STATISTICAL ANALYSIS OF LONGITUDINAL DATA AND APPLICATIONS TO NEURO-IMAGING

Rudrasis Chakraborty[1] *and Baba C. Vemuri*[1]

[1]Department of Computer and Information Science and Engineering
University of Florida, Gainesville, FL, USA

ABSTRACT

Longitudinal data analysis is an often encountered problem in Medical Image Analysis. A differential geometric treatment of such problems has been reported in literature in the recent past. However, most of these methods require that the trajectory characterizing the evolution of features over time lie on a geodesic emanating from the initial time point on the manifold containing the trajectory. This is a stringent and not necessarily a meaningful requirement. Further, most of these methods impose the restriction that the number of samples on a trajectory be the same across the members of a group of trajectories. At times, this restriction is hard to meet from a practical view point. In this paper, we present a novel formulation of the trajectory analysis problem that overcomes the aforementioned limitations. We represent the trajectories by embedding them in a product Riemannian manifold and endowing it with a Riemannian metric, thereby facilitating the statistical analysis. Finally, we present real data (from MR brain scans of dementia patients) examples depicting the performance of our algorithms.

Index Terms— Longitudinal studies, Trajectories, Manifolds, PGA.

1. INTRODUCTION

In this paper, we present a novel method to perform statistical analysis on the "space of trajectories" which is a Riemannian manifold. We will formally define "space of trajectories" in Section 2, but one can view a trajectory as a path consisting of discrete points on a Riemannian manifold. It is very common in longitudinal data analysis applications in medical imaging to formulate this problem by considering the "trajectory" to lie on a Riemannian manifold. For example, it is often possible to identify patients with dementia by taking a time course of magnetic resonance (MR) brain scans over time, assessing and comparing the structural changes from the norm in the corpus callosum. But, as the "space of trajectories" does not have a well-known geometric structure, we first give the "space of trajectories" a topological structure by identifying it with a product space of two Riemannian manifolds. In the rest of paper, we use the term "space of trajectories" to denote the identification with the product space of two Riemannian manifolds. In this work, we have used the publicly available OASIS data [1], where each patient has multiple brain MR scans from different visits over time. For each patient, we track the changes in their corpus callosum and map them on to the space of trajectories. Then, we perform statistical analysis including computation of Fréchet mean (FM) [2] and the modes of variation to classify patients with dementia versus people without dementia. The salient features of our proposed

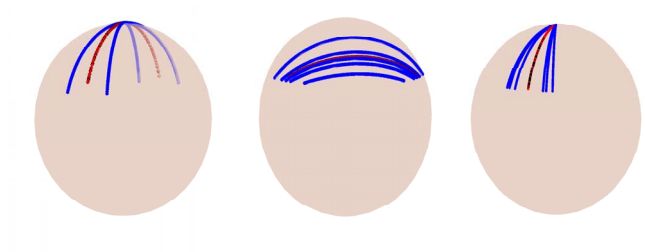

Fig. 1: Toy example

method are, (1) as our identification is a bijection, we can recover the trajectory completely, hence we can compute the mean trajectory or the "trajectory atlas" of a population of trajectories. (2) Unlike most of the existing works in literature, our method does not assume that each trajectory is along a geodesic on the manifold. (3) Further, unlike the previous works, our formulation allows for trajectories with variable number of points on them. *To the best of our knowledge, our technique is the first method without the linearity assumption, i.e., the entire trajectory is not required to lie on a geodesic, further, the number of points on each trajectory is not required to be equal.* In the "toy" example of Figure (1), we show trajectories on the sphere, $\mathbf{S}^2$, (shown in blue) and the mean of these trajectories is shown in red. It is evident that the mean trajectory goes through the "midpoint" of the the sample trajectories. In the rest of the paper, we first present the algorithm to get the mapping to identify the space of trajectories with the product space of two Riemannian manifolds. We have taken multiple MR scans of both demented and non-

This research was funded in part by the NSF grant IIS-1525431 to BCV.

demented patients and map them on to trajectories that lie on a hypersphere. Then, we compute the FM and the principal geodesics to classify demented and non-demented patients. Before we discuss our framework to compute statistics on the space of trajectories, we will briefly discuss some of the recent related work.

In [3], the authors developed a framework to compute "trajectory atlas" and registration of trajectories. They modeled each trajectory as a smooth curve on a Riemannian manifold. In this work, the authors formulated the computation of "trajectory atlas" as an optimization. From a computational perspective, this is not as efficient as our proposed method, which does not need any optimization. In [4], authors identify each trajectory as a point on the tangent bundle and then perform statistical analysis on the tangent bundle using the Sasaki metric. In [5], authors performed a principal geodesic analysis (PGA) on the tangent bundle to achieve PGA of the longitudinal dataset. Authors in [6] performed the segmentation of motion characterized by trajectories on a Riemannian manifold.

2. STATISTICS ON THE SPACE OF TRAJECTORIES

In this section, we first briefly discuss the geometry of Stiefel manifold, henceforth denoted by $\mathrm{St}(p,n)$ which will be required throughout the paper. Then, we will define the space of trajectories and compute statistics on this space.

Stiefel manifold : The set of all full column rank $(n \times p)$ dimensional real matrices form a Stiefel manifold, $\mathrm{St}(p,n)$, where $n \geq p$. A compact Stiefel manifold is the set of all column orthonormal real matrices. In many Computer Vision and Medical Imaging applications, we encounter the compact Stiefel manifold and hence in the rest of the paper, we concern ourselves only with this manifold. With a slight abuse of notation, henceforth, we denote the compact Stiefel manifold by $St(p,n)$ defined formally as, $\mathrm{St}(p,n) = \{X \in \mathbf{R}^{n\times p} | X^T X = I_p\}$, where I_p is the $p \times p$ identity matrix. At any $X \in \mathrm{St}(p,n)$, the tangent space $T_X\mathrm{St}(p,n)$ is defined as follows $T_X\mathrm{St}(p,n) = \{U \in \mathbf{R}^{n\times p} | X^T U + U^T X = 0\}$. Now, given $U, V \in T_X\mathrm{St}(p,n)$, the canonical Riemannian metric on $\mathrm{St}(p,n)$ is defined as follows $\langle U, V\rangle_X = trace\Big(U^T V\Big)$.

Given $X \in \mathrm{St}(p,n)$, we can define the Riemannian retraction and lifting map within an open neighborhood of X. We will use an efficient Cayley type retraction and lifting maps on $\mathrm{St}(p,n)$ as defined in [7].

Here, we use the fact that the special orthogonal group, $\mathrm{SO}(n)$ acts on $\mathrm{St}(p,n)$ by pre-multiplication, i.e., if $g \in \mathrm{SO}(n)$, $gX \in \mathrm{St}(p,n)$. Note that, $\mathrm{SO}(n) = \{X \in \mathbf{R}^{n\times n} | \ X^T X = I_n, \det(X) = 1\}$. Given $X \in \mathrm{St}(p,n)$, we define the lifting map $\mathrm{Exp}_X^{-1} : \mathrm{St}(p,n) \to T_X\mathrm{St}(p,n)$ by $\mathrm{Exp}_X^{-1}(Y) = \begin{bmatrix} C & -B^T \\ B & 0 \end{bmatrix}$ where, C is a $p \times p$ skew-symmetric matrix and B is a $(n-p)\times p$ matrix defined as follows: $C = 2(X_u^T + Y_u^T)^{-1}\mathrm{sk}(Y_u^T X_u + X_l^T Y_l)(X_u + Y_u)^{-1}$ and $B = (Y_l - X_l)(X_u + Y_u)^{-1}$ where, $X = [X_u, X_l]^T$, and $Y = [Y_u, Y_l]^T$ with $X_u, Y_u \in \mathbf{R}^{p\times p}$, and $X_l, Y_l \in \mathbf{R}^{(n-p)\times p}$, provided that $X_u + Y_u$ is nonsingular. And $\mathrm{sk}(M)$ is defined as $\frac{1}{2}(M^T - M)$ and, $Y \in St(p,n)$. Given $W \in \mathfrak{so}(n)$, the Cayley map is a conformal mapping, $Cay : \mathfrak{so}(n) \to \mathrm{SO}(n)$ defined by $Cay(W) = (I_n + W)(I_n - W)^{-1}$.

Using the Cayley mapping, we can define the Riemannian retraction map $\mathrm{Exp}_X : T_X\mathrm{St}(p,n) \to \mathrm{St}(p,n)$ by $\mathrm{Exp}_X(W) = \mathrm{Cay}(W)X$. Hence, given $X, Y \in \mathrm{St}(p,n)$ within a regular geodesic ball, i.e., the geodesic ball does not include the cut locus, of appropriate radius (henceforth, we always assume geodesic ball to be regular), we can define the unique geodesic between X and Y, denoted by $\Gamma_X^Y : [0,1] \to \mathrm{St}(p,n)$ as $\Gamma_X^Y(t) = \mathrm{Exp}_X(t\ \mathrm{Exp}_X^{-1}(Y))$. Also, we can define the distance between X and Y as $d(X,Y) = \sqrt{\langle \mathrm{Exp}_X^{-1}(Y), \mathrm{Exp}_X^{-1}(Y)\rangle_X}$. Note that, the Cayley map used is defined within a geodesic ball of appropriate radius.

The Space of trajectories and its geometry: Let $(\mathcal{M}, g)$ be a Riemannian manifold equipped with a Riemannian metric g [8]. Let d be the distance on $\mathcal{M}$ induced by g. We define a trajectory on $\mathcal{M}$ to be a path λ consists of a set of discrete points on $\mathcal{M}$. Let $\mathfrak{T}(\mathcal{M})$ be the set of trajectories. Before identifying the space of trajectories as a topological space (in fact as a Riemannian manifold), we first define the following statistics.

Definition 2.1. Fréchet mean (FM): *Given $\{X_i\}_{i=1}^N \subset \mathcal{M}$, define "the" Fréchet mean (FM) [2], M to be the minimizer of sum of squared geodesic distances, i.e., $M = \arg\min_\mu \sum_i d^2(\mu, X_i)$. FM exists and is unique within a regular geodesic ball of appropriate radius [9].*

Definition 2.2. PGA: *Given $\{X_i\}_{i=1}^N \subset \mathcal{M}$ and FM M, the principal vectors, $\{\mathbf{v}_i\} \subset T_M\mathcal{M}$ are recursively defined as $\mathbf{v}_i = \arg\max_{\|\mathbf{v}\|=1, \mathbf{v}\in V_{i-1}^\perp} \sum_{j=1}^N d^2(M, \Pi_{S_i}(X_j))$, where V_i is the subspace spanned by $\{\mathbf{v}_1, \cdots, \mathbf{v}_i\}$, and $S_i = Exp_M(span\{V_{i-1}, \mathbf{v}_i\})$. $\Pi_{S_i}(X_j)$ is the projection of X_j on S_i. The i^{th} principal geodesic submanifold (PG) given by the i^{th} principal vector, $\mathbf{v}_i$ is $Exp_M(\mathbf{v}_i)$.*

In the above definition of principal geodesic analysis (PGA), we are essentially searching for a direction along which the data variance is maximized. This is called "variance maximization" formulation of PGA [10, 11]. An alternative definition of PGA is given in by [12],

$$\mathbf{v}_i = \underset{\|\mathbf{v}\|=1, \mathbf{v}\in V_{i-1}^\perp}{\arg\min} \sum_{j=1}^N d^2(X_j, \Pi_{S_i}(X_j)), \qquad (1)$$

i.e., we are searching for a direction s.t. the projection of X_j along the direction is closed to X_j. This formulation is known as "reconstruction error minimization" formulation of PGA.

Given $\{\mathbf{x}_i\}$ on the Euclidean space, $\mathbf{R}^n$, the *SD line* [13] is defined to be a straight line, $\mathbf{l}$ through the mean of the points s.t., sum of the squared perpendicular distances from each $\mathbf{x}_i$ on $\mathbf{l}$ (MSE) is minimized. One can naturally generalize the notion of the *SD line* on a Riemannian manifold, we will denote it by *SD geodesic*.

Definition 2.3. SD geodesic: *Given $\{X_i\}_{i=1}^N \subset \mathcal{M}$ and FM M, the principal geodesic through M is defined to be the SD geodesic.*

Proposition 2.1. *SD geodesic defined above minimizes the sum of squared perpendicular distances from data points.*

Proof. It follows from the alternative "reconstruction error minimization" formulation of PGA. □

Given $\{X_i\}_{i=1}^N \subset \mathcal{M}$, we can recursively compute SD geodesics as follows:

Algorithm 1: Algorithm to compute the SD geodesics

1 $k \leftarrow 1$;
2 $\{\bar{X}_i^1\} \leftarrow \{X_i\}$;
3 **while** $k \leq K$ **do**
4 Compute $\mathbf{v}_k$ using Eq. 1 on $\{\bar{X}_i^k\}$;
5 k^{th} SD geodesic is given by the geodesic passing through M tangential to $\mathbf{v}_k$;
6 Project each $\bar{X}_i^k$ on the subspace orthogonal to the current geodesic subspace to get $\bar{X}_i^{k+1}$;
7 $k \leftarrow k+1$;
8 **end**

Definition 2.4. *Given a trajectory, γ, consisting of N points on a Riemannian manifold $\mathcal{M}$ of dimension m, identify γ with $(X_\gamma, Y_\gamma) \in \mathcal{M} \times St(K,m)$, where X_γ is the FM of the points of γ and i^{th} column of Y_γ consists of the tangent vector corresponding to i^{th} SD geodesics. We will denote this mapping from $\mathfrak{T}(\mathcal{M})$ to $\mathcal{M} \times St(K,m)$ by Φ.*

Observe that in the above definition columns of Y_γ are mutually orthogonal as they are principal vectors. Moreover, the product manifold $\mathcal{M} \times \mathrm{St}(K,m)$ is equipped with a product metric (distance) given by: $d((X_{\gamma_1}, Y_{\gamma_1}), (X_{\gamma_2}, Y_{\gamma_2})) = d(X_{\gamma_1}, X_{\gamma_2}) + d(Y_{\gamma_1}, Y_{\gamma_2})$. It is easy to see that the above identification is well-defined and using it along with the product Riemannian metric gives $\mathfrak{T}(\mathcal{M})$ a Riemannian manifold structure.

Proposition 2.2. *Φ is surjective. Moreover, Φ is injective* iff *K is sufficiently large to reconstruct γ.*

In the experimental section, we have used trajectories that lie on hypersphere. The projection operator used to get SD geodesics (in line 6 of Algorithm 1) is given in [10]. Further, an analytic expression for reconstruction of data points from principal vectors is given in [10].

3. EXPERIMENTAL RESULTS

In this section, we use the OASIS data [1] to automatically discriminate between demented (D) and non-demented (ND) patients using our proposed framework. This dataset contains at least two MR brain scans of each of the 150 subjects, aged between 60 to 96 years old. For each patient, scans are separated by at least one year. The dataset contains patients of both sexes. In order to avoid gender effects, we have taken MR scans of male patients alone from three visits, which resulted in the dataset containing 69 MR scans of 11 subjects with dementia and 12 subjects without dementia. We first construct an atlas from 36 "normal" MR scans, i.e., MR scans from subjects without dementia.

	ND	D
ND	10	2
D	2	9

(a) using FM

	ND	D
ND	11	1
D	1	10

(b) using PGA

Table 1: Confusion matrix of demented vs. non-demented classification

First, we rigidly register each MR scan to the atlas and then compute the displacement field from each MR scan after non-rigidly registering them to the atlas. Thus, from each patient, we obtain three displacement fields (collected from three visits) to get a path on the space of diffeomorphisms. It can be shown that the space of diffeomorphism group quotiented out by the normal subgroup of volume preserving diffeomorphisms is isomorphic to hypersphere. Hence, we use this identification map to the hypersphere (of appropriate dimension) from the space of diffeomorphisms (as was done in [14]). It is well-known that the geometry of the space of diffeomorphisms is complicated, hence we can perform statistical analysis on the hypersphere instead of on the space of diffeomorphisms after quotienting out the volume preserving diffeomorphisms. Now, for each patient, we have a trajectory on the hypersphere, in our experiments, $\mathbf{S}^{892}$. We then use two techniques to classify demented vs. non-demented patients. In Figure (2), the segmented corpus callosa for three visits of two subjects from the two classes (demented and non-demented) are shown.

We first compute the mean trajectory for each of the two classes and classify each data (trajectory) to the nearest mean trajectory (analogous to nearest neighbor classification technique) in a leave-one-out fashion. In the second technique, we also perform leave-one-out classification. For each class in the training set, we compute first principal geodesic, i.e., we compute the leading principal geodesics for each of the two classes using the exact-PGA for constant curvature manifolds [10]. Then, for a test sample, we project the sample on the principal geodesic and assign the sample to the class in which the projection error is minimum. The projection error is com-

puted as follows. Given $\mathbf{x} \in \mathbf{S}^{892}$, and the principal vector $\mathbf{v} \in T_{\mathbf{m}}\mathbf{S}^{892}$ ($\mathbf{m}$ is the FM), the projection error, $E(\mathbf{x}, \mathbf{v})$ is computed analytically using,

$$E(\mathbf{x}, \mathbf{v}) = \arccos(\mathbf{x}^t \bar{\mathbf{x}})$$

where,

$$\bar{x} = \cos\left(\arctan\left(\theta\right)\right)\mathbf{m} + \sin\left(\arctan\left(\theta\right)\right)\mathbf{v}/\|\mathbf{v}\|$$

where, $\theta = \frac{\mathbf{x}^t\mathbf{v}/\mathbf{v}^t\mathbf{m}}{\|\mathbf{v}\|}$.

We compare both of our proposed classification schemes in Table 1. From the confusion matrix in Table 1, using FM, we get the sensitivity, specificity and classification accuracy as 0.83, 0.82 and 82.61% respectively. On the other hand, using PGA, the sensitivity, specificity and classification accuracy are 0.92, 0.91 and 91.30% respectively. As evident and per expectation, the exact-PGA [10] based classification yields a better accuracy.

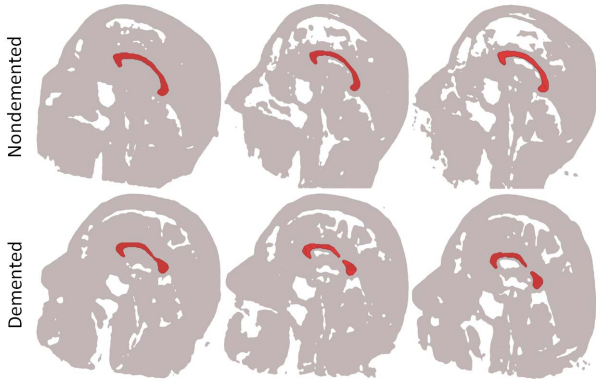

Fig. 2: Change in corpus callosa shapes in two classes

4. CONCLUSIONS

In this paper, we presented a novel formulation for performing statistical analysis of trajectories characterizing longitudinal data e.g., a time course of brain MR scans of dementia patients and others. The salient features of our formulation are, (i) it does not require the trajectory to lie on a geodesic, (ii) it does not require that the number of time points on each trajectory across the members in a group of trajectories be the same. Our novel formulation embeds the trajectories on manifold $\mathcal{M}$ in a product Riemannian manifold that is endowed with a Riemannian metric. The specific product manifold here is $\mathcal{M} \times \mathrm{St}(p, n)$. We presented algorithms to compute statistics on this product manifold and used it in conjunction with a nearest neighbor classifier to discriminate between subjects scans with and without dementia. Preliminary results presented here are quite encouraging and provide the impetus for further research.

5. REFERENCES

[1] "OASIS," `http://www.oasis-brains.org/`.

[2] Maurice Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'institut Henri Poincaré*, pp. 215–310, 1948.

[3] Jingyong Su, Anuj Srivastava, Fillipe DM de Souza, and Sudeep Sarkar, "Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition," in *CVPR*, 2014, pp. 620–627.

[4] Prasanna Muralidharan and P Thomas Fletcher, "Sasaki metrics for analysis of longitudinal data on manifolds," in *CVPR*, 2012, pp. 1027–1034.

[5] Yi Hong, Nikhil Singh, Roland Kwitt, and Marc Niethammer, "Group testing for longitudinal data," in *IPMI*, 2015, pp. 139–151.

[6] Alvina Goh and René Vidal, "Segmenting motions of different types by unsupervised manifold clustering," in *CVPR*, 2007, pp. 1–6.

[7] Tetsuya Kaneko, Simone Fiori, and Toshihisa Tanaka, "Empirical arithmetic averaging over the compact stiefel manifold," *IEEE TSP*, vol. 61, no. 4, pp. 883–894, 2013.

[8] Manfredo Perdigao do Carmo Valero, *Riemannian geometry*, 1992.

[9] Bijan Afsari, "Riemannian L^p center of mass: Existence, uniqueness, and convexity," *Proc. AMS*, vol. 139, no. 2, pp. 655–673, 2011.

[10] Rudrasis Chakraborty, Dohyung Seo, and Baba C. Vemuri, "An efficient exact-pga algorithm for constant curvature manifolds," `http://arxiv.org/abs/1603.03984/`, 2016.

[11] P.T. Fletcher, C. Lu, et al., "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE TMI*, pp. 995–1005, 2004.

[12] Stefan Sommer, François Lauze, Søren Hauberg, and Mads Nielsen, "Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations," in *Computer Vision–ECCV 2010*, pp. 43–56. Springer, 2010.

[13] Ronald Christensen, *Plane answers to complex questions: the theory of linear models*, Springer Science & Business Media, 2011.

[14] Dohyung Seo, Jeffrey Ho, and Baba C Vemuri, "Computing diffeomorphic paths for large motion interpolation," in *CVPR*, 2013, pp. 1227–1232.