STATISTICS ON THE SPACE OF TRAJECTORIES FOR LONGITUDINAL DATA ANALYSIS

Rudrasis Chakraborty[1], *Monami Banerjee*[1] *and Baba C. Vemuri*[1]

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

ABSTRACT

Statistical analysis of longitudinal data is a significant problem in Biomedical imaging applications. In the recent past, several researchers have developed mathematically rigorous methods based on differential geometry and statistics to tackle the problem of statistical analysis of longitudinal neuroimaging data. In this paper, we present a novel formulation of the longitudinal data analysis problem by identifying the structural changes over time (describing the trajectory of change) to a product Riemannian manifold endowed with a Riemannian metric and a probability measure. We present theoretical results showing that the maximum likelihood estimate of the mean and median of a Gaussian and Laplace distribution respectively on the product manifold yield the Fréchet mean and median respectively. We then present efficient recursive estimators for these intrinsic parameters and use them in conjunction with a nearest neighbor (NN) classifier to classify MR brain scans (acquired from the publicly available OASIS database) of patients with and without dementia.

Index Terms— Longitudinal studies, Trajectories, Manifolds.

1. INTRODUCTION

In this paper, we present a novel method to compute statistics on the space of manifold-valued trajectories. We will first give the space of trajectories a Riemannian manifold structure. It is very common in medical imaging applications to have data lie on a path residing on a Riemannian manifold, specifically longitudinal or time-series data. For example, it is often possible to identify patients with dementia by taking multiple magnetic resonance (MR) scans of the brain over time, assessing and comparing the structural changes from the norm in the corpus callosum. 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 identify the space of trajectories with a product space of two well-studied Riemannian manifolds and perform clustering on this product space. Further, we also present a intrinsic mean and median computation techniques on this space of trajectories. The salient features of our proposed method are (1) Both the mean and median computation algorithms are recursive and efficient. (2) We can recover the mean trajectory or the "trajectory atlas" since our identification of trajectory on the product manifold is a bijection. (3) Our proposed mean and median estimators are recursive, hence, in an online streaming of trajectory data, our method can compute the mean and median very effectively without storing the whole data. In the "toy" example of Figure (1), we show two trajectories on the sphere, $\mathbf{S}^2$, (shown in blue) and the mean of these two trajectories is shown in red. It is evident that the mean trajectory goes through the "midpoint" of the two trajectories. Thus, our goal for the rest of the paper is to define the space of trajectories and identify it with a known Riemannian manifold. Then, we generalize two well-known distributions, namely the Gaussian and the Laplace to this space. Since our identification maps the space of trajectories to a product of two homogeneous spaces, in order to define a distribution on the product space, we first generalize the two well-known distributions to a homogeneous space. Then, we sample from these distribution and state (proof not included due to space limitations) that the maximum likelihood estimator (MLE) of the location parameter of Gaussian distribution yields the Fréchet mean (FM) [2] of the samples, while the MLE of the location parameter of Laplace distribution yields the Fréchet median (FMe) [2] of the samples. Equipped with the tools to define distributions on the space of trajectories, we propose an efficient recursive estimator to compute FM and FMe on this space. Moreover, we claim that our proposed estimator is (weakly) consistent. Due to space limitations, we simply state the theorems without proof and will include them in a future publication. We use publicly available OASIS data [1] to perform classification of patients with and without dementia using the FM on the space of trajectories in conjunction with a nearest-neighbor (NN) classifier. Further, when the

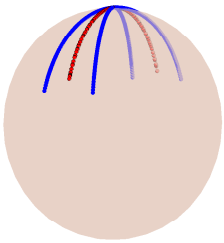

Fig. 1: Toy example

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

data are corrupted with outliers, we replace the FM with FMe the above classifier. Comparisons are presented between the proposed FM and FMe-based classification to depict the robustness of the FMe based scheme in the presence of outliers. 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 proposed a framework to compute "trajectory atlas" and registration of trajectories. They modeled each trajectory as a smooth curve on a Riemannian manifold. The method to compute the "trajectory atlas" involves an optimization problem, hence is not as computationally efficient as our proposed method, which does not involve any optimization. In [4], authors equipped the tangent bundle with a Sasaki metric and identify the longitudinal data as a point on the tangent bundle. In [5], authors performed a principal geodesic analysis (PGA) on the tangent bundle to achieve PGA of the longitudinal dataset. In [6], authors performed the segmentation of motion characterized by trajectories on a Riemannian manifold.

2. STATISTICS ON THE SPACE OF TRAJECTORIES

In this section, we will first briefly discuss the geometry of two homogeneous spaces, namely Stiefel and SPD(n) (space of $n \times n$ symmetric positive-definite matrices). *These two spaces will be needed later in our geometric formulation of the space of trajectories*. Thus, in the following paragraphs, we present a "bare bones" background differential geometry material needed to present our formulation and refer the reader to [7] for more details.

Let $(\mathcal{M}, \mathfrak{g})$ be a Riemannian manifold with a Riemannian metric $\mathfrak{g}$. Let d be the metric induced by the Riemannian metric $\mathfrak{g}$. Let G be the set of all isometries of $\mathcal{M}$, i.e., given $g \in G$, $d(g.X, g.Y) = d(X, Y)$, for all $X, Y \in \mathcal{M}$. Let $O \in \mathcal{M}$ and let $H = \text{Stab}(O) = \{h \in G | h.O = O\}$. We say G acts *transitively* on $\mathcal{M}$, iff given $X, Y \in \mathcal{M}$, there exists $g \in \mathcal{M}$ such that $Y = g.X$.

Definition 2.1. *Let M be a Riemannian manifold. Let $G = I(\mathcal{M})$ acts transitively on $\mathcal{M}$ and $H = Stab(O)$, $O \in \mathcal{M}$ (called the "origin" of $\mathcal{M}$) is a normal subgroup of G. Then, $\mathcal{M}$ is a homogeneous space and can be identified with the quotient space G/H under the diffeomorphic mapping $gH \mapsto g.O, g \in G$ [7].*

A compact Stiefel manifold, $\text{St}(p, n)$, is the set of all $(n \times p)$ dimensional column orthonormal real matrices, where $n \geq p$. $\text{St}(p, n)$ is a homogeneous space and can be identified with $SO(n)/SO(n-p)$, $SO(n)$ is the group of special orthogonal matrices. Given $X \in \text{St}(p, n)$, we can define an efficient Cayley type Riemannian retraction and lifting map within an open neighborhood of X as defined in [8]. SPD(n) is the space of $n \times n$ symmetric positive definite matrices. SPD(n) is a homogeneous space and can be identified with $GL(n)/O(n)$, $O(n)$ is the group of orthogonal matrices. For the Exponential and Inverse Exponential maps on SPD(n), we refer the readers to [9].

From the definition of a homogeneous space, we know that the Riemannian metric $\mathfrak{g}$ at X is invariant under the group operation $X \mapsto g.X$, hence the volume element $d\nu$ is also preserved.

Proposition 2.1. *Let $F : \mathcal{M} \to \mathbf{R}$ be an integrable function. Then, $\int F(g.X) d\nu(g.X) = \int F(X) d\nu(X)$.*

Gaussian and Laplace distribution: Now, we will define the Gaussian and Laplace distributions on a homogeneous space with the following density functions with respect to $d\nu$ as follows, here $\sigma, b > 0$. These distributions will be required subsequently in the statement of the theorem on *maximum likelihood estimation*.

$$\textbf{(Gaussian)}\ f_X(M, \sigma) = \exp(-\frac{d^2(X, M)}{2\sigma^2})/Z(\sigma) \quad (1)$$

$$\textbf{(Laplace)}\ f_X(M, b) = \exp(-\frac{d(X, M)}{\sigma})/Z(b) \quad (2)$$

Theorem 2.1. *The normalizing constants in Eqs. (1, 2), i.e., $Z(M, \mu) = \int f_X(M, \sigma)\, d\nu(X)$ and $Z(M, b) = \int f_X(M, b)\, d\nu(X)$ are constants and independent of M, i.e., the functions in Eqs. (1, 2) are valid probability densities.*

Due to limited space, we will provide the proof of this theorem in a subsequent journal version of this paper. Given a set of N samples, $\{X_i\}_{i=1}^N$, the Fréchet mean (FM) [2], M is defined as $M^* = \arg\min_Y \sum_{i=1}^N d^2(X_i, Y)$. The Fréchet median (FMe) [2] on a set of N samples, $\{X_i\}_{i=1}^N$, is defined as $\widetilde{M} = \arg\min_Y \sum_{i=1}^N d(X_i, Y)$. For the rest of the paper, we will assume that the samples lie within a regular geodesic ball of appropriate radius so that both FM and FMe exist and are unique [10].

Theorem 2.2. *(a) Given a set of i.i.d samples $\{X_i\}_{i=1}^N \in \mathcal{M}$ drawn from the Gaussian distribution on $\mathcal{M}$, the maximum likelihood estimate (MLE) of M is the Fréchet mean (FM) of all the samples.*

(b) Given a set of i.i.d samples $\{X_i\}_{i=1}^N$ drawn from the Laplace distribution, the MLE of M is the Fréchet median (FMe) of all the samples.

Proof of this theorem will be included in a subsequent journal version of this paper. We will now give an efficient recursive FM and FMe estimator, M_k (we denote both FM and FMe by M_k with a slight abuse of notation) on $\text{St}(p, n)$ and SPD(n).

$$M_1 = X_1, \qquad M_k = \text{Exp}_{M_{k-1}}(v_k/k) \quad (3)$$

where, $v_k = \text{Exp}^{-1}_{M_{k-1}}(X_k)$ for FM and for FMe, $v_k = \text{Exp}^{-1}_{M_{k-1}}(X_k)/d(M_{k-1}, X_k)$. Here $\{X_i\} \subset \text{St}(p, n)$ or

$\{X_i\} \subset \text{SPD}(n)$. The proof of consistency of this FM estimator on $\text{St}(p,n)$ will be given in the companion journal version. The proof of consistency of FM estimator on $\text{SPD}(n)$ is given in [9]. We refer the readers to [11] for the proof of consistency of FMe estimator on any Riemannian manifold.

The Space of trajectories and its geometry: We define a trajectory, γ to be a path (consists of a discrete set of points) on a Riemannian manifold $\mathcal{M}$. For the rest of this section, we will identify the space of trajectories on a Riemannian manifold, denoted by $\mathfrak{T}(\mathcal{M})$, with a Riemannian manifold and define statistics on $\mathfrak{T}(\mathcal{M})$ using the theory discussed above.

Definition 2.2. *Given that a trajectory, α, consists of a set of p points, on a Riemannian manifold $\mathcal{M}$, identify α with a matrix, A, of dimension $n \times p$, where n is the dimension of the ambient space (isomorphic to $\mathbf{R}^n$) of $\mathcal{M}$. Without any loss of generality, we assume that $n \geq p$ and moreover, each point on the path, i.e., each column of A, are i.i.d. samples drawn from a probability distribution on $\mathbf{R}^n$. Now, define a map $\Psi : \mathfrak{T}(\mathcal{M}) \to St(p,n) \times SPD(n)$ as $A \mapsto (Q,P)$ where $P = R^T R$, $QR = A$ is the qr-decomposition of A [12]. Here, $St(p,n) \times SPD(n)$ is the product manifold of $St(p,n)$ and $SPD(n)$ with product metric defined as: $d((Q_1,P_1),(Q_2,P_2)) = d(Q_1,Q_2) + d(P_1,P_2)$, here $Q_1, Q_2 \in St(p,n)$, $P_1, P_2 \in SPD(n)$.*

Proposition 2.2.

(a) Ψ is a well-defined map.

(b) Ψ is a bijection.

The proposition can be proved using the uniqueness of qr-decomposition for full column rank matrices and uniqueness of Cholesky factorization of symmetric positive definite (SPD) matrices respectively. Now, as we can identify $\mathfrak{T}(\mathcal{M})$ with the product manifold of $\text{St}(p,n)$ and $\text{SPD}(n)$ using Ψ. W can then define the Gaussian and the Laplace distributions on $\mathfrak{T}(\mathcal{M})$ using Eqs. (1, 2) with respect to the product measure. Moreover, given a set of N i.i.d. samples drawn from a Gaussian (Laplace) distribution, we can define the FM and FMe using the product metric as defined above. We can also define the FM and FMe estimators as given by Eq. (3), for this product manifold.

3. EXPERIMENTAL RESULTS

In this section, we use OASIS data [1] to address the classification of demented (D) vs. non-demented (ND) patients using our proposed framework. This dataset contains at least two MR brain scans of 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 compute an atlas (using the method in [13]) from the $36 (= 12 \times 3)$ MR scans of patients without dementia.

	ND	D
ND	10	2
D	1	10

(a) using FM

	ND	D
ND	9	3
D	2	9

(b) using eFM

Table 1: Confusion matrix (without outliers)

After rigidly registering each MR scans to the atlas, we compute the displacement field of each MR scan required to non-rigidly register them to the atlas. So, from each patient, we get three displacement fields (collected from three visits) to get a path on the space of diffeomorphisms. As the geometry of this space is complicated, we quotient out the volume preserving diffeomorphisms to map each displacement field on to a hypersphere of dimension 892 (as was done in [14]). Now, for each patient, we have a trajectory on the hypersphere $\mathbf{S}^{892}$. We then 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. We compare our proposed FM and FMe with *extrinsic FM (eFM)* and *extrinsic FMe (eFMe)* [15] respectively. The confusion matrix [16] is given in Table (1). *Using FM (eFM), the sensitivity, specificity and classification accuracy are* 0.83 *(*0.67*),* 0.91 *(*0.82*) and* 86.96% *(*73.91%*) respectively*. In Figure (2), the segmented corpus callosa for three visits of two subjects from the two classes (demented and non-demented) are shown.

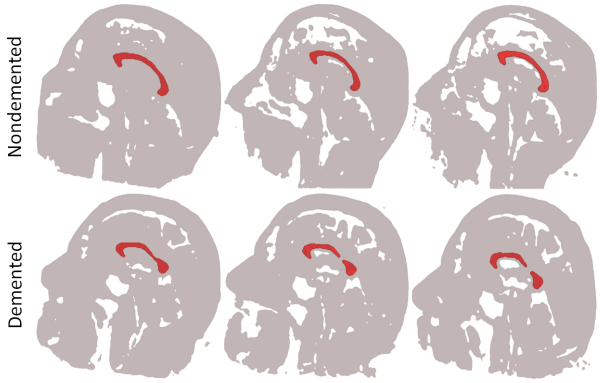

Fig. 2: Change in corpus callosa shapes in two classes

	ND	D
ND	9	3
D	1	10

(a) using FM

	ND	D
ND	8	4
D	2	9

(b) using eFMe

	ND	D
ND	10	2
D	1	10

(c) using FMe

Table 2: Confusion matrix (10% outliers)

This OASIS data also has some samples from the "converted" group, i.e., converted/"cured" (after treatment) from non-demented to demented during the study. We have used

	ND	D
ND	10	2
D	4	7

(a) using FM

	ND	D
ND	9	3
D	6	5

(b) using eFMe

	ND	D
ND	10	2
D	2	9

(c) using FMe

Table 3: Confusion matrix (20% outliers)

this "converted" group as an outlier in our original two class data. In the "converted" group, there are 7 patients having at least three visits. We have added 10% and 20% outliers to the data, i.e., added 2 and 4 samples from the converted group to the original data. We now do the leave-one-out classification on the outlier corrupted data using both FM and FMe. The comparison results are shown in Table (2) (10% outlier) and (3) (20% outlier). *The sensitivity, specificity, accuracy for FM with* 10% *outliers are* 0.75*,* 0.91*, and* 82.61% *respectively. Whereas, when using eFMe (FMe) we get these values to be* 0.67 *(*0.83*),* 0.82 *(*0.91*) and* 73.91% *(*86.96%*) respectively*. Note that, on data with 10% outliers, FMe performs as good as FM (without outliers). Thus confirming the robustness of FMe. On data corrupted with 20% outliers, using FM (FMe), the sensitivity, specificity and accuracy values are 0.83 (0.83), 0.64 (0.82), 73.91% (82.61%). Whereas, using eFMe the respective values are 0.75, 0.45 and 60.86% respectively. The comparative performance analysis of FM, FMe and eFMe clearly indicate the superior performance of FMe over both FM and eFMe for data with outliers.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented a novel framework to compute statistics on the space of trajectories, which was given a Riemannian product manifold structure. Further, we defined the Gaussian and the Laplace distributions on this space. This was achieved by defining these distributions on a Riemannian homogeneous space. Then, we sampled from these distributions and claimed that the maximum likelihood estimator (MLE) of the location parameter of the Gaussian and the Laplace distributions respectively are the Fréchet mean (FM) and Fréchet median (FMe). Further, efficient recursive estimators for computing the FM and the FMe on the space of trajectories were presented. The usefulness of our estimators was shown by applying them in conjunction with a nearest-neighbor classifier to the classification of demented vs. non-demented patient scans acquired from the OASIS database [1]. The classification results have shown improved performance of FMe over FM on outlier corrupted data. Our future work will focus on the computation of the variance and principal geodesic analysis (PGA) on this space of trajectories.

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] Sigurdur Helgason, *Differential geometry, Lie groups, and symmetric spaces*, vol. 80, 1979.

[8] 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.

[9] Jeffrey Ho, Guang Cheng, Hesamoddin Salehian, and Baba C Vemuri, "Recursive karcher expectation estimators and geometric law of large numbers.," in *AISTATS*, 2013, pp. 325–332.

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

[11] Silvere Bonnabel, "Stochastic gradient descent on riemannian manifolds," *IEEE TAC*, vol. 58, no. 9, pp. 2217–2229, 2013.

[12] Gene H Golub and Charles F Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.

[13] Brian B Avants, Nick Tustison, and Gang Song, "Advanced normalization tools (ANTS)," *Insight J*, vol. 2, pp. 1–35, 2009.

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

[15] Rabi Bhattacharya and Vic Patrangenaru, "Large sample theory of intrinsic and extrinsic sample means on manifolds. i," *Annals of statistics*, pp. 1–29, 2003.

[16] Stephen V Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.