Extrinsic Data Analysis on Sample Spaces with a Manifold Stratification

R. N. Bhattacharya∗-U Arizona, Tucson, AZ, USA,
M. Buibaş-Brain Corporation, San Diego, CA, USA
I. L. Dryden-U SC, Columbia, SC, USA
L. A. Ellingson†TTU, Lubbock, TX, USA,
D. Groisser-U Florida, Gainesville, FL, USA,
H. Hendriks -Radboud U, Nijmegen, Netherlands,
S. Huckemann‡IMS, Göppingen, Germany,
Huiling Le-University of Nottingham, UK,
X. Liu-FSU, Tallahassee, FL, USA§
D. E. Osborne¶FSU, Tallahassee, FL, USA,
V. Pătrângenaru∥FSU, Tallahassee, FL, USA,
J. S. Marron∗∗UNC, Chapel Hill, NC, USA,
A. Schwartzman††Harvard, Cambridge, MA, USA,
H. W. Thompson-LSU, New Orleans, LA, USA,
A.T.A. Wood-University of Nottingham, UK

March 8, 2012

∗Research partially supported by NSF Grants DMS-0806011, DMS-1107053
†Research partially supported by NSF Grant DMS-0805977
‡Research partially supported by DFG grants SFB 755 and HU 1575/2
§Research partially supported by NSF grant DMS-0713012
¶Research partially supported by NSF Grant DMS-1106935
∥Research partially supported by NSF Grants DMS-0805977, DMS-1106935
∗∗Research partially supported by NSF grant DMS-0854908
††Research partially supported by NIH Grant 1R21EB012177-01
Abstract

Given a random object on a stratified space embedded in a numerical space, often times the extrinsic sample means stick to a stratum that is closest to the population mean of the corresponding probability distribution in the ambient numerical space via this embedding. In the case of an open book, the limiting distribution of the extrinsic sample means supported on this possibly lower dimensional stratum, is extrinsically Gaussian on that stratum. Applications of the classical single stratum case to various types image data analysis, including to medical imaging, are also given.

Introduction

In general, given a complete separable metric space $(M, \rho)$, the Borel $\sigma$-algebra $B_M$ is the $\sigma$-algebra $\sigma(T_M)$ generated by the topology of $(M, \rho)$. Assume $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space. A random object (r.o.) on $M$ is function $X : \Omega \to M$, with the property that for any Borel set $B \in B_M, X^{-1}(B) \in \mathcal{A}$. To any r.o. $X$ we associate a probability measure $Q$ on $M$, given by $Q(B) = \mathbb{P}(X^{-1}(B))$. Given a r. o. $X$ in $M$, its Fréchet mean set is the set of all minimizers of the expected square distance of $X$ (assuming this is finite) to an arbitrary point on $M$. If the Fréchet mean set consists in one minimizer, the minimizer is called Fréchet mean of $X$ (see Fréchet (1948)[16], Ziezold (1977)[34]). Obviously the Fréchet mean (set) depends on the distance on $M$: The case in point here is when the distance is the chord distance induced by an embedding $j : M \to \mathbb{E}^N$, given in terms of the Euclidean norm by $\rho(x_1, x_2) = \|j(x_1) - j(x_2)\|$. In this case the Fréchet mean (set) is called extrinsic mean (set). If $\mu = E(j(X))$ and we set $M(j) = \{p \in j(M), \rho(p, \mu) = \rho(j(M), \mu)\}$, then the extrinsic mean set is $j^{-1}(M(j))$, and if $M(j)$ consists in a singleton, then $X$ has an extrinsic mean, which is labeled $\mu_j$. Given a sample $x_1, \ldots, x_n$ of objects on $M$, the extrinsic sample mean (set) is the extrinsic mean (set) of the empirical distribution $\hat{Q}_n = \sum_{i=1}^n \frac{1}{n} \delta_{x_i}$, where $\delta_x$ assigns probability one to $x$. The extrinsic sample mean, when it exists, will be labeled by $\bar{x}_{j,n}$. Assume $X_1, \ldots, X_n$ are independent identically distributed random objects (i.i.d.r.o.’s) on $M$. For extrinsic data analysis purposes, the first objective is the understanding of the asymptotic distribution of the extrinsic sample means $\bar{X}_{j,n}$. 2
This paper is a result of collaboration within the Working Group on Data Analysis on Sample Spaces with a Manifold Stratification, that ran under the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2010–2011 program on Analysis of Object Data. The group, in which most of the authors were members, considered practical applied questions from evolutionary biology, medical imaging, and shape analysis (see for example Patrangenaru and Ellingson (2011) [29]). Here are some key examples of stratified spaces considered:

- Projective shape spaces - sets of equivalence classes of configurations under actions of the pseudogroup of projective transformations, useful in analyzing 3D scenes from pinhole camera images outputs (Patrangenaru et al. (2010) [30]), affine shape spaces, useful in analyzing aerial or satellite images (Patrangenaru and Mardia (2003)[31]), or similarity shape spaces, relevant in medical imaging (see Bandulasiri et. al. (2009)[3] or Dryden and Mardia (1998, Chapter 6)[12], and in other applications (see Huckemann et. al (2010)[22] for example);

- Spaces of covariance matrices, arising as data points in diffusion tensor imaging (see Basser and Pierpaoli(1996)[4], Schwartzman(2008)[32] or Schwartzman et. al.(2008)[33] for example); and

- Tree spaces, representing spaces of metric phylogenetic trees on fixed sets of taxa (see Billera et. al.(2001)[8], Aydin et al. (2011)[1], for example).

In this paper, we briefly summarize extrinsic object data analysis, by reversing the order from the most recent to the more “classical” types of random objects. In Section 1 we are concerned with the asymptotic distribution in the case when $M$ is a stratified space, that is a metric space $(M, \rho)$ that decomposes as finite disjoint unions of manifolds (strata) in such a way that the singularities of the $M$ are constant along each stratum (see Goresky and MacPherson (1988, Section 1.4). The trivial case, of a manifold of dimension $d$ (stratified space having one $d$ dimensional stratum only) was settled by Hendricks and Landsman (1998)[19] and Bhattacharya and Patrangenaru (2005)[7], who showed that the asymptotic distribution of the tangential component of the extrinsic sample means $\hat{X}_{j,n}$ in...
the tangent space at the \((j)\) image of the extrinsic population mean, has a multivariate normal distribution \(\mathcal{N}_d(0_d, \frac{1}{n} \Sigma_j)\), where \(\Sigma_j\) is the extrinsic population covariance matrix.

For details see Bhattacharya and Patrangenaru (2005)[7], where it is also shown that the intrinsic sample means have a similar asymptotic distribution. A new behavior is exhibited when the stratified space has two strata: recently, Hotz et. al.(2012)[20], considered the asymptotic distribution of a certain intrinsic sample mean on an open book, which a space obtained by gluing disjoint copies of a Euclidean half-space along their boundaries (spine of the open book). They showed that, for a natural metric on the open book, if the intrinsic mean sits on the spine, than, often times the intrinsic sample means stick to the spine and have an asymptotically multivariate normal distribution there. In Section 1 we show that, for \(d = 1\), a similar asymptotic behavior is exhibited by the extrinsic sample means on open book, with embeddings literally reminding of an infinitely thin open paper back book in 3D.

In Section 2, we introduce the general concept of locally homogeneous space, and show that due to consistency of Fréchet sample means, any two sample problem can be reasonably tackled on such a sample space. Results concerning two sample tests for extrinsic means of random objects on locally homogeneous spaces are also given in this section, the data analysis being transported on a simply locally transitive Lie group of isometries. We apply our new methodology to a concrete Diffusion Tensor Imaging example, using a small size data set previously analyzed by Schwartzman et al.(2008)[32]. Section 3 is dedicated to 3D projective shape analysis of a scene from its from digital camera images. We notice that for \(k \geq 5\), a manifold of projective shapes \(k\)-ads in in general position in 3D has a structure of \(3k - 15\) dimensional manifold, that is locally modeled on a Lie group that is equivariantly embedded in an Euclidean space, therefore testing for mean change amounts to a one sample test for extrinsic means on this Lie group. The Lie group technique leads to a large sample and nonparametric bootstrap test for one population extrinsic mean on a projective shape space, as recently developed by Patrangenaru et. al.(2010)[30] and Crane and Patrangenaru (2011)[10]. On the other hand, in absence of occlusions, the 3D projective shape of a spatial configurations can be retrieved from a stereo pair of images, thus allowing to test for mean glaucomatous 3D projective shape change detection from standard stereo
pair of a scene, in particular of a face, via RANSAC, as noticed by Buibas et al. (2012)[9]. Sections 4 provides the main ingredients for asymptotic nonparametric inference based on *Schoenberg means* on $\Sigma_{p,0}^k$ and $\Sigma_{p,0}^{k}$. Bandulasiri et al. (2009)[3] showed that the necessary and sufficient condition for a probability measure $Q$ on $\Sigma_{p,0}^k$ or on $\Sigma_{p,0}^{k}$ to be Schoenberg-nonfocal is that, in their decreasing order, the eigenvalues of rank $p$ and $p + 1$ of the mean matrix of the push forward distribution are distinct. Although on $\Sigma_{p,0}^{k}$ this condition is the same as that for the existence of an *MDS mean* in Dryden et. al. (2008)[11], the extrinsic mean differs from the MDS mean in that paper. Section 4 illustrates this theory by applications with real CT scan data. Besides a high dimensional example of extrinsic sample mean reflection-size-and-shape of protein structures from the Protein Data Bank (PDB), first considered by Ellingson (2011), we derive simultaneous bootstrapped Bonferroni confidence regions are constructed for the mean similarity shape of nine facial landmarks from virtual reconstructions of 3D skulls from about 100 CT slices per head of an individual in a sample of twenty plus healthy men. This new medical imaging technique was first considered in D. Osborne et al. (2012)[28], and can be used for designing helmets to minimize the exposure of delicate structure of the head around the eyes region. The paper concludes with a discussion on other topics of the 2010/2011 SAMSI-AOD work group of Data Analysis on Sample Spaces with a Manifold Startification.

1 Extrinsic Central Limit Theorem on Open Books

Set $S = \mathbb{R}^d$, the real vector space of dimension $d$ with the standard Euclidean metric. If $\mathbb{R}_{\geq 0} = [0, \infty)$, then the closed half-space

$$\hat{H}_+ = \mathbb{R}_{\geq 0} \times S$$

is a metric subspace of $\mathbb{R}^{d+1} = \mathbb{R} \times S$, with boundary $S$, which we identify with $\{0\} \times S$, and interior $H_+ = \mathbb{R}_{> 0} \times S$. We consider the following equivalence relationship

$$\mathcal{R} \subset (\hat{H}_+ \times \{1, \ldots, K\}) \times (\hat{H}_+ \times \{1, \ldots, K\}) :$$

a pair $((x, k), (y, j)) = (x^{(0)}, x^{(1)}, \ldots, x^{(d)}, k), (y^{(0)}, y^{(1)}, \ldots, y^{(d)}, j)$ is in $\mathcal{R}$, if $x^{(0)} = y^{(0)} = 0$ and $x^{(i)} = y^{(i)}, \forall i = 1, \ldots, d$ or $x^{(0)} = y^{(0)} > 0$ and $x^{(i)} = y^{(i)}$ and $k = j$. The
open book $\mathcal{O}$ is the quotient $(\bar{H}_+ \times \{1, \ldots, K\})/\mathcal{R}$. The following definition summarizes and introduces terminology.

**DEFINITION 1.1 (Leaves and spine)** The open book $\mathcal{O}$ consists of $K \geq 3$ leaves $L_k$, for $k = 1, \ldots, K$, each of dimension $d + 1$. $L_k$ is the set of equivalence classes of points in $\bar{H}_+ \times \{k\}$. The leaves are joined together along the spine $L_0$, set of equivalence classes of points in $\{0\} \times S \times \{1, \ldots, K\}$, i.e. $L_0$ can be identified with the space $S = \mathbb{R}^d$.

**EXAMPLE 1.2** A piece of the open book $\mathcal{O}$ with $d = 1$ and $K = 5$ embedded in $\mathbb{R}^3$ is depicted in figure 1. Each embedded leaf continues to infinity away from the spine.

If for simplicity, we consider the case $d = 1, K$ arbitrary, and an embedding $J : \mathcal{O} \to \mathbb{R}^3$ similar with the one in example 1.2, we will restrict ourselves to random objects $X$ on $\mathcal{O}$ having a probability mass with the property $(p)$ that the mean $E(J(X))$ is closer to the spine $J(L_0)$, that to any of the open leaves $J(L_k \setminus L_0), k = 1, \ldots, K$.

**THEOREM 1.3 (Extrinsic Sticky CLT on $\mathcal{O}$)** Assume the independent identically distributed random objects $X_1, \ldots, X_n$ on $\mathcal{O}$, are such that $X_1$ has the property $(p)$. Let $P_{J(L_0)}$ be the projection on $J(L_0)$, and assume $J^{-1}(P_{J(L_0)}(J(X_1)))$ has finite variance $\sigma_{L_0}^2$. Then for
n large enough, the extrinsic sample mean $\bar{X}_{J,n}$ sticks to the spine $L_{0}$ and, on the spine $\sqrt{n}\sigma_{L_{0}}^{-1}(\bar{X}_{J,n} - \mu_{J})$ has asymptotically a standard normal distribution.

The proof of Theorem 1.3 follows from the fact that due to consistency, with probability one, for $n$ large enough, the sample mean $\bar{J}(X_{n})_{n}$ is closest to the spine than to any of the open leaves, thus the projection $P_{J}(X_{n})$ sticks to $J(L_{0})$. Since by the CLT, $\bar{J}(X_{n})_{n}$ has a trivariate normal distribution, the projection $P_{J(O)}(\bar{J}(X_{n})_{n})$ has a normal distribution, and the result follows by standardizing this asymptotic distribution.

2 Diffusion Tensor Imaging Data Analysis

An example of a medical imaging data analysis problem is data analysis of spatially registered Diffusion Tensors from MRI images, to track the water flow along axons in the brain. A diffusion tensor can be represented as $3 \times 3$ symmetric positive definite matrix, as shown by Basser and Pierpaoli (1996)[4]. Such matrices, when vectorized, form an open convex subset in $\mathbb{R}^{6}$ which has a locally homogeneous space structure under the additive group action of $\mathbb{R}^{6}$. Nonparametric bootstrap provides a reasonable estimation of the of the mean. This approach is used in a concrete example for detection from DTI of a condition called dyslexia, that is relatively frequent in humans. Following Schwartzman et al. (2008)[32], we consider an experiment in which DTI maps acquired for two groups of subjects, a group of healthy children and a group of children suffering from dyslexia, differ on average. In figure 2 we compare the bootstrap distribution of the sample means in these two populations, which are clearly showing a difference in the DTI means. Similar results were obtained by Schwartzman et al. (2008)[32] using a parametric approach. A nonparametric intrinsic approach for the same data was used recently in Osborne and Patrangenaru (2011) [27].
3 3D projective Shape Data Analysis

Based on a key result Computer Vision due to Faugeras(1992)[14] and Hartley et. al.(1992)[18], on the 3D reconstruction of a scene from two regular camera images, Patrangenaru et al.(2010)[30] showed that in absence of occlusions, the projective shape of a 3D configuration $\mathcal{R}$ reconstructed from a pair of matched 2D configurations in noncalibrated cameras images of a 3D configuration $\mathcal{C}$, and the projective shape of $\mathcal{C}$ are the same. For this reason, for a 3D projective shape analysis of a scene, it suffices to reconstruct a random sample of size $n$ of projective shapes of a 3D scene from its $2n$ digital camera images. If the configuration under investigation, consists in $k$ points in general position its projective shape is uniquely determined by the projective coordinates of $k - 5$ projective point, relative to a projective frame, that for together this configuration, so that $P^k_3$, the manifold of such projective shapes, is diffeomorphic with $(\mathbb{RP}^3)^{k-5}$, and, this manifold that can be embedded in $(S(m + 1))^{k-5}$ via the Veronese-Whitney embedding

$$j_k([x_1], ..., [x_q]) = (j([x_1]), ..., j([x_q])), \quad (3.1)$$

where $x_s \in \mathbb{R}^{m+1}, x_s^T x_s = 1, \forall s = 1, ..., q$ and $j([x]) = \frac{x x^T}{x^T x}$. Given a random projective shape $Y$ of a $k$-ad in $\mathbb{RP}^3$ is given in axial representation by the multivariate random axes

$$(Y^1, ..., Y^q), Y_s^{*} = [X_s^*], \quad (X_s^*)^T X_s^* = 1, \forall s = 1, ..., q = k - 5, \quad (3.2)$$

Bhattacharya and Patrangenaru (2003)[6] or Patrangenaru et. al.(2010)[30] showed that the extrinsic mean projective shape of $(Y^1, ..., Y^q)$ with respect to the Veronese-Whitney embedding (3.1) exists if $\forall s = 1, ..., q$, the largest eigenvalue of $E(X_s^*(X_s^*)^T)$ is simple. In this case $\mu_{j_k}$ is given by $\mu_{j_k} = ([\lambda_1(4)], ..., [\gamma_q(4)])$, where $\lambda_a(a)$ and $\gamma_s(a), a = 1, ..., 4$ are the eigenvalues in increasing order and the corresponding unit eigenvectors of $E(X_s^*(X_s^*)^T)$, and its sample counterpart is given by a similar formula:

$$\overline{Y}_{j_k,n} = ([g_1(4)], ..., [g_q(4)]). \quad (3.3)$$

This allows to derive various computational procedures for estimation and testing for mean projective shapes, as shown in Patrangenaru et al.(2010)[30] or in Bhattacharya and Pat-
trangenaru (2005)[7]. These procedures were extended to paired random samples of projective shapes, in Crane and Patrangenaru (2010)[10] using Lie group structure of the projective shape manifold $P_{\Sigma^k 3}$.

Face analysis from the projective shape retrieved from digital camera images was pursued by Buibas et al. (2012) [9] In Figure 3 one considers a configuration on a face, the 3D projective shape of the chin of that person, as retrieved from that configuration.

![Figure 3: Random landmarks displayed that are used for 3D face reconstruction of a face (left) and two views of reconstructed 3D configurations of points on the chin.](image)

### 4 3D Data Analysis in Medical Imaging

Gazing into the human body via medical imaging devices is an essential part of our health care system. The ability to produce medical images of the human body has lead to various new areas of research that studies the size-and-shape of all parts of the human body, such as the chest, skull, belly, pelvis, spinal cord, bladder, liver, lungs, pancreas, intestines, kidneys, heart, etc. Various techniques and processes are used to create images of the human body or parts and function thereof for clinical and medical science purposes. One important reason for using various medical imaging procedures is to reveal, diagnose, and examine diseases in the human body and maybe in some cases, in an animal body. Another important reason for using medical imaging procedures is to study the normal anatomy and physiology of the human body for future medical treatment. Nonparametric Statistical Analysis on Manifolds underwent a rapid development in recent years as far as applications to Medical Imaging are concerned. Statisticians are working more and more with nonlinear object data in medical
imaging, regarding their observations as points on manifolds. In this section, we provide a concrete example of such a medical imaging object data analysis. The data consists in twenty CT scans, all including the part of an individual’s head above the mandible. Figure 4 displays a CT Scan of one such observation. Here, we are interested in preforming a landmark based statistical analysis based on 3D virtual skulls; however, before such a goal can be executed, various imaging processing techniques must be applied to the CT data in order to reconstruct the 3D virtual skulls. This must be done before one can select a group of matched landmarks on the 3D virtual skulls. Details on the reconstruction of the head bone structure from these data can be found in Osborne (2012)[26].

In order to perform landmark based statistical analysis of the 3D virtual skulls, one suggestion is to model the CT image analysis on the Size-and-Reflection Shape Space. Consider nine matched landmarks around the eyes displayed in Figure 5. Nonparametric statistical analysis on the 3D size-and-reflection shape space can be preformed based on the Schoenberg embedding, Bandulasiri and Patrangenaru (2005) [2] Bandulasiri et.al. (2009)[3], fol-

Figure 4: One CT Scan of the head

Figure 5: Left: A group of nine landmarks around the eye. Right: a 95% confidence region for these nine landmarks.
lowed by calculating the Schoenberg Extrinsic 3D Mean Shape Configuration, given in Osborne et.al. (2012)[28].

5 Discussion

At this time we had to omit additional extrinsic object data analysis examples that were pursued in our work group of the SAMSI AOD program 2010/2011. Those include 2D direct similarity shape analysis of regular contours, nonparametric bootstrap on stratified spaces, including bootstrap on tree spaces, stratifications of affine shape spaces and affine shape analysis, extrinsic PCA on manifolds, hyperbolic data analysis and 3D reconstruction of a 3D scene from a pair of its calibrated digital camera images. These additional topics will be presented elsewhere.

Acknowledgements

The authors are grateful to Ezra Miller and Frits Ruymgaart for their advice and help. We are grateful to NSF and the organizers of the 2010-2011 program of Analysis of Object Data at SAMSI, for giving us the opportunity to work jointly on this research and related topics.

References


Applications in Medical Imaging. *Journal of Multivariate Analysis*. **100** 1867-1882.


[5] Bhattacharya, R.N; Ellingson, L; Liu, X; Patrangenaru, V; and Crane, M. *Extrinsic Analysis on Manifolds is Computationally Faster than Intrinsic Analysis, with Application to Quality Control by Machine Vision*. To appear *ASMB*.


