Group Testing for Longitudinal Data

Yi Hong[1], Nikhil Singh[1], Roland Kwitt[2], and Marc Niethammer[1]

[1] Department of Computer Science, UNC Chapel Hill, NC, USA
[2] Department of Computer Science, University of Salzburg, Austria

Abstract. We consider how to test for group differences of shapes given longitudinal data. In particular, we are interested in differences of *longitudinal models* of each group's subjects. We introduce a generalization of principal geodesic analysis to the tangent bundle of a shape space. This allows the estimation of the variance and principal directions of the distribution of *trajectories* that summarize shape variations within the longitudinal data. Each trajectory is parameterized as a point in the tangent bundle. To study statistical differences in two distributions of trajectories, we generalize the Bhattacharyya distance in Euclidean space to the tangent bundle. This not only allows to take second-order statistics into account, but also serves as our test-statistic during permutation testing. Our method is validated on both synthetic and real data, and the experimental results indicate improved statistical power in identifying group differences. In fact, our study sheds new light on group differences in longitudinal corpus callosum shapes of subjects with dementia versus normal controls.

Keywords: Longitudinal data; distribution of trajectories; tangent bundle; group testing; Bhattacharyya distance

1 Introduction

Longitudinal data designs frequently arise in medical research that involves repeated measurements during follow-up studies. Analysis of such longitudinal data often involves constructing statistical models to summarize growth, aging and disease progression over time. For example, longitudinal studies in new-borns and young children use imaging at multiple follow-up visits to understand the process of early brain development [6]. Similarly, recent collective efforts have enabled longitudinal data collection to facilitate the study of neurodegeneration due to aging and age-related neurological disorders, such as the Alzheimer's disease [11]. Conventional cross-sectional models of regression that do not take into account the temporal dependencies of measurements are inappropriate for modeling such longitudinal data designs.

Recent methods for analyzing longitudinal, manifold-valued data have enabled modeling and even detection of changes over time [4,7,13]. These methods allow for the estimation of trajectories, *i.e.*, smooth paths estimated from the longitudinal data of subjects. Building upon these methods, Riemannian approaches for computing averages of trajectories have been proposed [12,16]. The

registration and comparison of trajectories has also been studied in [3,17,18]. In general, statistical methods for longitudinal manifold-valued data focus on first-order statistics, such as computing the mean, which only captures limited information of the data distribution. Capturing higher-order statistics on the trajectories themselves would be useful for a more comprehensive description of the underlying distributions and for designing test-statistics that go beyond a simple comparison of means; an example would be testing differences in variances.

Motivated by this, we develop an approach that leverages second-order statistics of shape trajectories for group testing. In particular, we propose a generalization of principal component analysis (PCA) and principal geodesic analysis (PGA) [5] to the tangent bundle [9] of a shape space. Similar to PCA/PGA, the first principal direction characterizes the dominant variability in a *population of trajectories*, and each point along this principal direction is a trajectory. This differs from previous studies which have focused on computing averages on the tangent bundle. Incorporating second-order statistics additionally allows to identify differences between groups of trajectories in situations where the average longitudinal trend over time is similar (or equal) between two groups. We refer to this approach as *principal geodesic analysis on the tangent bundle.*

Contribution. We extend principal geodesic analysis to the tangent bundle of a shape to estimate both variance and principal directions of shape trajectories. We then introduce a generalization of the Bhattacharyya distance to manifold-valued data, which enables the assessment of statistical differences between groups of trajectories. We validate our approach on both synthetic and real shapes. The results indicate improved statistical power in distinguishing groups with different distributions, especially for cases with similar means but different variances.

Organization. The paper is organized as follows: Section 2 discusses the basic principles of our approach in Euclidean space. Section 3 then generalizes these concepts to manifolds and discusses group testing on the tangent bundle. Section 4 presents our experimental study and Section 5 concludes the paper with a summary of the main points, open problems and an outlook on future work.

2 Distribution of trajectories in Euclidean space

We first illustrate the concept of analyzing populations of trajectories in Euclidean space, which is a trivial case of a Riemannian manifold.

Consider the case of two groups of subjects such that each subject is measured at multiple points in time. Such a data configuration is also referred to as a *staggered longitudinal design*, see Fig. 1(b). If we ignore the within-subject correlations and model the data with a cross-sectional design, illustrated in Fig. 1(a), the two groups cannot be separated using statistical tests that rely on a comparison of means only (*cf*. Table 1). Hence, to leverage longitudinal information, we first estimate linear regression models on each subject to summarize its trend. The regression line, a smooth trajectory approximating a subject's data points, is parameterized the tuple of *slope* and *intercept*, which can be represented as a point in the space of trajectories. As shown in Fig. 1(c), representing the data

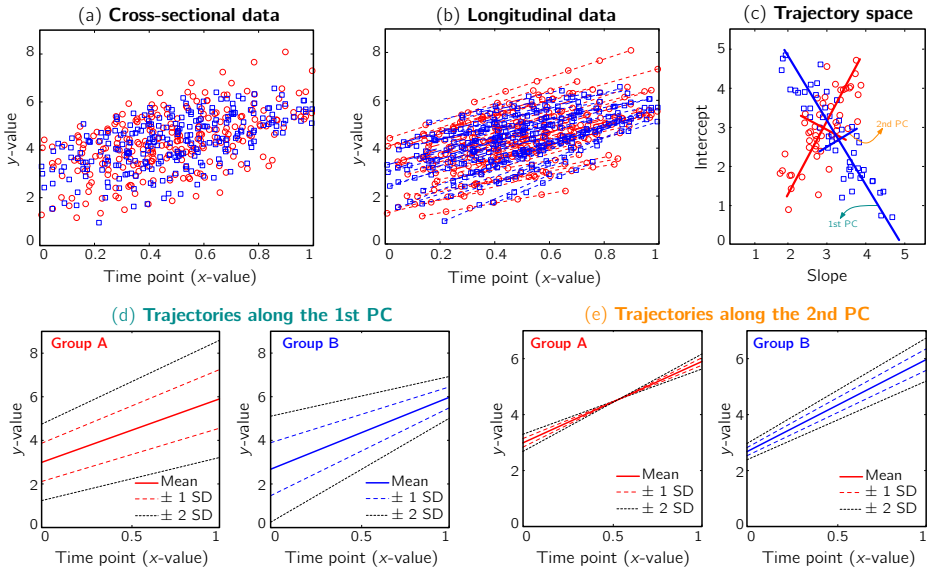


Fig. 1: A toy example in Euclidean space. *Top*: (a) Cross-sectional data of two groups, illustrated as red circles and blue squares; (b) the same data *with* longitudinal information (middle) where points on the same line are observations from one subject; (c) the trajectory space, represented by a slope and an intercept. Every point in this space corresponds to a straight line in (b). *Bottom*: (d) Trajectories generated by points along the 1st principal component (PC) of standard PCA in trajectory space with $\{0, \pm 1, \pm 2\}$ standard deviations (SD); (e) trajectories generated along the 2nd PC (best-viewed in color).

in this trajectory space separates the populations (at least visually) in this example. In fact, Table 1 indicates that including longitudinal information allows us to identify differences between the two groups statistically.

To further analyze the group differences, we explore the distribution of trajectories within the (slope, intercept) space, *i.e.*, the trajectory space. Under a Gaussian assumption, principal component analysis (PCA) is a standard tool to estimate the variance and principal directions of a sample. By applying PCA to (slope, intercept) data, we obtain a representation of the population of trajectories, namely their variances and their principal components. For example, the solid lines with different colors in Fig. 1(c) show the principal components of the two groups, respectively. By moving along these two principal components, we generate new points in the trajectory space such that each point represents a straight line in the original space of the data points. Figures 1(d) and (e) visualize the trajectories along the principal components for different standard deviations. The five trajectories in Figure 1(d), for instance, show the five points along the first principal component in the trajectory space for each group. This

	Cross-sectional data			Longitudinal data		
	$\bar{D}_E$	$\bar{D}_M$	D_B	$\bar{D}_E$	$\bar{D}_M$	D_B
Distance	0.0003	0.0047	0.0077	0.2438	0.3332	0.6722
p-value	0.9232	0.7487	0.1249	0.0347	0.0186	**1e-4**

Table 1: Distances and estimated p-values (10000 random permutations) on toy data using (1) the mean difference in Euclidean space ($\bar{D}_E$), (2) the Mahalanobis distance ($\bar{D}_M$), and (3) the Bhattacharyya distance (D_B) as a test-statistic.

Euclidean case illustrates that the proposed approach is a potentially useful tool in the analysis of longitudinal time-varying data.

Bhattacharyya distance. Visualization of trajectories along principal directions can qualitatively demonstrate differences between groups. However, to quantitatively assess the differences, we need a suitable distance measure that serves as a test-statistic. An appropriate candidate for this is the Bhattacharyya distance [1], which measures the similarity of two probability distributions. Given two multivariate Gaussians, with means (μ_1, μ_2) and covariance matrices (Σ_1, Σ_2), the Bhattacharyya distance D_B has the closed-form expression

$$D_B((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \frac{1}{8}(\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)^\top + \frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}\right), \quad (1)$$

where $\Sigma = (\Sigma_1 + \Sigma_2)/2$, and $|\cdot|$ denotes the matrix determinant. The first term in Eq. (1) measures the separability of the distributions w.r.t. their means. It is related to the squared Mahalanobis distance [10], which can be considered a special case of Eq. (1) when the difference between the covariances (as measured by the second term in the summation) is not considered. This additional term makes D_B more suitable, compared to the Mahalanobis distance, in cases where the distributions differ in variances. In particular, the Mahalanobis distance is zero when two distributions have equal means. However, as D_B only satisfies three conditions of a distance metric (non-negativity, identity of indiscernibles, and symmetry), but lacks the triangle inequality, it is only a semi-metric.

In fact, Eq. (1) allows us to compute a distance between the two distributions (assuming Gaussianity) in Fig. 1(c), and thereby to define a test-statistic to test for group differences in a permutation testing setup. The null-hypothesis H_0 of the permutation test is that the two distributions (say P, Q) to be tested are the same, *i.e.*, $H_0 : P = Q$. We estimate the empirical distribution of the test-statistic under H_0 by repeatedly permuting the group labels of the points in Fig. 1(c), and re-computing D_B between the two groups that result from the permuted labels. The p-value under H_0 then is the proportion of the area under the empirical distribution of samples for which the distance is less than the one estimated for the original (unpermuted) label assignments. In Table 1, D_B, tested on the longitudinal data, exhibits the best performance in separating the groups with an estimated p-value of <1e-4 under 10000 permutations.

3 Distribution of trajectories on manifolds

To explore the distribution of trajectories for manifold-valued data, *e.g.*, images or shapes, we need to generalize the statistical test of the previous section from Euclidean space to manifolds. Specifically, let $\{P_{i,j,k}\}$ be a population of longitudinal data on the same manifold, where i is the group identifier, j is the subject identifier, and k identifies the time point. Further assume we have N groups: group i has S_i subjects ($i = 1, \ldots, N$), and each subject has multiple time points, $\{t_{i,j,k}\}, k = 1, \ldots, T_{i,j}$. Our objective is to characterize the distribution of trajectories for each group, $\{D_i\}$, *i.e.*, to estimate its variance and principal directions, and to assess whether two groups are significantly different.

Individual trajectories for longitudinal data. To perform statistical tests on subjects with associated longitudinal data, our first step is to summarize the variations within a subject as a smooth trajectory. The parametric geodesic regression approaches for data in Kendall's shape space [4], or images [13,7], which generalize linear regression in Euclidean space, provide a compact representation of the continuous trajectory for each subject. The trajectory of subject j from group i is parametrized by the initial point $\hat{p}_{i,j}$ and the initial velocity $\hat{u}_{i,j}$. This trajectory minimizes the sum-of-squared geodesic distances between the observations and their corresponding points on the trajectory, *i.e.*,

$$(\hat{p}_{i,j}, \hat{u}_{i,j}) = \underset{(p_{i,j}, u_{i,j})}{\arg\min} \sum_{k=1}^{T_{i,j}} d_g^2(\mathrm{Exp}(p_{i,j}, t_{i,j,k} \cdot u_{i,j}), P_{i,j,k}) \ , \tag{2}$$

where $d_g(\cdot, \cdot)$ is the geodesic distance and $\mathrm{Exp}(\cdot, \cdot)$ denotes the exponential map on some manifold $\mathcal{M}$ [4]. This compact representation, $(\hat{p}_{i,j}, \hat{u}_{i,j})$, is a point in the tangent bundle $\mathcal{TM}$ of $\mathcal{M}$. $\mathcal{TM}$ is also a smooth manifold, which can be equipped with a Riemannian metric, such as the *Sasaki metric* [15]. Since each subject's longitudinal data is represented as a point on $\mathcal{TM}$, we work in this space, instead of the space of the data points, to perform group testing.

Principal geodesic analysis (PGA) for trajectories. We generalize principal geodesic analysis to estimate the variance and the principal directions of trajectories on the tangent bundle for each group. We follow the definitions of the exponential- and the log-map on $\mathcal{TM}$ in [12] and use the Sasaki metric. Specifically, given two points $(p_1, u_1), (p_2, u_2) \in \mathcal{TM}$, the log-map outputs the tangent vector such that $(v, w) = \mathrm{Log}_{(p_1,u_1)}(p_2, u_2)$. The exponential map enables us to shoot forward with a given base point and a tangent vector, *i.e.*, $(p_2, u_2) = \mathrm{Exp}_{\mathcal{TM}}((p_1, u_1), (v, w))$. Furthermore, using the log-map, the geodesic distance on $\mathcal{TM}$ can be computed as $d_{\mathcal{TM}}((p_1, u_1), (p_2, u_2)) = \| \mathrm{Log}_{(p_1,u_1)}(p_2, u_2) \|$.

Before computing the variance and the principal directions, we first need to estimate the mean of the trajectories for each group. This is done by minimizing the sum-of-squared geodesic distances, for each group, on $\mathcal{TM}$ as

$$\forall i : (\bar{p}_i, \bar{u}_i) = \underset{(p_i, u_i)}{\arg\min} \sum_{j=1}^{S_i} d_{\mathcal{TM}}^2((p_i, u_i), (\hat{p}_{i,j}, \hat{u}_{i,j})) \ . \tag{3}$$

Then, following the PGA algorithm of [5], we compute the variance and principal directions w.r.t. the estimated mean of the trajectories. Specifically, we first compute the tangent vector from the mean of group i to the trajectory of its subject j, $(v_{i,j}, w_{i,j}) = \mathrm{Log}_{(\bar{p}_i,\bar{u}_i)}(\hat{p}_{i,j}, \hat{u}_{i,j})$ and then calculate the covariance matrix $\Sigma_i = \frac{1}{S_i-1}\sum_{j=1}^{S_i}(v_{i,j}, w_{i,j})(v_{i,j}, w_{i,j})^\top$. The principal decomposition of Σ_i results in the eigenvalues $\lambda_{i,q} \in \mathbb{R}_0^+$ and eigenvectors $(v_{i,q}, w_{i,q}) \in \mathcal{T}_{(\bar{p}_i,\bar{u}_i)}\mathcal{M}$ with $q = 1, \ldots, Q_i$ for group i. As a result, we can identify the distribution of trajectories for each group by $D_i = \{(\bar{p}_i, \bar{u}_i), \Sigma_i\}$ with $i = 1, \ldots, N$. By moving along a principal direction, we can generate points on $\mathcal{TM}$, which correspond to trajectories on the manifold of the data points.

Generalized Bhattacharyya distance. Since we can characterize the distribution of trajectories on $\mathcal{TM}$ for each group, to measure the distance between them, we generalize the Bhattacharyya distance from Euclidean space to $\mathcal{TM}$. Again, the distribution D_i on $\mathcal{TM}$, is identified by a mean $\mu_i = (\bar{p}_i, \bar{u}_i) \in \mathcal{TM}$, and a covariance matrix Σ_i with respect to the mean μ_i.

Generalizing the first term of the Bhattacharyya distance in Eq. (1), *i.e.*, the pooling of covariance matrices $\Sigma = (\Sigma_1 + \Sigma_2)/2$, is not as straightforward on $\mathcal{TM}$ as it is in Euclidean space because the covariance matrices Σ_1 and Σ_2 of the two groups reside in tangent spaces at different points on $\mathcal{TM}$. Hence, we follow the strategy in [12], and replace the first term with the average of two squared-Mahalanobis distances, *i.e.*, $(\mathrm{Log}_{\mu_1}\mu_2\Sigma_1^{-1}\mathrm{Log}_{\mu_1}\mu_2{}^\top + \mathrm{Log}_{\mu_2}\mu_1\Sigma_2^{-1}\mathrm{Log}_{\mu_2}\mu_1{}^\top)/2$. Furthermore, because most manifold-valued data in medical applications is high dimensional and low sample size, the resulting covariance matrix is usually semi-positive-definite (SPD) with zero eigenvalues. This means that in many applications Σ_1 and Σ_2 are not invertible[3]. To address this issue, we approximate the covariance matrix via eigen-decomposition by dropping the eigenvalues that are smaller than a cutoff value, ϵ[4]. In this way, the covariance matrix can be decomposed approximately as $\Sigma_i \approx U_{i,Q_i}\Lambda_{i,Q_i}U_{i,Q_i}^\top$, where $\lambda_{i,q} < \epsilon$ if $q > Q_i$, resulting in $\Sigma_i^{-1} \approx U_{i,Q_i}\Lambda_{i,Q_i}^{-1}U_{i,Q_i}^\top$ [14].

To generalize the second term of the Bhattacharyya distance, which involves the computation of the determinant of a covariance matrix, we use the pseudo-determinant, *i.e.*, the product of all non-zero eigenvalues of a square matrix. For consistency, the same number of eigenvalues as for the first term is used, *i.e.*, $|\Sigma_i| = \prod_{q=1}^{Q_i}\lambda_{i,q}$. Since it is non-trivial to compute the pooled covariance matrix Σ, we replace its determinant in Eq. (1) with the averaged determinants of Σ_1 and Σ_2. While this changes the original definition of the Bhattacharyya distance, its properties are kept (see Appendix A). Also, it can be shown that the value of the second term increases as the difference in the determinants gets larger. Hence, the generalized second term can serve as a distance measure of generalized variances of covariance matrices on $\mathcal{TM}$. In summary, we define the

[3] A better estimate of the covariance matrix may be obtained, *e.g.*, by using [8] or [2].
[4] The threshold ϵ varies with the application. In our experiments, we set it to 1e-6. Usually, the eigenvalues larger than ϵ cover almost 99% of the variances.

generalized Bhattacharyya distance between two Gaussians D_1, D_2 on $\mathcal{TM}$ as

$$D_B^{\mathcal{TM}}(D_1, D_2) = \frac{1}{16}(D_M^{\mathcal{TM}}(\mu_1, D_2) + D_M^{\mathcal{TM}}(\mu_2, D_1)) + \frac{1}{2}\ln\left(\frac{(|\Sigma_1| + |\Sigma_2|)}{2\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}\right) \quad (4)$$

where $D_M^{\mathcal{TM}}$ is a generalized version of the squared Mahalanobis distance, given by $D_M^{\mathcal{TM}}(\mu_i, D_j) = \langle \mathrm{Log}_{\mu_j} \mu_i, U_{j,Q_j} \rangle \Lambda_{j,Q_j}^{-1} \langle \mathrm{Log}_{\mu_j} \mu_i, U_{j,Q_j} \rangle^\top$, and $\langle \cdot, \cdot \rangle$ is the inner product on the tangent bundle. $D_B^{\mathcal{TM}}$ is a *pseudo-semimetric*, *i.e.*, it satisfies (1) non-negativity, (2) symmetry, and (3) $D_B^{\mathcal{TM}}(D_i, D_i) = 0$ for all D_i (required for the identity of indiscernibles); see Appendix A for a detailed proof of these properties. As shown in the proof, although Eq. (4) does not satisfy the positivity property, *i.e.*, for all $D_1 \neq D_2$, $D_B^{\mathcal{TM}}(D_1, D_2) > 0$, only the distance between two distributions with equal mean *and* generalized variance is zero. Consequently, we can distinguish two distributions of trajectories that have different means and/or different determinants of the covariance matrices.

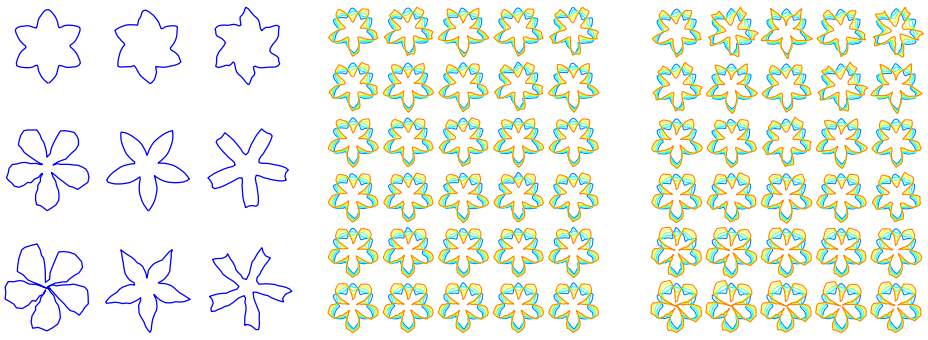
We use Eq. (4) as our test-statistic in the same permutation testing setup as described in Section 2. The null-hypothesis H_0 is that the samples of trajectories from the two groups were drawn from the same underlying distribution. The distribution of test-statistics under H_0 is estimated by randomly permuting the group label assignments. We then count the number of times that the distance is larger than the one computed without permutation to obtain a p-value estimate. Compared to the Hotelling T^2 statistic used in [12], which tests for difference in sample means (based on the squared Mahalanobis distance), our permutation test is based on Eq. (4), which is more appropriate in situations where two distributions have similar means but different variances.

4 Experiments

We demonstrate our method on (1) a toy example in Euclidean space, (2) a 2D example with synthetic shapes, and (3) real corpus callosum shapes. All shapes are represented in (2D) Kendall's shape space.

Toy example in Euclidean space. Fig. 1 shows the generated toy data and the qualitative comparison between two groups using PCA in the trajectory space. Both groups have 50 subjects each, measured at 3 to 7 time points. Table 1 reports the quantitative comparison, *i.e.*, permutation testing with 10000 permutations and three different distances: the Euclidean distance $\bar{D}_E$ (*i.e.*, the squared mean differences), the Mahalanobis distance $\bar{D}_M$ (*i.e.*, the squared mean difference based on the pooled covariance matrix), and the Bhattacharyya distance D_B. The results of the cross-sectional *vs.* longitudinal tests indicate that leveraging the longitudinal information greatly improves our ability to identify differences, as indicated by low p-values. Besides, among the three evaluated distance measures, the Bhattacharyya distance most clearly highlights this difference with a p-value of <1e-4 (given the number of permutations).

Synthetic shapes in Kendall's shape space. To verify the advantage of the generalized Bhattacharyya distance over the generalized Mahalanobis distance,



(a) Basic shapes **(b)** Group A **(c)** Group B

Fig. 2: Synthetic shapes: (a) Basic shapes used to generate the population on the right; (b) and (c) show the two groups of trajectories (best-viewed in color).

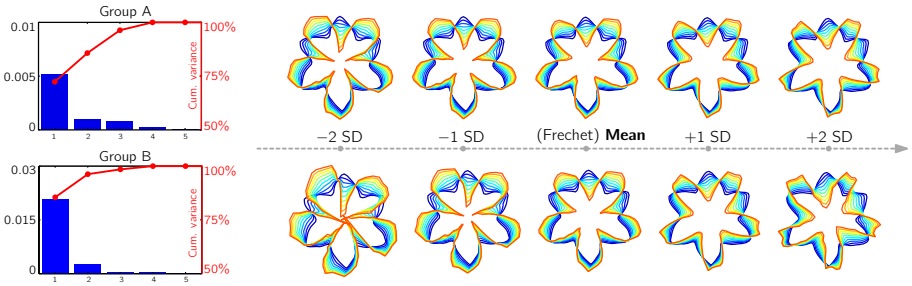


Fig. 3: Visualization of the variances (left) and principal directions (right) of trajectory distributions for the synthetic data (best-viewed in color).

we generate two groups of 2D shapes with similar mean trajectories but different variances, see Figs. 2(b) and 2(c). Hence, the distributions are different by design. In particular, we use the three shapes in the first row of Fig. 2(a) to uniformly sample 60 shapes within the triangle region in Kendall's shape space, spanned by the three shapes[5]. We call them the *base shapes*. In the same way, the shapes in the second and third row are used to sample 30 shapes each; we refer to these shapes as the *target shapes*. In summary, we have 60 base shapes from the same distribution and two groups of target shapes from two different distributions. By splitting the 60 base shapes into two subsets of equal size and connecting each base shape with one target shape (via a geodesic), we obtain 30 trajectories per group. Assuming every base shape is at time 0 and every target shape is at time 1, we sample 5 shapes along each trajectory to represent one subject. To make

[5] We use two geodesics to connect three given shapes and uniformly sample points on these two geodesics. Then, by connecting opposing points, we obtain new geodesics which are located within the triangle region to sample a population of shapes.

	$(\hat{p}, \hat{u})$		$(\hat{p}, 0)$		$(0, \hat{u})$	
	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$
Distance on $\mathcal{TM}$	0.7212	2.2833	0.0232	0.0152	0.7439	2.3057
p-value	0.1817	**0.0234**	0.8486	0.6801	0.1650	**0.0297**

Table 2: Distances and estimated p-values (10000 random permutations) on synthetic shapes using the averaged Mahalanobis distance ($\bar{D}_M^{\mathcal{TM}}$) and the generalized Bhattacharyya distance ($D_B^{\mathcal{TM}}$). The last two columns report the test results when dropping one of the initial conditions.

sure these two groups of trajectories have similar means, the shapes in the third row of Fig. 2(a) are not picked randomly, but generated using the shapes in the second row. This is done by computing the mean of the shapes in the second row, then shooting a geodesic from the mean to each of the three shapes and continuing to move beyond time 1 (for two times) to generate the shapes in the third row. Essentially, this has the effect that the means of the trajectories of both groups are similar, but the variances differ.

Fig. 3 shows the results of PGA in trajectory space for the synthetic shapes. The largest eigenvalue of the trajectories in *Group A* is 0.005 at 72% cumulative variance, compared to the largest eigenvalue of 0.02 at 85% cumulative variance in *Group B*. Also, as expected, the trajectories visualized by 10 shapes in Fig. 3 show that the shapes of *Group B* change faster than in *Group A*. Table 2 reports the quantitative measures of the difference between the two groups. Since, by design, the mean trajectories are similar, it is difficult to identify significant deviations from the null-hypothesis H_0 using the generalized Mahalanobis distance; this is indicated by the relatively high p-values of $\bar{D}_M^{\mathcal{TM}}$ in Table 2[6]. As desired, $D_B^{\mathcal{TM}}$ is sensitive w.r.t. differences in variance, indicated by the relatively low p-value. This would allow to reject H_0 at the customary significance level of 0.05.

Furthermore, since all base shapes are uniformly sampled from within the shape triangle spanned by the first row of Fig. 2(a), *i.e.*, the initial points of the two groups have similar means, it is *not* possible to only use the initial points to separate the two groups; this is confirmed by the high p-values for both distance measures in the $(\hat{p}, 0)$ column of Table 2. In fact, even when specifically testing for differences in the initial velocity, the generalized Bhattacharyya distance exhibits better behavior than the generalized Mahalanobis distance in terms of lower p-values (*cf*. column $(0, \hat{u})$ of Table 2).

Corpora callosa in Kendall's shape space. The longitudinal corpus callosum dataset used in [12], contains 23 subjects, 11 of which are males with dementia, and the rest are normal controls. Every subject has been measured at three time points within the age range of 60 to 92 years old, and each corpus callosum shape is represented by 64 2D boundary landmarks.

[6] The average of two generalized squared-Mahalanobis distances is related to the first term of the generalized Bhattacharyya distance in Eq. (4).

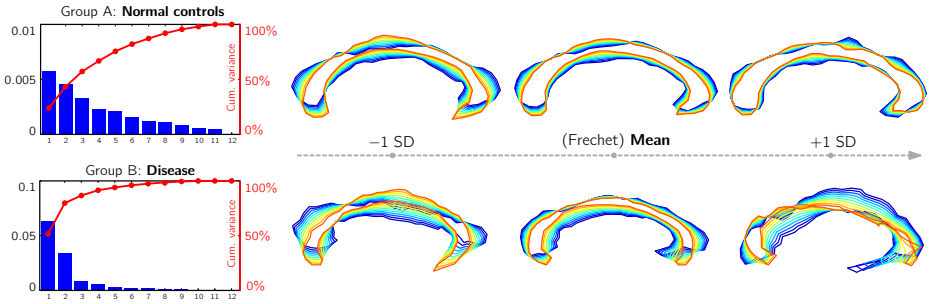


Fig. 4: Visualization of the variances (left) and principal directions (right) of trajectory distributions for the *normal control* (top) and *disease* group (bottom) of corpus callosum shapes (best-viewed in color, blue to red: young to old).

	$(\hat{p}, \hat{u})$		$(\hat{p}, 0)$		$(0, \hat{u})$	
	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$
Distance on $\mathcal{TM}$	3.1817	4.0029	3.7377	3.6863	4.1537	4.3765
p-value	0.0241	**0.0054**	0.2014	0.0654	0.0319	**0.0046**

Table 3: Distances and estimated p-values (10000 random permutations) on corpora callosa using the averaged Mahalanobis distance ($\bar{D}_M^{\mathcal{TM}}$) and the generalized Bhattacharyya distance ($D_B^{\mathcal{TM}}$). The last two columns report the test results of dropping one of the initial conditions during the distance computation.

Fig. 4 demonstrates the variances and the principal directions of the trajectories from the normal controls and the disease group. As shown in Fig. 4, the largest eigenvalue of the normal control group only accounts for 24% variability with a numeric value of 0.006, while the largest eigenvalue of the disease group accounts for 52% variability with a numeric value of 0.06. Fig. 4 (right) further shows the trajectories of each group along the first principal direction with standard deviations changing from -1 to 1. The plots indicate that the corpora callosa with dementia degenerate faster than the normal controls.

Table 3 reports the quantitative measures of the group tests on the corpus callosum shapes with 10000 permutations. Compared to the generalized squared-Mahalanobis distance, the generalized Bhattacharyya distance consistently exhibits better behavior in identifying the group differences. Similar to the experiments on the synthetic shapes, during the distance computation we drop one term of the initial conditions to measure which one plays a more important role in the group tests. As shown in Table 3, regardless of the distance measure, the initial velocity is most relevant in identifying group differences; this is consistent with [12]. If we declare the statistical significance at the level of 0.01, the p-value of the generalized Bhattacharyya distance, either using both

initial conditions or only the initial velocity, indicates that the disease group of corpus callosum shapes is significantly different from the normal control group.

5 Discussion

We have proposed an approach for studying group differences in the distributions of shape *trajectories*, estimated from longitudinal data. By means of a generalized version of the Bhattacharyya distance, we demonstrated, on both real and toy data, that taking second-order statistics into account can be beneficial in assessing group differences. However, the proposed approach also has limitations. For instance, although the compact representation of a trajectory is an efficient way to summarize longitudinal data, its accuracy inevitably influences the test-statistics. Currently, the adopted regression approach for estimating a trajectory is a generalization of linear regression in Euclidean space. Hence, we expect poor fitting performance on data that cannot be represented by a geodesic. For that reason, our test-statistic may not be appropriate under such a model. Furthermore, our real dataset only contains a limited number of subjects, which does not allow strong conclusions and requires to interpret results in the context of the low sample size. A potential direction for future work is to apply our method to other types of longitudinal data, *e.g.*, images, which is straightforward but slightly more involved due to the complexity of the tangent bundle.

Acknowledgements This work was supported by NSF EECS-1148870 and NSF EECS-0925875.

Appendix

A Properties of the generalized Bhattacharyya distance

Non-negativity. In the first term of Eq. (4), $D_M^{\mathcal{TM}}$ is the generalized squared-Mahalanobis distance which is non-negative; consequently, the first term in Eq. (4) is non-negative. Furthermore, the determinant of a covariance matrix in the second term is also non-negative, since it is the product of all non-negative eigenvalues. Besides, it is easy to demonstrate that $(|\Sigma_1|+|\Sigma_2|)/(2\sqrt{|\Sigma_1||\Sigma_2|}) \geq 1$, indicating the second term is non-negative. Hence, $D_B^{\mathcal{TM}}(D_1, D_2) \geq 0$.

Identity of indiscernibles. If $D_1 = D_2$, *i.e.*, $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$, we see that (1) $\mathrm{Log}_{\mu_1} \mu_2$ and $\mathrm{Log}_{\mu_2} \mu_1$ are zero tangent vectors, and (2) $|\Sigma_1| = |\Sigma_2|$. Hence, $D_M^{\mathcal{TM}}(\mu_1, D_2) = D_M^{\mathcal{TM}}(\mu_2, D_1) = 0$, *i.e.*, the first term of Eq. (4) is 0; also, the second term is 0. Now, if $D_1 = D_2$ then $D_B^{\mathcal{TM}}(D_1, D_2) = 0$. On the other hand, assuming $D_B^{\mathcal{TM}}(D_1, D_2) = 0$, we can only obtain $\mu_1 = \mu_2$ and $|\Sigma_1| = |\Sigma_2|$, because of the non-negativity properties of the two terms in Eq. (4). But, we *cannot* draw the conclusion that the two covariance matrices are equal. Therefore, if $D_1 = D_2$ then $D_B^{\mathcal{TM}}(D_1, D_2) = 0$, but it is possible that $D_B^{\mathcal{TM}}(D_1, D_2) = 0$ for some $D_1 \neq D_2$, if $\mu_1 = \mu_2$ and $|\Sigma_1| = |\Sigma_2|$.

Symmetry. Because both terms of Eq. (4) are symmetric, the sum of them is also symmetric, *i.e.*, $D_B^{\mathcal{T}\mathcal{M}}(D_1, D_2) = D_B^{\mathcal{T}\mathcal{M}}(D_2, D_1)$.

Triangle inequality. Since, Eq. (1) in $\mathbb{R}^n$ does not satisfy the triangle inequality, our generalized variant will not satisfy it either.

References

1. A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946. 4
2. P.J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Stat.*, pages 2577–2604, Dec. 2008. 6
3. S. Durrleman, X. Pennec, A. Trouvé, J. Braga, G. Gerig, and N. Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *IJCV*, 103(1):22–59, May 2013. 2
4. P.T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *IJCV*, 105(2):171–185, Nov. 2013. 1, 5
5. P.T. Fletcher, C. Lu, S.M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005, Aug. 2004. 2, 6
6. J.H. Gilmore, F. Shi, S.L. Woolson, R.C.. Knickmeyer, S.J. Short, W. Lin, H. Zhu, R.M. Hamer, M. Styner, and D. Shen. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cereb. Cortex*, 22(11):2478–2485, Nov. 2012. 1
7. Y. Hong, S. Joshi, M. Sanchez, M. Styner, and M. Niethammer. Metamorphic geodesic regression. In N. Ayache, H. Delingette, P. Golland, and K. Mori, editors, *MICCAI, 2012, Part III. LNCS, vol. 7512*, pages 197–205, 2012. 1, 5
8. O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88(2):365–411, Feb. 2004. 6
9. J. Lee. *Introduction to smooth manifolds*. Springer, 2012. 2
10. P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, Apr. 1936. 4
11. D.S. Marcus, A.F. Fotenos, J.G. Csernansky, J.C. Morris, and R.L. Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *J. Cognitive Neurosci.*, 22(12):2677–2684, Dec. 2010. 1
12. P. Muralidharan and P.T. Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *CVPR*, pages 1027–1034, 2012. 1, 5, 6, 7, 9, 10
13. M. Niethammer, Y. Huang, and F.X. Vialard. Geodesic regression for image time-series. In G. Fichtinger, A. Martel, and T. Peters, editors, *MICCAI 2011, Part II, LNCS, vol. 6892*, pages 655–662, 2011. 1, 5
14. D.S. Oliver. Calculation of the inverse of the covariance. *Math. Geol.*, 30(7):911–933, 1998. 6
15. S. Sasaki. On the differential geometry of tangent bundles of Riemannian manifolds. *TMJ*, 10(3):338–354, 1958. 5
16. N. Singh, J. Hinkle, S. Joshi, and P.T. Fletcher. A hierarchical geodesic model for diffeomorphic longitudinal shape analysis. In *IPMI*, pages 560–571, 2013. 1
17. J. Su, S. Kurtek, E. Klassen, and A. Srivastava. Statistical analysis of trajectories on riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *Ann. Appl. Stat.*, 8(1):530–552, Mar. 2014. 2
18. J. Su, A. Srivastrava, F. de Souza, and S. Sarkar. Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition. In *CVPR*, pages 620–627, 2014. 2