_{\mu} P_{j_k}(r e_a^b) \right]_{(s=1, \dots, q), (a=1, \dots, m)} \cdot \Sigma \\ &\quad \cdot \left[e_{(s,a)}(P_{j_k}(\mu)) \cdot D_{\mu} P_{j_k}(r e_a^b) \right]_{(s=1, \dots, q), (a=1, \dots, m)}^T. \end{aligned} \quad (12)$$

Large Sample Theory for Projective Shapes-2

Assume Y_1, \dots, Y_n is a random sample from a j_k -nonfocal probability measure on $(\mathbb{R}P^m)^q$ and μ_{j_k} in (??) is the extrinsic mean of Y_1 . The expression of the entries of the extrinsic sample covariance matrix $G(j_k, Y)$ was also given in Patrangenaru et. al (2009). We arrange the pairs of indices (s, a) , $s = 1, \dots, q$; $a = 1, \dots, m$, in their lexicographic order, then with respect to a convenient orthonormal basis, the entries of $G(j_k, Y)$ are given by

$$G(j_k, Y)_{(s,a),(t,b)} = n^{-1} (d_S(m+1) - d_S(a))^{-1} (d_t(m+1) - d_t(b))^{-1} \cdot \sum_{r=1}^n (g_S(a)^T X_r^S)(g_t(b)^T X_r^t)(g_S(m+1)^T X_r^S)(g_t(m+1)^T X_r^t). \quad (13)$$

From Bhattacharya and Patrangenaru (2005) it follows that $G(j_k, Y)$ is a strongly consistent estimator of the population extrinsic covariance matrix Σ_{j_k} . In preparation for an asymptotic distribution of $\bar{Y}_{j_k, n}$ we set

$$D_S = (g_S(1) \dots g_S(m)) \in \mathcal{M}(m+1, m; \mathbb{R}), s = 1, \dots, q. \quad (14)$$

Large Sample Theory for Projective Shapes-3

If $\mu = ([\gamma_1], \dots, [\gamma_q])$, where $\gamma_s \in \mathbb{R}^{m+1}$, $\gamma_s^T \gamma_s = 1$, for $s = 1, \dots, q$, we define a Hotelling's T^2 type-statistic

$$T(\bar{Y}_{j_k, n}; \mu) = n(\gamma_1^T D_1, \dots, \gamma_q^T D_q) G_n^{-1} (\gamma_1^T D_1, \dots, \gamma_q^T D_q)^T. \quad (15)$$

Large and Small Sample Theory for Projective Shapes

Assume $(Y_r)_{r=1,\dots,n}$ are i.i.d.r.o.'s on $(\mathbb{R}P^m)^q$, and Y_1 is j_k -nonfocal, with $\Sigma_E > 0$. Let $\lambda_s(a)$ and $\gamma_s(a)$ be the eigenvalues in increasing order and corresponding unit eigenvectors of $E[X_1^a(X_1^a)^T]$. If $\lambda_s(1) > 0$, for $s = 1, \dots, q$, then $T(\bar{Y}_{j_k, n}; \mu_{j_k})$ converges weakly to χ_{mq}^2 .

For small samples, use Efron's bootstrap.

A Lie Group Structure on a Manifold of 3D Projective Shapes

- ▶ Note that unlike in other dimensions the projective shape manifold $P\Sigma_3^k$ **has a Lie group structure**.
- ▶ This Lie group is homeomorphic to $M = (\mathbb{R}P^3)^q$, where $q = k - 5$. Therefore with this identification, $P\Sigma_3^k \sim (\mathbb{R}P^3)^q$ *inherits a Lie group structure* from the group structure p-quaternions $\mathbb{R}P^3$ given in (17) on the next slide. The multiplication in $(\mathbb{R}P^3)^q$ is given by

$$([h_1], \dots, [h_q]) \cdot_q ([h'_1], \dots, [h'_q]) := ([h_1] \cdot [h'_1], \dots, [h_q] \cdot [h'_q]) = ([h_1 \cdot h'_1], \dots, [h_q \cdot h'_q]) \quad (16)$$

- ▶ The identity element is given by $1_{(\mathbb{R}P^3)^q} = ([0 : 0 : 0 : 1], \dots, [0 : 0 : 0 : 1])$, and given a point $\mathbf{h} = ([h_1], \dots, [h_q]) \in (\mathbb{R}P^3)^q$, from (18), its inverse is $\mathbf{h}^{-1} = \bar{\mathbf{h}} = ([\bar{h}_1], \dots, [\bar{h}_q])$.

Quaternions and P-Quaternions - detail

If a real number x is identified with $(0, 0, 0, x) \in \mathbb{R}^4$, and if we label the quadruples $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, respectively $(0, 0, 1, 0)$ by \vec{i} , \vec{j} , respectively \vec{k} , then the multiplication table given by

\cdot	\vec{i}	\vec{j}	\vec{k}
\vec{i}	-1	\vec{k}	$-\vec{j}$
\vec{j}	$-\vec{k}$	-1	\vec{i}
\vec{k}	\vec{j}	$-\vec{i}$	-1

can be extended by linearity to a multiplication \cdot of \mathbb{R}^4 , and $(\mathbb{R}^4, +, \cdot)$ has a structure of a noncommutative field, the field of *quaternions*, usually labeled by \mathbb{H} , in honor of William Rowan Hamilton. Note that if $h, h' \in \mathbb{H}$, then $\|h \cdot h'\| = \|h\| \|h'\|$, therefore the three dimensional sphere inherits a group structure, the *group of quaternions of norm one*.

Moreover since $\mathbb{R}P^3$ is the quotient $S^3/x \sim -x$

$$[x] \cdot [y] =: [x \cdot y], \quad (17)$$

is a well defined *Lie group* operator on $\mathbb{R}P^3$, called the group of **p-quaternions**. Note that if $h = t + x\vec{i} + y\vec{j} + z\vec{k}$, its *conjugate* is $\bar{h} = t - x\vec{i} - y\vec{j} - z\vec{k}$, and it turns out that the inverse of h is given by

$$h^{-1} = \|h\|^{-2} \bar{h}, \quad (18)$$

Quaternions and P-Quaternions - as Riemannian Homogeneous Spaces

Note that $\mathbb{R}P^3$ is isomorphic with $SO(3)$, and the Lie algebra $so(3)$ is

$$[X_1, X_2] = X_3, [X_2, X_3] = X_1, [X_3, X_1] = X_2, \quad (19)$$

therefore the Left invariant dual Pfaff forms satisfy

$$d\omega^1 + \omega^2 \wedge \omega^3 = 0, d\omega^2 + \omega^3 \wedge \omega^1 = 0, d\omega^3 + \omega^1 \wedge \omega^2 = 0 \quad (20)$$

Using the **Cartan triple method**, Patrangenaru (1994) showed that any homogeneous metric on $SO(3)$ is isometric with a metric of the form

$$g = \lambda_1(\omega^1)^2 + \lambda_2(\omega^2)^2 + \lambda_3(\omega^3)^2, \quad (21)$$

for some positive constants $\lambda_j, j = 1, 2, 3$. This space has constant curvature (is a quotient of the round sphere S^3), only if $\lambda_1 = \lambda_2 = \lambda_3$. The choice of equal lambdas in intrinsic projective shape analysis, is meant only to ease computational aspects of the intrinsic data analysis, which, even in this case, is slower than the VW analysis.

Two sample test for VW-mean 3D projective shapes from stereo images.

- ▶ Testing the hypothesis $H_0 : \mu_1 = \mu_2$ on a Lie group G is same as testing $H_0 : \mu = 1_G$, where $\mu = \mu_2^{-1} \mu_1$.
- ▶ From general considerations in large sample theory one should first derive the asymptotic distribution of $\hat{\mu} = \hat{\mu}_2^{-1} \hat{\mu}_1$, where $\hat{\mu}_a$ is a consistent estimator of μ_a , under H_0 , **via a delta method on manifolds**.
- ▶ In particular this distribution is determined in the case when $G = (\mathbb{R}P^3)^q$, with $q = k - 5$, and used to derive two sample tests for independent populations of 3D projective shapes, based on large samples.
- ▶ For small samples, we derive nonparametric bootstrap confidence regions for μ above based on

$$(\hat{\mu}_2^*)^{-1} \hat{\mu}_1^* \tag{22}$$

- ▶ This methodology will be applied to two a sample test for mean 3D projective shape from images of a face.

Half Bust Data



Figure: Epicurus bust images.

Hypothesis Testing on the Projective Shape Space

- ▶ Set the data of the first 16 images as H_1 , the rest 8 images as H_2 . The null and alternative hypothesis are

$$H_0 : \mu_{1,9}^{-1} \odot \mu_{2,9} = 1_4. \quad (23)$$

- ▶ Select landmarks 1, 4, 5, 6, 8 were to construct the projective frame. For the confidence region, we compute 2,000,000 bootstrap V-W sample means, based on landmarks 2, 3, 7, 9 by (??).
- ▶ $(0, 0, 0)^T$ is in the 4 boxes, which means we fail to reject the null; the two busts portrayed the same person.

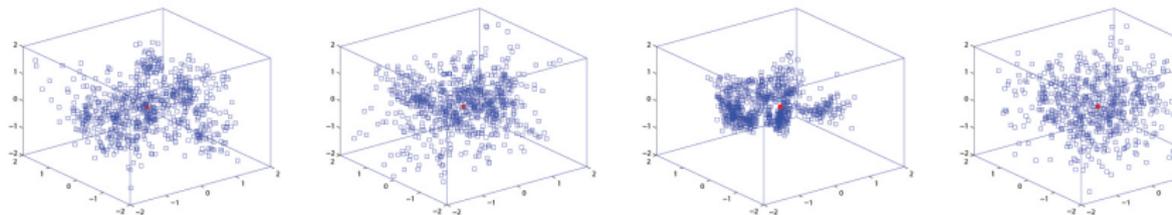


Figure: Simultaneous confidence regions for the statue data.

Shape of Contours Data

Shape of contours data is collected from digital images of 2D scenes, via edge maps extractors, and smoothing. Consider an example for one sample test for extrinsic mean shape is performed for sting ray contours. In this case, the sample extrinsic mean shape for a sample of contours of $n = 20$ sting rays is the shape shown on the left hand side in Fig. 16.

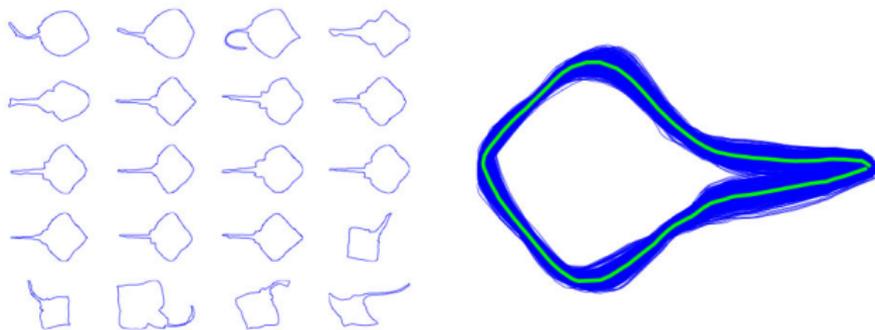


Figure: 95% bootstrap c.r. for the extrinsic mean shape based on 500 contour reps and 600 bootstrap resamples

3D virtual skull reconstructions from C T scans

Medical Imaging can be used for reconstructive plastic surgery of the skull.

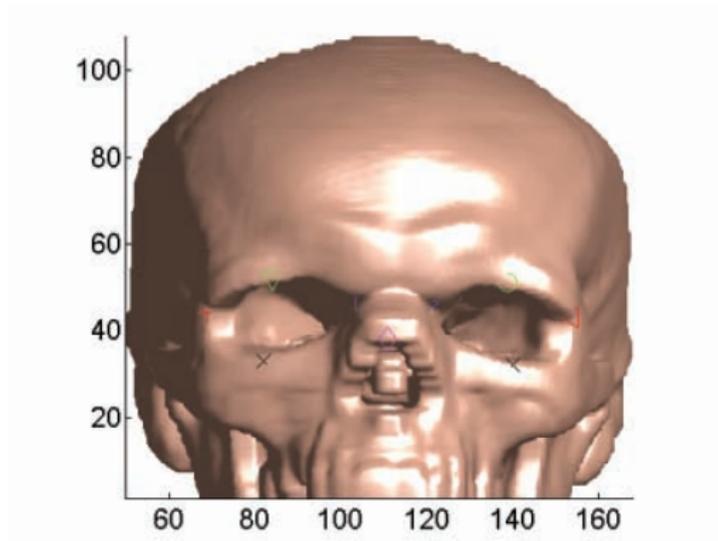


Figure: One (out of twenty) 3D virtual skull reconstruction with midface landmarks

3D mean size and shape of midface landmark configurations from 3D virtual skull reconstructions

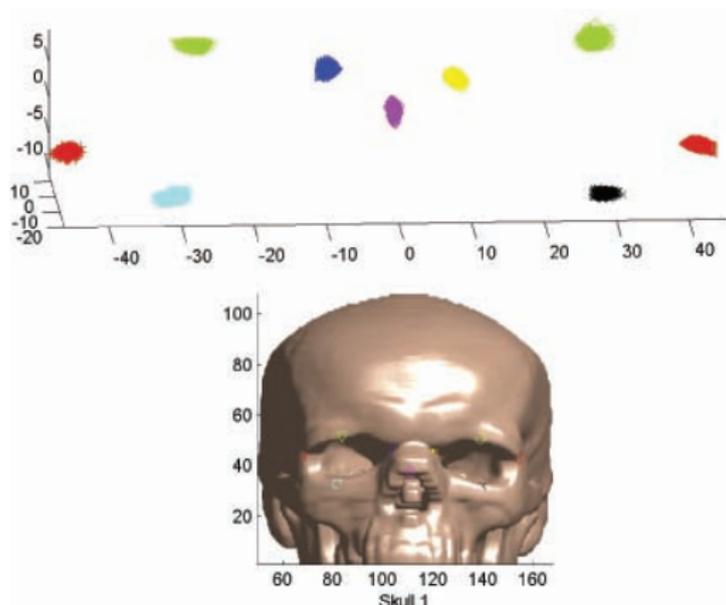


Figure: Simultaneous bootstrap c.r.'s for mean midface landmark configurations

Stereo data of eye fundus - courtesy H. W. Thompson (LSU-Biostatistics)

An L S U experiment for glaucomatous change detection using both Heidelberg Retina Tomograph and regular stereo eye fundus data was performed on a group of Rhesus monkey around Y2K. Increased Internal Ocular Pressure (IOP) was induced in one treated eye, with the other eye used as a control. All monkeys survived the experiment, still they could not be moved in time when Katrina hit downtown New Orleans. Stereo eye fundus data was recovered much later(2008), given that the PI himself (Claude Burgoyne) relocated in Portland after Katrina.).

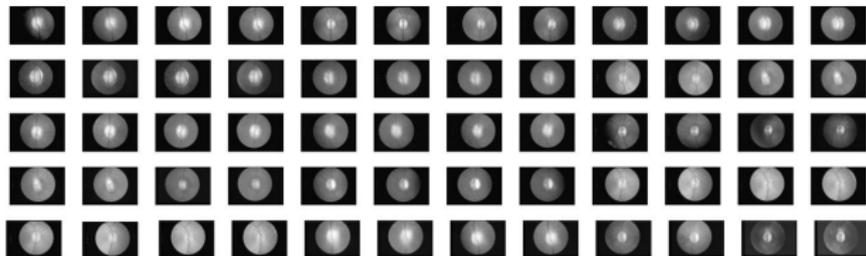


Figure: Rhesus monkeys eye fundus stereo data

Glaucomatous projective shape change detection

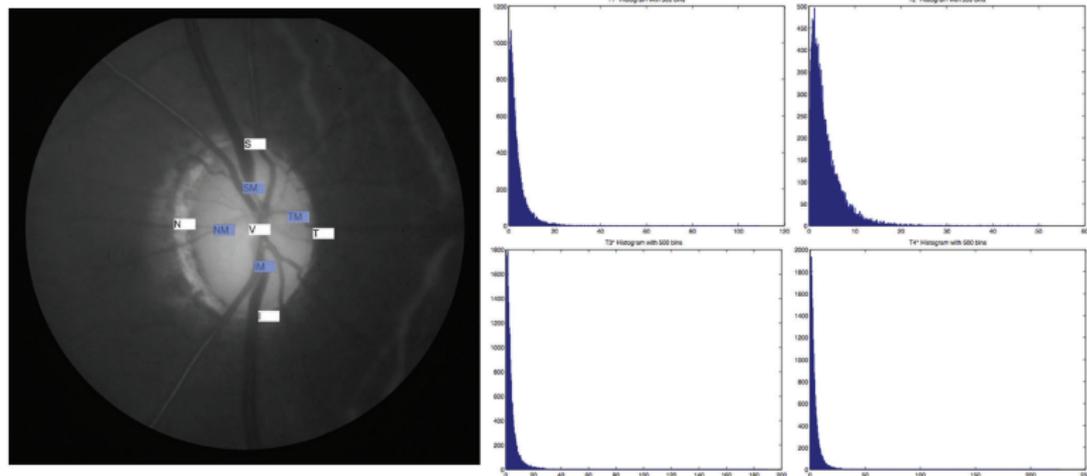


Figure: Detection of 3D projective shape change in 9 anatomical landmarks using 3D projective shape reconstructions from eye fundus stereo data, and bootstrap on projective quaternion reps

Fréchet's program for a Statistical Analysis of Elements

- ▶ The question of studying random element (nowadays called **random objects (r. o.'s)**), other than random vectors, was first raised by Maurice Fréchet (1948).
- ▶ As an example, Fréchet suggested to analyze **the shape of a contour of a closed curve, or the shape of an egg selected at random from an wire egg basket.**
- ▶ Fréchet's approach to AOD consists in identifying an object with a point in a metric space (M, d) .
- ▶ Next, given a r.o. X on M , he defined what we call today the *Fréchet functions* on M , given by $F_{d,q}(p) = E(d^q(X, p))$.
- ▶ A minimizer of $F_{d,2}$ above is a **Cartan-Fréchet mean**, and the minimum value of $F_{d,2}$ is the **Fréchet total variance**.
- ▶ Ziezold (1977): **the Cartan-Fréchet sample mean set is a consistent estimator of the Cartan-Fréchet population mean set.**

Nonparametric Statistics on Manifolds, prior to AOD program

- ▶ Function and density estimation on a manifold (Beran(1968), Wellner, Gine, Henricks, Kim) was the first target of data analysis on manifolds
- ▶ Secondly the asymptotic behavior of Fréchet sample means on manifold was possible, since *manifold* is locally modeled on its *tangent space* \mathbb{H} . The first results are due to Hendricks and Landsman and Patrangenaru($\dim\mathbb{H} < \infty$), and to Ellingson, Patrangenaru and Ruymgaart (2013)($\dim\mathbb{H} = \infty$)

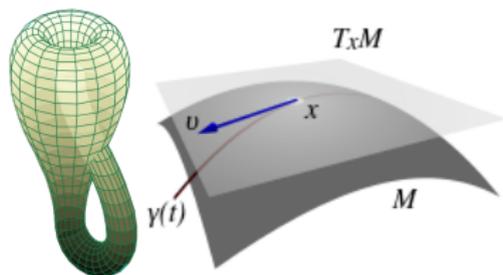


Figure: Left: A 2D manifold - Klein bottle. Right: a tangent space at a point

Fréchet's program - the case of Manifolds

- ▶ A *manifold* modeled on a Hilbert space \mathbb{H} , is locally, but not globally homeomorphic to \mathbb{H} .
- ▶ If $\dim \mathbb{H} < \infty$, manifold data, are a natural extension of multivariate data.
- ▶ If $\dim \mathbb{H} = \infty$, data on Hilbert manifolds are an extension of functional data.
- ▶ Dimension reduction via geodesic PCA for HDLSS data on certain Riemannian manifolds (Huckemann, Hotz and Munk (2010)).
- ▶ Extrinsic analysis on nonflat manifolds is computationally **faster** then intrinsic analysis (Bhattacharya, Crane, Ellingson, Liu and Patrangenaru (2012)).

Nonparametric Statistics on Manifolds (NSSS 1.0)

- ▶ A *manifold* modeled on a Hilbert space \mathbb{H} , is locally, but not globally homeomorphic to \mathbb{H} .
- ▶ If $\dim \mathbb{H} < \infty$, manifold data, are a natural extension of multivariate data.
- ▶ If $\dim \mathbb{H} = \infty$, data on Hilbert manifolds are an extension of functional data.
- ▶ Large sample theory of *extrinsic* sample means ($d =_j d$, where $j : M \rightarrow L$ is an *embedding* of M), are due to Hendricks and Landsman(1996,1998) and Patrangenaru(1998). A CLT for *intrinsic* sample means ($d = d_g$, where g is a Riemannian tensor on the manifold) was given in Bhattacharya and Patrangenaru (2005).
- ▶ The multivariate normal asymptotic behavior of the Fréchet sample means in the tangent space at the population Fréchet mean, assuming the later exists, where also given in Bhattacharya and Patrangenaru (2005).

Stratified Spaces

- ▶ A **stratified space (space with a manifold stratification)** is a metric space M that admits a *filtration*
 $\emptyset = F_{-1} \subseteq F_0 \subseteq F_1 \dots \subseteq F_n \subset \dots = M = \bigcup_{i=0}^{\infty} F_i$, by closed subspaces, such that for each $i = 1, \dots, F_i \setminus F_{i-1}$ is empty or is an i -dimensional manifold, called the i -*th stratum*.
- ▶ The *regular part* of M is the highest dimensional stratum. At each regular point the stratified space has a tangent space
- ▶ The dimension of the stratified space is m if $M = F_m \neq F_{m-1}$, otherwise $\dim M = \infty$. All the strata of dimension lower than m are *singular*: at each of their points, the stratified space does not have a tangent point

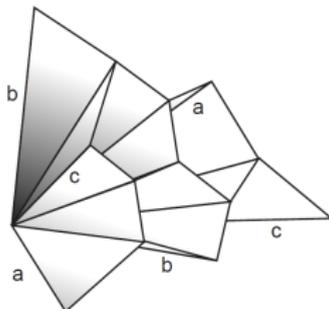


Figure: A 2D stratified space - T_4 , space of trees with four leaves

Phylogenetic Trees

- ▶ More recently the explosion of genetic data available through molecular biology has made tree-building even more popular. This presentation aims to give a motivation for working with metric trees data.
- ▶ The data that biologists use, usually comes from one homogenous sequence. When they talk about homogenous, they talk about problems between gene trees and genes that are made from one tree. The gene sequence might be about 200 base pairs long.
- ▶ One of the ideas that biologists believe in, is that the way evolution works is that there would only be **one species tree**. Different genes have different histories, so you get different gene trees. Putting them together is a statistical problem that helps understanding a possible evolutionary process

Example of Species Tree

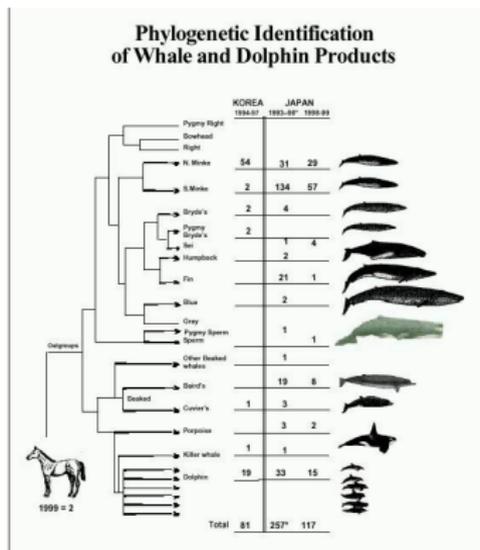


Figure: Phylogenetic tree of cetaceans (courtesy Susan Holmes)

The Space of Phylogenetic Trees with p Leaves

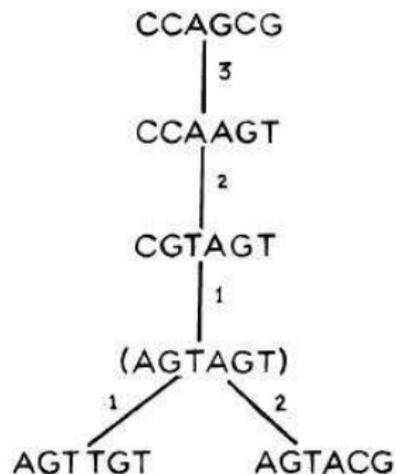
A phylogenetic tree with p leaves is an equivalence class based on a certain equivalence, of a DNA-based connected directed graph of species with no loops, having an unobserved *root* (common ancestor) and p observed *leaves* (current observed species of a certain family of living creatures).

```
Human   MVHLTPEEKSAVTALWGVVNDVEVGGEALGRLLVVPWTQRFFESPGDLSTPDVAMGNPK
Gorilla MVHLTPEEKSAVTALWGVVNDVEVGGEALGRLLVVPWTQRFFESPGDLSTPDVAMGNPK
Rabbit  MVHLSSEKSAVTALWGVVNDVEVGGEALGRLLVVPWTQRFFESPGDLSSANAVMNNPK
Cow     M. .LTAEKAAVTAFWGKVKVDEVGGEALGRLLVVPWTQRFFESPGDLSTADAVMNNPK
Goat    M. .LTAEKAAVTCFWGKVKVDEVGGEALGRLLVVPWTQRFFESPGDLSSADAVMNNPK
House   MVHLTDAEKAAVSVCLWGVVNDVEVGGEALGRLLVVPWTQRFFESPGDLSSASAIMGNPK
Chicken MVHWTAEKQLITGLWGVVNAECGAEALARLLIVVPWTQRFFASPGNLSPTAILGNPK
Carp    MVEWTDARSAIIGLWGLNPDGLGQALARCLIVVPWTQRFFASPGNLSPPAAMGNPK
```

```
61 120
Human   VKAHGKVLGAFSDGLAHLDNLKTGFATLSELHCCKLHVDPENFRLLGNVLYCVLARHFG
Gorilla VKAHGKVLGAFSDGLAHLDNLKTGFATLSELHCCKLHVDPENFKLLGNVLYCVLARHFG
Rabbit  VKAHGKVLGAFSEGLSHLDNLKTGFATLSELHCCKLHVDPENFRLLGNVLYVLSRHFG
Cow     VKAHGKVLGAFSDGKHLDDLKTGFATLSELHCCKLHVDPENFKLLGNVLYVLSRHFG
Goat    VKAHGKVLGAFSDGKHLDDLKTGFATLSELHCCKLHVDPENFKLLGNVLYVLSRHFG
House   VKAHGKVLGAFSDGKHLDDLKTGFATLSELHCCKLHVDPENFRLLGNVLYVLSRHFG
Chicken VRAHGKVLGAFSDGKHLDDLKTGFATLSELHCCKLHVDPENFRLLGNVLYVLSRHFG
Carp    VAHGRTVMGLERAIKIMDNIKATYAPLSVMHSEKLVDPDNFRLLADCTVCAAMGNPK
```

```
121 148
Human   .KEFTPPVQAAYQKVVAGVANALAHKYH
Gorilla .K.....
Rabbit  .KEFTPQVQAAYQKVVAGVANALAHKYH
Cow     .KEFTPVLQADFQKVVAGVANALAHKYH
Goat    .SEFTPLQAEFQKVVAGVANALAHKYH
House   .KDFTPAAQAAYQKVVAGVANALAHKYH
Chicken .KDFTECQAAYQKVVAGVANALAHKYH
Carp    PSCGFSFNVQEAQKFLSVVVSALCRQYH
```

Length of Edges on a Phylogenetic Tree



Tree Spaces

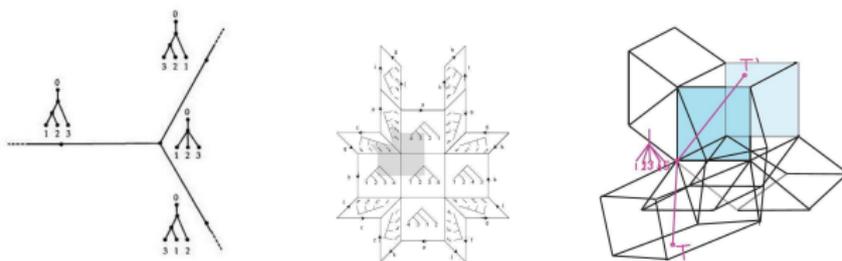


Figure: Tree spaces T_3, T_4, T_5 .

- ▶ A **tree with p leaves** is a connected, simply connected graph, with a distinguished vertex, labeled o , called the *root*, and p vertices of degree 1, called leaves, that are labeled from 1 to p . In addition, we assume that with all interior edges have positive lengths. (An edge of a p -tree is called interior if it is not connected to a leaf.) See Billera et. al.(2001).
- ▶ Now consider a tree T , with interior edges e_1, \dots, e_r of lengths l_1, \dots, l_r respectively. If T is binary, then $r = p - 2$, otherwise $r < p - 2$. The vector $(l_1, \dots, l_r)^T$ specifies a point in the positive open *orthant* $(0, \infty)^r$.

Tree Spaces are Stratified Spaces

- ▶ An p -tree has the maximal possible number of interior edges (namely $p-2$) and thus determines the largest possible dimensional orthant, when it is a binary tree; in this case the orthant is $p - 2$ -dimensional. The orthant corresponding to each tree which is not binary appears as a boundary face of the orthants corresponding to at least three binary trees; in particular the origin of each orthant corresponds to the (unique) tree with no interior edges. We construct the space T_p by taking one $p - 2$ -dimensional orthant for each of the $(2p - 3)!!$ possible binary trees, and gluing them together along their common faces.
- ▶ Note that tree spaces are *not* manifolds. Singularities (points where the space does not have a tangent space) are present in the tree space structure. For the stratification of the low dimensional tree spaces, see figure 24.

AOD = Intro to Statistics on Stratified Spaces (SSS)

- ▶ Recall that the space of trees with 3 leaves is $T_3 = S_3$, a 3-spider, union of three line segments with a common end.
- ▶ For a probability measure on S_ρ , if none of the the “legs” of the ρ -spider has a dominant probability mass, then the Fréchet mean is the star tree, which was the motivation for studying the asymptotics of Fréchet sample means of distributions on T_ρ . (Basrak(2010) and Hotz et. al.(2013)).
- ▶ There is no clear biological interpretation though.
- ▶ In shape analysis, in 3D computational examples lead to the idea of *sticky sample mean*, a new phenomenon.
- ▶ Computational examples from tree and shape data led to the creation of the Working Group on **Data Analysis on Sample Spaces with a Manifold Stratification**.
- ▶ One initial goal, was to understand the asymptotics of the Fréchet means for distributions on such stratified spaces.

Probability measures on Spiders

- ▶ $X_i, i = 1, \dots, n$ i.i.d.r.o.'s on S_p , of legs $L_a, a = 1, \dots, p$, and center C .
- ▶ The Fréchet mean $\mu_{X_1, F}$ exists
- ▶ The Fréchet variance is finite.
- ▶ Any probability measure Q on S_p decomposes uniquely as a weighted sum of probability measures Q_k on the legs L_k and a probability measure Q_0 is the atomic probability at C . More precisely, there are nonnegative real numbers $\{w_k\}_{k=0}^p$ summing to 1 such that, for any Borel set $A \subseteq S_p$, the measure Q takes the value

$$Q(A) = w_0 Q_0(A \cap C) + \sum_{k=1}^p w_k Q_k(A \cap L_k).$$

- ▶ The nontrivial case at: at least of the moments $\mu_a = E(Q_a), a = 1, \dots, p$ are positive.

CLT on Spiders (NSSS 2.1)

Hotz et. al. (2010).

- ▶ If $\exists a \in \overline{1, p}$ such that $\mu_a > \sum_{b \neq a} \mu_b$ then $\mu_{X_1, F} \in L_a$ and $\bar{X}_{n, F} \in L_a$, and $\sqrt{n}(\bar{X}_{n, F} - \mu_{X_1, F})$ has asymptotically a normal distribution with finite mean.
- ▶ If $\exists a \in \overline{1, p}$ such that $\mu_a = \sum_{b \neq a} \mu_b$, then asymptotically, after **folding the legs** $L_b, b \neq a$ into one **half line opposite to** L_a , $\sqrt{n}(\bar{X}_{n, F})$ has asymptotically the distribution of the absolute value of a normal distribution with finite mean.
- ▶ If $\forall a \in \overline{1, p}, \mu_a < \sum_{b \neq a} \mu_b, \mu_{X_1, F} = C$ and there is n_0 s.t. $\forall n \geq n_0, \bar{X}_{n, F} = 0$ a.s. (this is the **stickiness** phenomenon).

The result above was recently extended to C. L. T. on *open books* (Hotz et. al (2013)).

CLT on Trees (NSSS 2.s, $s > 1$)

Basrak(2010) extended his result to distributions on binary metric trees. The following are the final comments of that paper: **Note that the results and proofs above extend directly to arbitrary k -spiders and locally finite nonbinary metric trees. The strong laws of large numbers hold unaltered in these cases too. Limit theorems will need minor adjustments, since on a general tree, the barycenter can split the tree into more than three subtrees. Nevertheless, asymptotically, the inductive mean will have one of the three types of behavior described in Theorem 3. In particular, the phase transition in the limiting behavior is still to be expected.**

Πανδωρα's box - CLT's on a graph including cycles

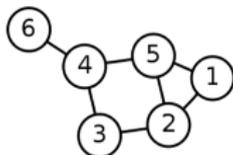


Figure: Bicyclic graph

- ▶ Basrak(2010) comment on the previous slide does not extend to graphs, since graphs usually have cycles.
- ▶ The key property of metric acyclic graphs (trees), is that they are $\text{Cat}(0)$, and a r.o. on a $\text{Cat}(0)$ space with finite Fréchet variance, has a Fréchet mean.
- ▶ There are r.o.'s on a circle having a Fréchet mean set with at least two objects. In fact the intrinsic mean set may contain an arbitrary number of points, or be even the entire circle.

Basrak's sufficient condition for CLT's on a graph

- ▶ Basrak(2010) condition for trees can be extended to the case of graphs in some general cases even if the graph G has cycles with positive mass on any arc of a cycle (Hendricks and Patrangenaru (2014)).
- ▶ The case when μ_F is on an edge (on the highest dimensional cycle is obvious).
- ▶ Assume μ_F is a vertex, and $G \setminus \mu_F$ at least two connected components of positive mass, and $Q(\omega_a(\mu_F, \varepsilon)) = 0$, where $\omega_a(\alpha, \varepsilon)$ is an ε - neighborhood of the farthest point to α on the connected component C_a of $G \setminus \alpha$.
- ▶ Assume C_1, \dots, C_r are the connected components of $G \setminus \mu_F$.
- ▶ Let $\mu_a = E(\tilde{Q}_a)$, where \tilde{Q}_a is the probability measure obtained from Q_a by folding the legs or arcs of C_a .

Under the assumptions above if X_1, \dots, X_n are i.i.d.r.o.'s on G with mean μ_F and finite total Fréchet variance. Then

- ▶ If $\exists a \in \overline{1, p}$ such that $\mu_a > \sum_{b \neq a} \mu_b$ then $\mu_{X_1, F} \in C_a$ and $\bar{X}_{n, F} \in C_a$, and $\sqrt{n}(\bar{X}_{n, F} - \mu_F)$ has asymptotically a normal distribution with finite mean.
- ▶ If $\forall a \in \overline{1, p}, \mu_a < \sum_{b \neq a} \mu_b$, then $\mu_{X_1, F} = C$ and, almost surely, for n large enough $\bar{X}_{n, F} = C$ as well. This is the *sticky sample mean*

Example: Sticky and Regular Means on Bicyclic Graph

(Loading movie...)

Drawbacks of an intrinsic analysis on tree spaces

- ▶ An analysis of a population of trees based on intrinsic means may be uninformative if the intrinsic sample mean sticks to a lower dimensional stratum of the tree space.
- ▶ On the other hand, an extrinsic analysis a population of trees based on extrinsic mean sets is may be helpful, given the consistency of the extrinsic sample mean sets. Each point in the extrinsic sample mean set has asymptotically a multivariate normal distribution around the point of the extrinsic population mean set in the corresponding orthant.
- ▶ Similarly, the intrinsic sample mean might often time stick to a vertex, thus making the data analysis on a graph difficult.
- ▶ There are no necessary and sufficient conditions for the existence of the intrinsic mean on a non simply connected graph.
- ▶ On the other hand there are such conditions for the existence of extrinsic means on graphs and tree spaces.

Sticky Mean on Tree Spaces

(Loading movie...)

Embeddings of Open Books

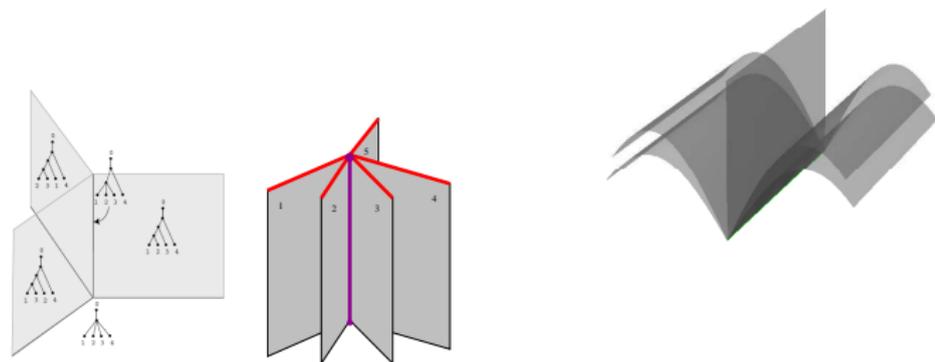


Figure: Embeddings of open books : part of T_4 , hard book and paper back book

The extrinsic mean set of a distribution on a carton open book (see figure 83) is on the spine if and only the distribution is entirely concentrated on the spine; such embeddings do not present the stickiness phenomenon. On the other hand, paperback open books present a stickiness phenomenon for a large family of distributions. In general the CLT for an embedding of stratified spaces is now known (see Bhattacharya et. al.(2013)).

Estimation of Means on Hilbert Manifolds - an Introduction

Motivated by a lack of nonparametric methods for inference in high level digital image analysis, we introduce a general extrinsic approach for data analysis on Hilbert manifolds. By embedding a Hilbert manifold into a Hilbert space, we can define extrinsic means and their sample analogues. To perform inference on these means, we appeal to the concept of neighborhood hypotheses from functional data analysis and derive a one-sample test. We apply these methods to the problem of estimating mean shapes of planar contours while considering the computational restrictions faced when utilizing digital imaging data. Comparisons of computational cost are provided to another method for analyzing contours.

Hilbert Manifolds

Note that a Hilbert space is a Banach space. Therefore differentiability in a Hilbert space is differentiability in the sense of Banach spaces.

A chart on a separable metric space (\mathcal{M}, ρ) is a one to one homeomorphism $\varphi : U \rightarrow \varphi(U)$ defined on an open subset U of \mathcal{M} to a Hilbert space \mathbf{H} . A **Hilbert manifold** is a separable metric space \mathcal{M} , that admits an open covering by domain of charts, such that the transition maps $\varphi_V \circ \varphi_U^{-1} : \varphi_U(U \cap V) \rightarrow \varphi_V(U \cap V)$ are differentiable.

Example - The projective space associated with a Hilbert space

The projective space $P(\mathbf{H})$ of a Hilbert space \mathbf{H} , the space of all one dimensional linear subspaces of \mathbf{H} , has a natural structure of Hilbert manifold modelled over \mathbf{H} . Define the distance between two vector lines as their angle, and, given a line $\mathbb{L} \subset \mathbf{H}$, a neighborhood $U_{\mathbb{L}}$ of \mathbb{L} can be mapped via a homeomorphism $\varphi_{\mathbb{L}}$ onto an open neighborhood of the orthocomplement \mathbb{L}^{\perp} by using the decomposition $\mathbf{H} = \mathbb{L} \oplus \mathbb{L}^{\perp}$. Then for two perpendicular lines \mathbb{L}_1 and \mathbb{L}_2 , it is easy to show that the transition maps $\varphi_{\mathbb{L}_1} \circ \varphi_{\mathbb{L}_2}^{-1}$ are differentiable as maps between open subsets in \mathbb{L}_1^{\perp} , respectively in \mathbb{L}_2^{\perp} . A countable orthobasis of \mathbf{H} and the lines $\mathbb{L}_n, n \in \mathbb{N}$ generated by the vectors in this orthobasis is used to cover $P(\mathbf{H})$ with the open sets $U_{\mathbb{L}_n}, n \in \mathbb{N}$. Finally, use the fact that for any line $\mathbb{L}, \mathbb{L}^{\perp}$ and \mathbf{H} are isometric as Hilbert spaces. The line \mathbb{L} spanned by a nonzero vector $\gamma \in \mathbf{H}$ is usually denoted $[\gamma]$ when regarded as a projective point on $P(\mathbf{H})$.

Embeddings of means on Hilbert manifolds

An embedding of a Hilbert manifold \mathcal{M} in a Hilbert space \mathbb{H} is a one-to-one differentiable function $j : \mathcal{M} \rightarrow \mathbb{H}$, such that for each $x \in \mathcal{M}$, the differential $d_x j$ is one to one, and the range $j(\mathcal{M})$ is a closed subset of \mathbb{H} and the topology of \mathcal{M} is induced via j by the topology of \mathbb{H} .

Example We embed $P(\mathbf{H})$ in $\mathcal{L}_{HS} = \mathbf{H} \otimes \mathbf{H}$, the space of Hilbert-Schmidt operators of \mathbf{H} into itself, via the Veronese-Whitney (VW) embedding j given by

$$j([\gamma]) = \frac{1}{\|\gamma\|^2} \gamma \otimes \gamma. \quad (24)$$

If $\|\gamma\| = 1$, this definition can be reformulated as

$$j([\gamma]) = \gamma \otimes \gamma. \quad (25)$$

Given that $\gamma^*(\beta) = \langle \beta, \gamma \rangle$ equation (25) is equivalent to

$$j([\gamma])(\beta) = \langle \beta, \gamma \rangle \gamma. \quad (26)$$

The range of this embedding is the submanifold \mathcal{M}_1 of rank one Hilbert-Schmidt operators of \mathbf{H} .

Extrinsic means on Hilbert manifolds

Consider a random object X on a Hilbert manifold \mathcal{M} embedded in a Hilbert space, that has an extrinsic mean set. Then (i) $j(X)$ has a mean vector μ and (ii) the extrinsic mean set is the set of all points $x \in \mathcal{M}$, such that $j(x)$ is at minimum distance from μ . (iii) In particular, μ_E exists if there is a unique point on $j(\mathbf{M})$ at minimum distance from μ , the projection $P_j(\mu)$ of μ on $j(\mathbf{M})$, and in this case $\mu_E = j^{-1}(P_j(\mu))$.

Extrinsic analysis on Hilbert manifolds

Consider a random object X on a Hilbert manifold \mathcal{M} embedded in a Hilbert space, that has an extrinsic mean set. Then (i) $j(X)$ has a mean vector μ and (ii) the extrinsic mean set is the set of all points $x \in \mathcal{M}$, such that $j(x)$ is at minimum distance from μ . (iii) In particular, μ_E exists if there is a unique point on $j(\mathbf{M})$ at minimum distance from μ , the projection $P_j(\mu)$ of μ on $j(\mathbf{M})$, and in this case $\mu_E = j^{-1}(P_j(\mu))$. A random object X on a Hilbert manifold \mathcal{M} embedded in a Hilbert space is j -nonfocal if there is a unique point p on $j(\mathcal{M})$ at minimum distance from $E(j(X))$.

Assume $X = [\Gamma]$ is a random object in $P(\mathbf{H})$. Then the VW mean of X exists if and only if $E(\frac{1}{\|\Gamma\|^2} \Gamma \otimes \Gamma)$ has a simple largest eigenvalue, in which case, the distribution is j -nonfocal. In this case the VW mean is $\mu_E = [\gamma]$, where γ is an eigenvector for this eigenvalue.

A one-sample test of the neighborhood hypothesis

We can now define the neighborhood hypothesis procedure for tests of extrinsic means. Assume Σ_j is the extrinsic covariance operator of a random object X on the Hilbert manifold \mathcal{M} , with respect to the embedding $j : \mathcal{M} \rightarrow \mathbb{H}$. Let \mathbf{M}_0 be a compact submanifold of \mathcal{M} . Let $\varphi_0 : \mathcal{M} \rightarrow \mathbb{R}$ be the function

$$\varphi_0(p) = \min_{p_0 \in \mathbf{M}_0} \|j(p) - j(p_0)\|^2, \quad (27)$$

and let $\mathbf{M}_0^\delta, \mathbb{B}_0^\delta$ be given respectively by

$$\begin{aligned} \mathbf{M}_0^\delta &= \{p \in \mathcal{M}, \varphi_0(p) \leq \delta^2\}, \\ \mathbb{B}_0^\delta &= \{p \in \mathcal{M}, \varphi_0(p) = \delta^2, \}. \end{aligned} \quad (28)$$

Since φ_0 is Fréchet differentiable and all small enough $\delta > 0$ are regular values of φ_0 , it follows that \mathbf{B}_0^δ is a Hilbert submanifold of codimension one in \mathcal{M} . Let ν_p be the normal space at a points $p \in \mathbf{B}_0^\delta$, orthocomplement of the tangent space to \mathbb{B}_0^δ at p . We define $\mathbb{B}_0^{\delta, X}$

$$\mathbb{B}_0^{\delta, X} = \{p \in \mathbf{B}_0, \Sigma_j|_{\nu_p} \text{ is positive definite}\}. \quad (29)$$

Definition The neighborhood hypothesis consists in the following two alternative hypotheses:

Dimension reduction : from infinity to one

Here, we consider neighborhood hypothesis testing for the particular situation in which the submanifold \mathbf{M}_0 consists of a point m_0 on \mathcal{M} .

We set $\varphi_0 = \varphi_{m_0}$, and since $T_{m_0}\{m_0\} = 0$ we have the following result. **Theorem.** If $M_0 = \{m_0\}$, the test statistic for the neighborhood hypotheses has an asymptotically standard normal distribution and is given by:

$$T_n = \sqrt{n}\{\varphi_{m_0}(\hat{\mu}_E) - \delta^2\}/s_n, \quad (31)$$

where

$$s_n^2 = 4\langle \hat{\nu}, S_{E,n}\hat{\nu} \rangle \quad (32)$$

and

$$S_{E,n} = \frac{1}{n} \sum_{i=1}^n (\tan_{\hat{\mu}} d_{j(X)_n} P_j(j(X_i) - \overline{j(X)_n})) \otimes (\tan_{\hat{\mu}} d_{j(X)_n} P_j(j(X_i) - \overline{j(X)_n})) \quad (33)$$

is the extrinsic sample covariance operator for $\{X_i\}_{i=1}^n$, and

$$\hat{\nu} = (d_{\hat{\mu}_E, n} j)^{-1} \widehat{\tan}_{j(\hat{\mu}_E, n)}(j(m_0) - j(\hat{\mu}_E, n)). \quad (34)$$

THANK YOU !

References I

-  Beran, R. J.(1968). Testing for uniformity on a compact homogeneous space, *J. Appl. Probability*, **5**, 177–195.
-  Cartan, É. (1945). *Léçons sur la Géométrie des Espaces de Riemann (in French)*. Gauthier-Villars, Paris.
-  Rabi N. Bhattacharya, Marius Buibas, Ian L. Dryden, Leif A. Ellingson, David Groisser, Harrie Hendriks, Stephan Huckemann, Huiling Le, Xiuwen Liu, James S.Marron, Daniel E. Osborne, Vic Patrângenaru, Armin Schwartzman, Hilary W. Thompson, and Andrew T. A.Wood. (2013). Extrinsic data analysis on sample spaces with a manifold stratification. In *Advances in Mathematics, Invited Contributions at the Seventh Congress of Romanian Mathematicians, Brasov, 2011*, Publishing House of the Romanian Academy (Editors: Lucian Beznea, Vasile Brîzanescu, Marius Iosifescu, Gabriela Marinoschi, Radu Purice and Dan Timotin), pp. 148–156.

References II

-  Billera, Louis J.; Holmes, Susan P.; Vogtmann, Karen (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** , no. 4, 733–767
-  Bojan Basrak (2010). *Limit theorems for the inductive mean on metric trees* J. Appl. Prob. **47**, 1136–1149.
-  Fréchet, Maurice (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10**, 215–310.
-  Wang, H.; and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, **35**, 1849-1873.
-  Ellingson, L.; Patrangenaru, V.; and Ruymgaart, F. (2013). Nonparametric Estimation of Means on Hilbert Manifolds and Extrinsic Analysis of Mean Shapes of Contours. **122**, 317–333.
-  Mardia, K. V.; Patrangenaru, V. (2005). Directions and projective shapes. *The Annals of Statistics*, **33**, 1666-1699.

References III

-  R. N. Bhattacharya and V. Patrangenaru (2014). Rejoinder of Discussion paper “Statistics on Manifolds and Landmarks Based Image Analysis: A Nonparametric Theory with Applications.” *Journal of Statistical Planning and Inference*. **145**, 42–48.
-  Bhattacharya, R.; Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Annals of statistics*,**31**, 1-29.
-  Bhattacharya, R.; Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Annals of statistics*,**33**, 1211-1245.
-  R. N. Bhattacharya, L. Ellingson, X. Liu and V. Patrangenaru and M. Crane (2012). Extrinsic Analysis on Manifolds is Computationally Faster than Intrinsic Analysis, with Applications to Quality Control by Machine Vision. *Applied Stochastic Models in Business and Industry*. **28**, 222–235.

References IV

-  M. Crane and V. Patrangenaru. (2011). Random Change on a Lie Group and Mean Glaucomatous Projective Shape Change Detection From Stereo Pair Images. *Journal of Multivariate Analysis*. **102**, 225-237.
-  Dryden, I. L.; Mardia, K. V. (1992). Size and shape analysis of landmark data. *Biometrika* **79** , 57–68.
-  Efron, B.(1975). Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion), *Annals of Statistics* , **3**, 1189–1242
-  Efron, B.(1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* , **7**, 1–26
-  L. Ellingson, V. Patrangenaru, H. Hendriks, P. S. Valentin (2014). CLT on Low Dimensional Stratified Spaces. *In Press at Proceedings of the First INSPS Conference*.

References V

-  N. I. Fisher, P. Hall, B. Y. Jing and A. T. A. Wood (1996). Properties of principal component methods for functional and longitudinal data analysis. *J. Amer. Statist. Assoc.* **91**, 1062–1070.
-  Fréchet, Maurice (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10**, 215–310.
-  Harrie Hendriks and Z. Landsman, *Asymptotic behaviour of sample mean location for manifolds*, *Statistics and Probability Letters.* **26** (1996), 169–178.
-  Hendriks, H. and Landsman, Z. (1998). Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions. *J. Multivariate Anal.* **67** no. 2, 227-243.
-  Hendriks, H. and Patrangenaru (2014). Mean location and sample mean location on graphs: asymptotics, tests, confidence regions. *in work*.

References VI

-  Patrangenaru, V; Liu, X. and Sugathadasa, S.(2010). Nonparametric 3D Projective Shape Estimation from Pairs of 2D Images - I, In Memory of W.P. Dayawansa. *Journal of Multivariate Analysis*. **101**, 11-31.
-  Munk, A.; Paige, R.; Pang, J. ; Patrangenaru, V. and Ruymgaart, F. H.(2008). The One and Multisample Problem for Functional Data with Applications to Projective Shape Analysis. *J. of Multivariate Anal.* . **99**, 815-833.
-  D. Osborne, V. Patrangenaru, L. Ellingson, D. Groisser and A. Schwartzman. (2013). Nonparametric Two-Sample Tests on Homogeneous Riemannian Manifolds, Cholesky Decompositions and Diffusion Tensor Image Analysis. *Journal of Multivariate Analysis*. **119**, 163-175.
-  Bhattacharya, A. (2008). Statistical analysis on manifolds: A nonparametric approach for inference on shape spaces. *Sankhya* Vol **70**, no. 2, Ser. A, 223–266.

References VII

-  Thomas Hotz, Stephan Huckemann, Huiling Le, James S. Marron, Jonathan C. Mattingly, Ezra Miller, James Nolen, Megan Owen, Vic Patrangenaru and Sean Skwerer (2013). Sticky Central Limit Theorems on Open Books. *Annals of Applied Probability*, **23**, 2238–2258.
-  Stephan Huckemann, Thomas Hotz, and Axel Munk. (2010). Intrinsic shape analysis: geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion), **20**, 1–100.
-  Johnson, R. A.; Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Sixth Edition, Upper Saddle River, NJ: Prentice Hall.
-  Paige R. L.; Patrangenaru, V.; and Qiu, M. (2013). 3D Projective Shapes of Leaves from Image Data. *Abstracts of the 29th European Meeting of Statisticians, Budapest July 21-25*. p.232.

References VIII

-  V. Patrangenaru (1996). Classifying 3 and 4 dimensional Homogeneous Riemannian Manifolds by Cartan Triples. *Pacific Journal of Mathematics*. **173**. 511 – 532.
-  V. Patrangenaru, M. Qiu and M. Buibas(2014). Two Sample Tests for Mean 3D Projective Shapes from Digital Camera Images. *Methodology and Computing in Applied Probability*. **16**, 485–506.
-  R. Martin, P. Sunley, C. Berndt and B. Klagge (2005). *Venture Capital Programmes in the UK and Germany: In what Sense Regional Policies?*, 39, pp.255-273, *Regional Studies*.
-  Kim, Peter T.; Koo, Ja-Yong (2005). Statistical inverse problems on manifolds. *J. Fourier Anal. Appl.* **11**, no. 6, 639–653. *Stat. Methodol.* **9**, 211–227.
-  Ziezold, H. (1994). Mean figures and mean shapes applied to biological figure and shape distributions in the plane. *Biometrical J.* **36**, no. 4, 491–510.