Riemannian Nonlinear Mixed Effects Models: Analyzing Longitudinal Deformations in Neuroimaging

Hyunwoo J. Kim[†] Nagesh Adluru[†] Heemanshu Suri[†]
Baba C. Vemuri[§] Sterling C. Johnson[†] Vikas Singh[†]
[†]University of Wisconsin-Madison [§]University of Florida
http://pages.cs.wisc.edu/~hwkim/projects/riem-mem

Abstract

Statistical machine learning models that operate on manifold-valued data are being extensively studied in vision, motivated by applications in activity recognition, feature tracking and medical imaging. While non-parametric methods have been relatively well studied in the literature, efficient formulations for parametric models (which may offer benefits in small sample size regimes) have only emerged recently. So far, manifold-valued regression models (such as geodesic regression) are restricted to the analysis of cross-sectional data, i.e., the so-called "fixed effects" in statistics. But in most "longitudinal analysis" (e.g., when a participant provides multiple measurements, over time) the application of fixed effects models is problematic. In an effort to answer this need, this paper generalizes non-linear mixed effects model to the regime where the response variable is manifold-valued, i.e., $f : \boldsymbol{R}^d \rightarrow \mathcal{M}$. We derive the underlying model and estimation schemes and demonstrate the immediate benefits such a model can provide — both for group level and individual level analysis — on longitudinal brain imaging data. The direct consequence of our results is that longitudinal analysis of manifold-valued measurements (especially, the symmetric positive definite manifold) can be conducted in a computationally tractable manner.

1. Introduction

A multitude of applications across computer vision and statistical machine learning involve operations on objects that are "structured"; for instance, probability densities, special classes of matrices/tensors, trees and graphs. It is generally accepted in the literature that treating these objects in their vectorized forms within off the shelf machine learning models may, more often than not, lead to unsatisfactory results, since the basic conditions of Euclidean geometry including distances and angles may not be satisfied. Defining machine learning models on such data requires care because it involves performing calculations while respecting the constraints inherent to the space where our samples come from [1]. This is challenging when the geometry of the space is not very well understood.

Fortunately, in certain cases, the canonical representations of the structured objects of interest lead to spaces that are relatively better characterized. For instance, when the space is a Riemannian manifold, the differential geometry provides a suitable Riemannian metric which serves as a building block for inference models. Within the last ten or so years, numerous authors have reported how respecting the geometry/structure of the data in tasks as simple as summarizing the observed variability in the empirical distribution [2] yields good empirical performance. There is a rapidly increasing body of work showing how manifold-based formulations yield powerful algorithms for problems including affine motion tracking [3, 4], human detection [5], modeling spatial distribution of earthquakes [6], classification tasks [7], robot localization [8], registration [9], and numerous settings in medical imaging [10, 11, 12].

Symmetric positive definite matrices (SPD) are a good representative example of structured data that lie on a Riemannian manifold. In computer vision, SPD matrices are often encountered due to covariance matrices which serve as feature descriptors in tracking [4] as well as for texture analysis and classification [5]. Diffusion tensor magnetic resonance imaging, a type of neuroimaging modality, yields an SPD matrix *at each image voxel* as the local diffusion tensor. But even aside from the SPD manifold, probability densities, shape spaces, and certain warping functions are other examples of Riemannian manifolds which commonly manifest themselves in vision applications [12, 13, 14, 15]. The need for metrics, estimators, hypothesis tests and other tools for inference tasks has led to much interest in Riemannian analogs of classical vision and machine learning methods. Consequently, we now have manifold versions of nearest neighbor queries, dimensionality reduction, dictionary learning, sparse coding, kernel methods, boosting, averaging and regression. These developments were recently

compiled within a handbook [16] which offers an overview of what has been accomplished so far as well as the yet unresolved technical issues.

A majority of the technical contributions in manifold-based algorithms within computer vision have focused on non-parametric formulations of the underlying problems. The parametric versions are recently receiving attention, motivated by applications in the life sciences where *sample sizes are smaller*. For instance, consider the simple parametric model that seeks to identify associations between a predictor (feature) x and a response y which is manifold-valued, $f : x \rightarrow y$. Efficient parametric algorithms for this problem were developed as recently as a few years back [17] where the model was used to identify the relationship between a "shape-based" manifold variable (e.g., shape of a segmented brain region) and a subject specific variable x (e.g., age). Only several months back, the generalization of this model for multi-dimensional predictors was obtained — to fit the response y by a linear combination of multiple predictors (or features), $x \in R^d$ or even when the predictor variable is manifold-valued. Such a construction also serves to identify associations between a set of covariates while controlling for a set of "nuisance" variables [18, 19].

Manifold-valued regression. In a longitudinal neuroimaging study of disease progression, the statistical models need to capture morphometric changes *over time* while controlling for the dependency of *repeated* measurements from the same subject. Consider the following analysis: we have two groups of subjects, "controls" and "disease" which correspond to healthy controls and individuals with a high risk of a disease. We want to identify the group-differences with respect to time (or disease progression). It is known that anatomical changes at a voxel X can be captured by spatial derivatives, i.e., the Jacobian matrix $J(X)$, of deformation maps of a subject (e.g., the changes between the first and the second time points, two years apart). The most widely used "deformation" feature is the log determinant of the Jacobian matrices, $\log(\det(J(X)))$ — a scalar voxel-wise value which captures the volumetric/anatomical changes. The so-called Cauchy deformation tensor (CDT) [20] represented as $\sqrt{J(X)^T J(X)}$ is a richer representation of $J(X)$, an object on the SPD(3) manifold, see Fig. 1. To understand how each specific voxel in the image is associated with a predictor variable (say, age or disease status), a regression between the predictor and CDT at each voxel will identify brain regions that are most affected by age or disease. This can be accomplished by recently proposed algorithms for geodesic regression [17, 19, 18] or kernel regression on manifolds [21].

Random effects in longitudinal analysis. Now, let us consider a slightly more involved setting where each subject provides data over multiple time points, a few years apart. In such a *longitudinal* setting, we obtain *one* CDT image between each consecutive time point (i.e., pairs). A standard linear regression (or its manifold-valued analog) is *agnostic to dependency of temporal samples*. Since subjects are examined multiple times within the study, the repeated measurements from the same subject as the subject specific "*random effect*". This dependency violates the i.i.d. assumptions of *fixed effects* models (e.g., generalized linear models), including the manifold versions [17, 18]. The fixed effects model assumes that all data are i.i.d. samples from the same underlying generating function with random noise on the response variable Y. As Fig. 2 shows, each subject may have a different trend. For example, subject A has an early disease onset (intercept). Subject B shows faster disease progression (slope). Also, based on the age-range of participants, there may be larger variability *between* subjects than the variability *within* a subject. So, a fixed effects linear model for the data in Fig. 2, is *prone to fit population level variability (black) than the trajectory of each subject (red)*. Indeed, such subject specific *random effects* can be modeled via the more general *mixed effects models*. The overarching **goal and contribution** of this work is to derive formulations/algorithms for the regime where the set of longitudinal responses Y is a manifold valued variable and the objective is to fit linear (or non-linear) mixed effects models.

Related work. As briefly described above, there is a growing body of work in vision that incorporates the specific geometry/structure of the data directly within the estimation problem. Various statistical constructs have been generalized to Riemannian manifolds: these include regression [22, 23], margin-based and boosting classifiers [7], classification [24], kernel methods [25], filtering [10] and dictionary learning [26, 27]. The literature also includes ideas related to projective dimensionality reduction methods. For instance, the generalization of Principal Components analysis (PCA) via the so-called PGA [28], Geodesic PCA [29], Exact PGA [30], CCA on manifolds [31], Horizontal Dimension Reduction [32] with frame bundles, and an extension of PGA to the product space of Riemannian manifolds, namely, tensor fields [24]. We should note that an important earlier work dealing with a univariate linear model on manifolds (related to geodesic regression) was studied for the group of diffeomorphisms using metrics which were not strictly Riemmanian [33, 34]. Despite these developments, the first results on successful generalization of parametric regression models to Riemannian manifolds is relatively recent: geodesic regression [17, 35], polynomial regression [36] and multi-variate linear regression with manifold-valued responses [18]. Finally, we point out that a recent independent result on mixed effects models in [37] deals with univariate manifolds $[0, 1]$, which is the unit interval in a real line $\mathbf{R}$ with a specifically designed metric to capture sigmoid function like patterns. This work does not deal

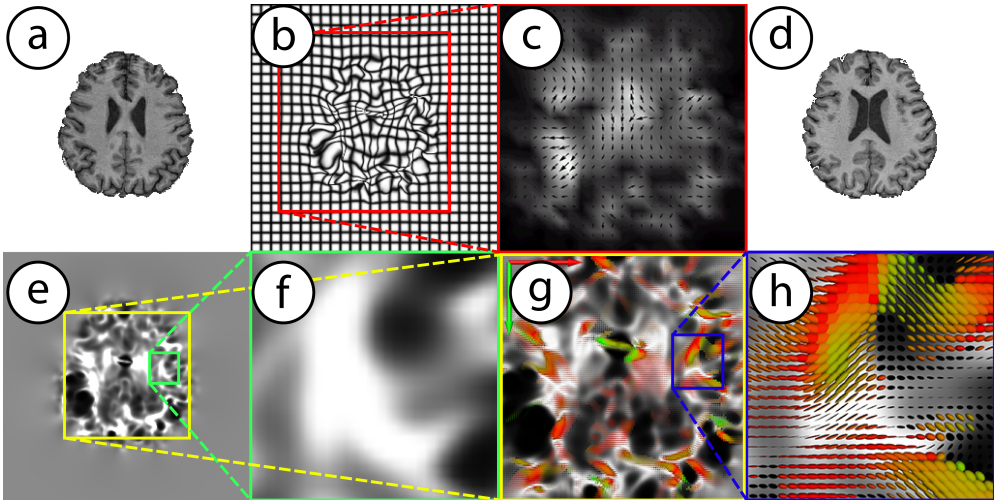

Figure 1. An example panel of data generated in morphometric studies. (a, d) The moving and fixed brain image respectively. (b) Warped spatial grid to move (a) to (d). (c) Vector field of local deformations. (e, f) A map of the $det(J)$ of the deformation field. (g, h) The Cauchy deformation tensor field (CDTs) ($\sqrt{J^T J}$). Among the different features of brain morphology that can be analyzed, CDTs are the focus of this paper.

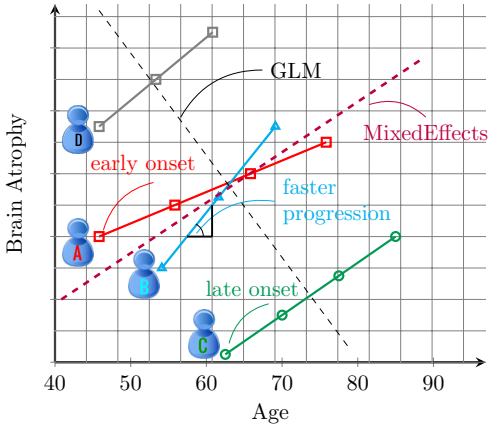

Figure 2. This figure demonstrates the key effects we are interested in capturing. Each subject has a different progression rate of the brain atrophy (acceleration effect) and has a different onset for the change (time shift). A regular general linear model (GLM) with fixed effects is insufficient to capture these effects while including random effects (subject-specific slope and intercept) in mixed effects models can capture these effects.

with actual manifold-valued variables (e.g., SPD); further it is computationally impractical for more than hundreds of voxels. In contrast, 3D CDT images we will analyze exceed 1M+ voxels.

2. Preliminary concepts and notations

We first briefly review *linear mixed effects models* and their estimation methods. Then, we summarize some basic *differential geometry* notations that we will use. The extended version of this paper and [38, 39] describe more details. Also, we introduce CDT and Jacobian matrices to capture longitudinal morphometric brain changes.

2.1. Linear mixed effects model

In general, the estimation of regression models (such as linear/polynomial) assumes that the data come from an underlying model with i.i.d. noise; so effects of the covariates/features are pertinent to the entire sample. These models are called *fixed effects*. For example, a linear regression model is also a fixed effects model given as

$$y = \beta^0 + \beta^1 x^1 + \cdots + \beta^p x^p + \epsilon, \tag{1}$$

where $y \in \mathbf{R}$, $\boldsymbol{x} \in \mathbf{R}^p$, $\boldsymbol{\beta} = [\beta^0, \ldots, \beta^p]^T \in \mathbf{R}^{p+1}$. We see that the coefficients are 'fixed' and the same over entire population. However, in longitudinal studies (see Fig. 2), the repeated measurements from the same subject are *no longer independent*. We need a more flexible specification – often covariates/features have different effects on individual subjects (or groups), which is called *random effects*. For example, the rate of brain atrophy and disease progression can vary over subjects given by

$$y_i = u_i^1 z^1 + \cdots + u_i^q z^q + \epsilon_i, \tag{2}$$

where $\mathbf{z}$ is a known vector specifying which subject (or group) a sample belongs to, and u_i^q is the q^{th} random effect for the i^{th} subject (or group) denoted by $\boldsymbol{u}_i$. This combination of *fixed* and *random* effects yields *mixed effects models* [40]. When the model is linear, we get linear mixed effects models, which we introduce next. We then work with its *nonlinear* analog. The nonlinear mixed effects models are an intermediate (but necessary) step in deriving our final models for manifold-valued data, introduced in Sec. 3.2.

Specifying the model. Let $\boldsymbol{y}_i = \left[y_{[\![ij]\!]}\right]_{j=1}^{n_i}$ be a set of n_i repeated observations of a response variable for subject i. Here $\boldsymbol{y}_i$ is a n_i dimensional vector, vertically stacked with $y_{[\![ij]\!]}$ responses for subject i. The notation $[\![i, j]\!]$ simply recovers the specific observation j for subject i. Similarly, let the subject-specific matrix X_i of size $n_i \times p$ be setup as $\left[x^1_{[\![ij]\!]}\ x^2_{[\![ij]\!]}\ \ldots\ x^p_{[\![ij]\!]}\right]_{j=1}^{n_i}$ where we collect for

subject i, all p measurements for all n_i visits as rows. The matrix Z_i will provide information on the number of longitudinal measurements for each subject (design matrix). Similar to X_i, we define Z_i by specifying rows as $Z_i = \left[z^1_{[\![ij]\!]}\ z^2_{[\![ij]\!]}\ \cdots\ z^q_{[\![ij]\!]}\right]_{j=1}^{n_i}$. These correspond to sets of p and q variables (features) for the i^{th} subject where one is interested in estimating fixed effects for the set X_i and random effects for the set Z_i on $\boldsymbol{y}_i$. In the classical setting, a linear mixed effects model ([40]) is given by

$$\begin{aligned} y_{[\![ij]\!]} =& \beta^0 + \beta^1 x^1_{[\![ij]\!]} + \cdots + \beta^p x^p_{[\![ij]\!]} + \\ & u^1_i z^1_{[\![ij]\!]} + \cdots + u^q_i z^q_{[\![ij]\!]} + \epsilon_{[\![ij]\!]}, \end{aligned}$$

where $\beta^1, \ldots, \beta^p$ are the fixed effects shared over the entire population and $u^1_i, \ldots, u^q_i$ are the (subject-specific) random effects for the i^{th} subject. The random effects $\boldsymbol{u}_i = [u^1_i\ u^2_i\ \cdots\ u^q_i]^T$ are assumed to follow a multivariate normal distribution. The "unexplained" random error $\epsilon_{[\![ij]\!]}$ comes from a normal distribution $\mathcal{N}(0, \Sigma^2_\epsilon)$. We can compactly write the model using matrix notation as

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{u}_i + \epsilon_i. \tag{3}$$

Let 'vstack$(\cdot)$' be the vertical stack of parameters. By denoting $\boldsymbol{y} = \text{vstack}(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)$, and similarly $X, Z, \boldsymbol{u}$, the final model for all N subjects can be expressed as,

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \sigma^2_\epsilon I,$$

where $\boldsymbol{u} \sim \mathcal{N}(0, \tilde{\Sigma})$ and $\tilde{\Sigma} = \text{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_N) = \Sigma \otimes I$ (when $\Sigma_i = \Sigma \forall i$), and $Z = \text{diag}(Z_1, Z_2, \ldots, Z_N)$. In general, estimation of linear mixed effects models does not have an analytic solution unless $\tilde{\Sigma}$ and σ^2_ϵ are known.

2.2. Basic differential geometry notations

Let $\mathcal{M}$ be a *differentiable (smooth) manifold* in arbitrary dimensions. A differentiable manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure. A *Riemannian manifold* $(\mathcal{M}, g)$ is a differentiable manifold $\mathcal{M}$ equipped with a smoothly varying inner product g. The family of inner products on all tangent spaces is known as the *Riemannian metric*, which defines various geometric notions on curved manifolds such as the length of a curve etc. A *geodesic curve* is a locally shortest path, which is analogous to straight lines in $\mathbf{R}^d$ — such a geodesic curve will be the object that defines the trajectory of our covariance matrices in the space of SPD matrices. Unlike the Euclidean space, note that there may exist multiple geodesic curves between two points on a curved manifold. So the *geodesic distance* between two points on $\mathcal{M}$ is defined as the length of the *shortest* geodesic curve connecting two points (i.e., SPD matrices). The geodesic distance helps in measuring the error of our trajectory estimation (analogous to a Frobenius or ℓ_2 norm based loss in the Euclidean setting). The geodesic curve from y_i to y_j is parameterized by a tangent vector in the tangent space anchored at y_i with an exponential map $\text{Exp}(y_i, \cdot) : T_{y_i}\mathcal{M} \to \mathcal{M}$. The inverse of the exponential map is the logarithm map, $\text{Log}(y_i, \cdot) : \mathcal{M} \to T_{y_i}\mathcal{M}$. These two operations move us back and forth between the manifold and the tangent space. Separate from the above notation, matrix exponential (and logarithm) are simply $\exp(\cdot)$ (and $\log(\cdot)$). Finally, *parallel transport* is a generalized parallel translation on manifolds. Given a differentiable curve $\gamma : \mathcal{I} \to \mathcal{M}$, where $\mathcal{I}$ is an open interval, the parallel transport of $v_0 \in T_{\gamma(t_0)}\mathcal{M}$ along curve γ can be interpreted as the parallel translation of v_0 on the manifold preserving the angle between $v(t)$ and γ. The parallel transport of v from y to y' is $\Gamma_{y \to y'} v$ (more details in the extended version).

3. Longitudinal analysis of CDT images

Let $\mathcal{I}_{i,j}$ denote the image acquired from subject i at time point j. Given images $\mathcal{I}_{i,j}$ and $\mathcal{I}_{i,j+1}$ for successive visits $(j, j+1)$, we can compute a deformation that aligns the two images [41, 42]. Let $\mathcal{I}_{i,1}$ give the subject-specific coordinate system denoted as Ω_i. This will provide the (intermediate) common coordinate system to represent the deformations undergone by subject i over time, $j = 1, 2, \cdots, n_i$. The global template where all $(n_i - 1)$ temporal deformations (i.e., CDT images) for each subject i will be represented is denoted as Ω. Then, a nonlinear deformation $\Phi(\texttt{vox})$ for voxels (spatial locations) $\texttt{vox} \in \Omega$ for each image (rather, for each $(\mathcal{I}_{i,j+1}, \mathcal{I}_{i,j})$ pair) is given as

$$\begin{aligned} &\Phi : \mathcal{I}_{i,j+1} \to \mathcal{I}_{i,j} \\ &\Phi(\texttt{vox} + d\texttt{vox}) = \Phi(\texttt{vox}) + J(\texttt{vox})d\texttt{vox} + \mathcal{O}(d\texttt{vox}^2), \end{aligned} \tag{4}$$

where $J(\texttt{vox})$ denotes the Jacobian of the deformations at position vox. A nice property of CDTs is that it preserves the determinant of $J(\texttt{vox})$, since $\det(J(\texttt{vox})) > 0$. So, a CDT representation introduced in Sec. 1, nicely symmetrizes $J(\texttt{vox})$ without affecting the volumetric change information, i.e., $\det(J) = \det(\sqrt{J^T J})$ (proof in the extended version). The CDT "image" comprised of voxel locations vox is an object of the same size as $\mathcal{I}_{1,1}$ and derived from a black-box diffeomorphism solver given as a 3×3 SPD matrix $\sqrt{J^T J}$ at each voxel. It provides the deformation field between two longitudinal images of a subject. Various results have described the benefits of CDT images for analysis; an example from our experiments is in Sec. 6.

Recall that for a pair of real numbers a and b, the prediction of a linear model "$y = ax + b$" is always a real number. But if a and b are SPD matrices (i.e., CDTs), this property no longer holds: "$y = ax + b$" may not be a valid SPD matrix (e.g., consider when $x < 0$). In addition to the validity of predictions, while in classical linear regression, the residual of a fit, $(y - (Ax + b))$, can be estimated directly,

the error (residual) on a SPD manifold must be measured via the *geodesic distance* on a manifold between the prediction $\hat{y}$ and the true response y. So, any machine learning model that needs to be extended for manifold-valued data must give particular attention to these issues, and often specialized optimization methods for parameter estimation are needed [17, 18, 21]. With these issues in mind, we first describe a linear mixed effects models with subject-specific intercepts and then transition to nonlinear mixed effects models for manifold valued data.

3.1. A model with subject specific intercepts

We know that in any longitudinal dataset, the errors/noise of repeated measurements are dependent. To consider this aspect, a common approach is to express the random effects as nuisance parameters. If the set $\{i = 1, i = 2, \cdots, i = N\}$ indexes the columns, we may write the design matrix Z as $\text{diag}(\mathbb{1}_{n_1}, \mathbb{1}_{n_2}, \cdots, \mathbb{1}_{n_N})$, where $\mathbb{1}_{n_i} = [1 \ \cdots 1]^T \in R^{n_i}$. Then, the model in (3) becomes

$$y_{[\![ij]\!]} = \beta^0 + \boldsymbol{\beta}^T \boldsymbol{x}_{[\![ij]\!]} + u_i, \tag{5}$$

where $y_{[\![ij]\!]}, \beta^0, u_i \in \mathbf{R}$, and $\boldsymbol{\beta}, \boldsymbol{x}_{[\![ij]\!]} \in \mathbf{R}^p$. Note that $\mathbf{z}_{[\![ij]\!]} \in \mathbf{R}^N$ recovers a specific row corresponding to subject i's visit j from matrix Z taking dot product with $\boldsymbol{u}$ gives us the subject specific random effects, $u_i = \mathbf{z}_{[\![ij]\!]}\boldsymbol{u}$.

This model poses two problems. **1)** It has the same slope $\boldsymbol{\beta}$ for the entire population, whereas subjects in the study may have different rate of disease progression; **2)** Another issue is the interpretation of u_i, which is viewed as subject-specific shift in the y space or x space, i.e., depending on whether we move it to the left or right of the equality in (5). In medical applications, readability of models is important to understand the disease. Our solution involves explicitly adding a subject-specific shift for x as well as a shift in y.

3.2. Nonlinear mixed effects models with $\psi_i(\boldsymbol{x})$

Based on the foregoing motivation, we can extend the linear mixed effects models with a subject-specific random function $\psi_i(\cdot)$ as

$$y_{[\![ij]\!]} = \beta_0 + \mathbf{z}_{[\![ij]\!]}u_i + \boldsymbol{\beta}.\psi_i(\boldsymbol{x}_{[\![ij]\!]}) \tag{6}$$

Depending on the form of $\psi_i(\cdot)$, (6) can be a nonlinear mixed effects model (NLMM). When ψ_i is the Identity, we simply get a linear mixed effects model. In our analysis, we use $\psi_i(\boldsymbol{x}) := \alpha_i(\boldsymbol{x} - \tau_i - t_0) + t_0$ motivated by [43] where each subject can have their own speed of disease progression (α_i) and different onset time τ_i, but $(\beta_0, \boldsymbol{\beta}, t_0)$ are common for the population. Then we have

$$y_{[\![ij]\!]} = \beta_0 + \mathbf{z}_{[\![ij]\!]}u_i + \boldsymbol{\beta}(\alpha_i(\boldsymbol{x}_{[\![ij]\!]} - \tau_i - t_0) + t_0). \tag{7}$$

Note that this extension is different from the *generalized linear mixed effects models* [44], e.g., $y_{[\![ij]\!]} = h^{-1}(\boldsymbol{x}_{[\![ij]\!]}\boldsymbol{\beta} + \mathbf{z}_{[\![ij]\!]}u_i), u_i \sim \mathcal{N}(0, \Sigma_i)$. Next, we extend the mixed effects models in (5) and (7) to manifold-valued data.

4. Mixed effects models on manifolds

The Linear Mixed Effects Model (LMM) can be extended in many ways to the manifold setting depending on the order of addition and interpretation. For instance, recall that the associativity of addition, $(a+b)+c = a+(b+c)$, in the Euclidean space is not directly translated to manifolds, i.e., $\text{Exp}(\text{Exp}(a, b), c) \neq \text{Exp}(\text{Exp}(a, c'), b')$, where b' and c' are parallelly transported tangent vectors of b and c respectively, so that they are in the right tangent spaces. A natural extension of LMM in (3) can be written as

$$y_{[\![ij]\!]} = \text{Exp}(\text{Exp}(\text{Exp}(B, V\boldsymbol{x}_{[\![ij]\!]}), U_i\mathbf{z}_{[\![ij]\!]}), \epsilon_{[\![ij]\!]}), \tag{8}$$

where $y_{[\![ij]\!]}, B, B_i \in \mathcal{M}$, $V \in T_B\mathcal{M}^p$, $U_i \in T_{h_{[\![ij]\!]}}\mathcal{M}^q$, $h_{[\![ij]\!]} = \text{Exp}(B, V\boldsymbol{x}_{[\![ij]\!]})$, $\boldsymbol{x}_{[\![ij]\!]} \in \mathbf{R}^p$ and $\mathbf{z}_{[\![ij]\!]} \in \mathbf{R}^q$. Recall that the base point B on the manifold $\mathcal{M}$ is the analog to the intercept β^0 in (5) whereas V (and U_i) corresponds to the slope $\boldsymbol{\beta}$ (and the random effects $\boldsymbol{u}_i$) respectively. Unfortunately, the model above has an issue related to U_i. Note that U_i is used in different tangent spaces at $h_{[\![ij]\!]}$. Also, especially on SPD(n) manifolds with the GL-invariant metric, the norm of the tangent vectors varies as a function of the base point B of the respective tangent spaces, i.e., $\|U\|_B^2 = \langle U, U\rangle_B = \text{tr}(UB^{-1}UB^{-1})$. So the corresponding scales might be different. As a result, the prior for U_i needs to be carefully designed [45] to have consistency over all tangent spaces. To address this problem, we change the order of the exponential maps and propose a mixed effects model with subject specific intercepts (shift in y) on manifolds. Also, unlike the Euclidean space, in general, there is no equivalence between the shift in x and the shift in y. So, we can explicitly add in the shift in x, denoted as τ_i. Then, our formulation on manifolds is given as

$$\boldsymbol{y}_{[\![ij]\!]} = \text{Exp}(\text{Exp}(B_i, \Gamma_{B\to B_i}(V)(\boldsymbol{x}_{[\![ij]\!]} - \boldsymbol{\tau}_i)), \boldsymbol{\epsilon}_{ij}), \tag{9}$$

$$B_i = \text{Exp}(B, U_i\mathbf{z}_{[\![ij]\!]}), \tag{10}$$

where $\boldsymbol{\tau}_i \in \mathbf{R}^p$, $B_i \in \mathcal{M}$, $V \in T_B\mathcal{M}^p$, $U_i \in T_B\mathcal{M}^q$, and the remaining variables are the same as before. As in the standard mixed effects models, U_i is assumed to follow a normal distribution. With this construction in hand, we may now include a subject-specific time shift in the onset time (similar to (7)) and assume that the progression of disease has the same overall pattern but only its speed/rate and onset time vary between subjects. The model is given by

$$y_{[\![ij]\!]} = \text{Exp}(\text{Exp}(B_i, \Gamma_{B\to B_i}(V)\alpha_i(x_{[\![ij]\!]} - \tau_i - t_0), \epsilon_{ij})))$$

$$B_i = \text{Exp}(B, U_i), \tag{11}$$

where α_i is the subject-specific acceleration, $\alpha_i < 1$ (and $\alpha_i > 1$ resp.) means slower (and faster resp.) than the population. Further, τ_i is the subject-specific shift in onset

time: $\tau_i > 0$ means a late onset time whereas t_0 is the global shift in onset time. Finally, U_i (or B_i) are the tangent vectors (or base points) that characterize the subject-specific shift in the response variable space (see the Euclidean case in Fig. 2). As in the classical setting, we may assume that the random effects are normally distributed, i.e., $\Gamma_{B\to I}U_i \sim \mathcal{N}_{SYM}(0,\sigma_U^2),\ \alpha_i \sim \mathcal{N}(1,\sigma_\alpha^2),\ \tau_i \sim \mathcal{N}(0,\sigma_\tau^2)$.

5. Parameter estimation procedure

The estimation of the mixed effects models on manifolds is challenging; provably accurate estimation methods are almost infeasible to run over the entire brain image. Even for Euclidean response variables, efficient estimation methods for *nonlinear* mixed effects models are still being actively studied in literature, e.g., Alternating algorithms [44], Laplacian and adaptive Gaussian quadrature algorithms [46], as well as generalized EM algorithms with MCMC [47]. Unfortunately, this issue only gets worse in the manifold setting. Fitting a nonlinear mixed effects model exactly, even for *univariate manifolds on the real line* takes about a day [37]. In our data set, the number of voxels is 1M+, it is impractical to perform exact analysis for the full brain. So, we present approximate algorithms based on a certain geometrical interpretation of the models.

5.1. Estimation of RNLMM

We observe that the main building block of our models, Riemannian nonlinear mixed effects models (RNLMMs), is a manifold-valued multivariate general linear model (MMGLM). This module has an efficient parameter estimation called the Log-Euclidean framework. It is known [19] that in practice the estimation can be well approximated in the tangent space at the Fréchet mean of the response variables Y with a centered X, i.e., $B \approx \bar{Y}, \tau \approx \bar{X}$. As in a global manifold-valued linear model [19], the parameter V will correspond to the full data set; however, we allow subject-specific variability for the base point B and τ via $B_i(r)$ and $\tau_i(r)$, where $r \in \mathbf{R}$ can be viewed as the mixing rate between the local models that share a global V. This is given by

$$y_{[\![ij]\!]} = \mathrm{Exp}(\mathrm{Exp}(B_i(r), \Gamma_{B\to B_i(r)}(V)(\boldsymbol{x}_{[\![ij]\!]} - \tau_i(r))), \epsilon). \tag{12}$$

In other words, $r \in \mathbf{R}$ is a weight to globally average the population subject specific base points $B_i(r)$ and time shifts $\tau_i(r)$ — all subjects share the fixed effects V but each subject corresponds to its own shifts $\tau_i(r)$ and $B_i(r)$ in x and y spaces. When $r = 0$, the model reduces to the model in [18] with only global intercepts (see the extended version).

Our estimation for (12) is summarized in Alg. 1, where $y^{\flat}_{[\![ij]\!]}$ is a tangent vector obtained by: taking the response $y_{[\![ij]\!]}$ and mapping it to the tangent space at $B_i(r)$ and parallel transporting that mapping to $T_B\mathcal{M}$. We now briefly describe how we can perform the estimation efficiently. First, in Step 2, we solve for the linear interpolation of two SPD matrices w.r.t the geodesic distance on the SPD manifold using the analytical form of the solution in [48] (note that when the number of samples is large, recursive schemes exist [49]). In Step 4, we transport the tangent vectors from B to I and vice versa using group action, which is known to be more efficient than parallel transport but equivalent [50].

Algorithm 1 Riemannian mixed effects models

1: Calculate the mean for each subject, $\bar{\boldsymbol{y}}_i$,

$$\bar{\boldsymbol{y}}_i = \operatorname*{argmin}_{\boldsymbol{y}\in\mathcal{M}} \sum_{j=1}^{n_i} \mathrm{d}(\boldsymbol{y}, \boldsymbol{y}_{[\![ij]\!]})^2. \tag{13}$$

Similarly calculate $\bar{\boldsymbol{y}}$ for the entire population.
2: Given r, solve for $B_i(r)$ (interpolation of $\bar{\boldsymbol{y}}_i$ and $\bar{\boldsymbol{y}}$) by

$$B_i(r) = \bar{\boldsymbol{y}}(\bar{\boldsymbol{y}}^{-1}\bar{\boldsymbol{y}}_i)^r = \bar{\boldsymbol{y}}_i(\bar{\boldsymbol{y}}_i^{-1}\bar{\boldsymbol{y}})^{1-r}, 0 \le r \le 1.$$

3: $\boldsymbol{y}^{\flat}_{[\![ij]\!]} = \Gamma_{\bar{B}_i(r)\to B}\mathrm{Log}(\bar{B}_i(r), \boldsymbol{y}_{[\![ij]\!]})$.
4: Transport $\boldsymbol{y}^{\flat}_{[\![ij]\!]}$ to I by group action.
5: Center $\boldsymbol{x}$ by $\boldsymbol{x}_{[\![ij]\!]}(r)$=$(1-r)(\boldsymbol{x}_{[\![ij]\!]} - \bar{\boldsymbol{x}}) + r(\boldsymbol{x}_{[\![ij]\!]} - \bar{\boldsymbol{x}}_i)$.
6: Calculate V^* using MMGLM [19] on transported $\boldsymbol{y}^{\flat}_{[\![ij]\!]}$ and $\boldsymbol{x}_{[\![ij]\!]}(r)$.
7: Prediction is given by

$$\hat{\boldsymbol{y}}_{ij} = \mathrm{Exp}(B_i(r), \Gamma_{B\to\bar{B}_i(r)}(V^*)\boldsymbol{x}_{ij}(r)). \tag{14}$$

5.2. Estimation of RNLMM with $\psi_i(x)$

The estimation of the model in (11) with the subject-specific random function $\psi_i(\cdot)$ involves few additional technical challenges. To reduce the problem complexity, we first find the main longitudinal change direction $\boldsymbol{\eta}$ controlling for the subject-specific random effects $\bar{Y}_i$ and $\bar{X}_i$ (since U_i and τ_i are random effects). This scheme is described in Alg. 2.

Algorithm 2 Calculate longitudinal change direction

1: Calculate the population Fréchet mean $\bar{\boldsymbol{y}}$ of response.
2: Calculate the Fréchet mean for each subject $\bar{\boldsymbol{y}}_i$.
3: Solve $\boldsymbol{y}^{\flat}_{[\![ij]\!]} = \Gamma_{\bar{\boldsymbol{y}}_i\to I}\mathrm{Log}(\bar{y}_i, \boldsymbol{y}_{[\![ij]\!]})$.
4: Solve $\boldsymbol{x}^{\flat}_{[\![ij]\!]} = \boldsymbol{x}_{[\![ij]\!]} - \bar{\boldsymbol{x}}_i$, where $\bar{\boldsymbol{x}}_i = \mathbb{E}_j[\boldsymbol{x}_{[\![ij]\!]}]$.
5: Collect $X^{\flat} = [\boldsymbol{x}^{\flat}_1; \ldots, \boldsymbol{x}^{\flat}_N]$, and $Y^{\flat} = [\boldsymbol{y}^{\flat}_1, \ldots, \boldsymbol{y}^{\flat}_N]$.
6: Calculate longitudinal change direction $\boldsymbol{\eta}$ by least squares estimation, $\boldsymbol{\eta}$=$((X^{\flat})^TX^{\flat})^{-1}((X^{\flat})^TY)$$\in T_I\mathcal{M}$.

Once the longitudinal change direction η (fixed effects for the entire population) is estimated, we solve for a subset of parameters at a time. This procedure is described in Alg. 3, where we solve for all parameters given the estimate of η. Note that for our downstream analysis, the bias induced by priors on parameters may reduce the statistical power. So, we simply used noninformative priors for all parameters. While Alg. 3 utilizes noninformative priors, with minor changes, we can easily incorporate normal distribution priors, see the extended version.

Algorithm 3 Riemannian mixed effects models with $\psi_i(x)$

1: Calculate the Fréchet mean $\bar{y} \in \mathcal{M}$ of population.
2: Calculate the Fréchet mean for each subject $\bar{y}_i \in \mathcal{M}$.
3: Main longitudinal change direction η by algorithm (2).
4: Calculate subject-specfic base points (random effects) $B_i = \text{Exp}(B, U_i^*)$, where $U_i^* = \text{argmin}_{U_i} \text{d}(\bar{y}_i, \text{Exp}(B, U_i))^2 + \lambda_{U_i} \|U_i\|_B^2$.
5: $y_{ij}^{\wr} = \Gamma_{B_i \to I} \text{Log}(B_i, y_{ij})$.
6: **while** until convergence **do**
7: Calculate the common speed of change $V = c\boldsymbol{\eta}$ and common time intercept $t_0 = b/c$ with fixed all other variables by

$$\begin{bmatrix} \sum_{ij} q_i^T q_i & \sum_{ij} p_{ij}^T q_i \\ \sum_{ij} p_{ij}^T q_i & \sum_{ij} p_{ij}^T p_{ij} \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{ij} q_i^T y_{ij}^{\wr} \\ \sum_{ij} p_{ij}^T y_{ij}^{\wr} \end{bmatrix},$$

where $b := t_0 c$, $q_i := \eta(1 - \alpha_i)$,$p_{ij} := \eta(\alpha_i x_{ij} - \alpha_i \tau_i)$.

8: Given V, t_0, calculate the subject-specific acceleration α_i, and time-shift τ_i by generalized least square estimation with the priors for α_i and $\tau_i = d_i/\alpha_i$

$$\begin{bmatrix} \sum_j W_{ij}^T W_{ij} & -\sum_j W_{ij}^T V \\ \sum_j W_{ij}^T V & -\sum_j V^T V \end{bmatrix} \begin{bmatrix} \alpha_i \\ d_i \end{bmatrix} = \begin{bmatrix} \sum_j \Upsilon_{ij}^T W_{ij} \\ \sum_j \Upsilon_{ij}^T V \end{bmatrix},$$

where $\Upsilon_{ij} := y_{ij}^{\wr} - V t_0$, $W_{ij} := V(X_{ij} - t_0)$ and $d_i = \alpha_i \tau_i$.
9: **end while**

Alg. 3 contains many steps in common with Alg. 1. In Step 7, we estimate the fixed effects V and t_0 by fixing all other variables (c is a dummy variable). In Step 8, we estimate the subject-specific random effects α_i and τ_i by fixing V and t_0 (d_i are dummy variables). Additional details including derivation are available in the extended version.

6. Experiments

Goals. The overarching goal of our experiments is to evaluate whether the proposed formulations can serve as core modules that drive longitudinal analysis of image datasets in neuroimaging. To this end, when conducting analysis of longitudinal data acquired in the context of a specific disease, the procedure should yield meaningful results for group analysis — for instance, when the population is split with a stratification variable (e.g., gender or disease risk factor), the "maps" of statistically significant group-wise differences in subject/voxel-specific "random" effects (especially, acceleration and spatial shift) should be scientifically interpretable, yet generally consistent with a baseline. Our experiments below show the extent to which the models satisfy this requirement.

Data. The CDT images (denoting subject-specific warps) were derived from a longitudinal neuroimaging study of pre-clinical Alzheimer's disease (AD). The longitudinal warps (or transformations) were obtained using *with-in* subject registration of T_1-weighted images between two consecutive visits i.e., $\Phi_{i,j} : \mathcal{I}_{i,j} \to \mathcal{I}_{i,j+1}$. Voxelwise CDTs were derived from the spatial derivatives $\nabla \Phi_{i,j}(\texttt{vox})$ of the deformation field. The details of minimizing spatio-temporal biases in the CDT estimation are presented in the extended verion due to space constraints.

CDT versus $\det(J)$**.** We first present a motivating experiment to demonstrate the rationale behind using CDTs instead of the determinants $\det(J)$ of Jacobian of the deformations, i.e., do CDTs actually carry more information? We test for group differences in longitudinal changes of the brain between groups of middle versus old aged individuals using CDTs and compare these results to those obtained via determinants. In order to avoid confounding factors in this comparison, we use the Cramér's test, a nonparametric test for univariate as well as manifold-valued data since it does not require any specification of the null distribution [51]. Fig. 3 clearly shows the improvements in statistical differences across the groups (higher sensitivity) when using CDTs (instead of $\det(J)$ maps).

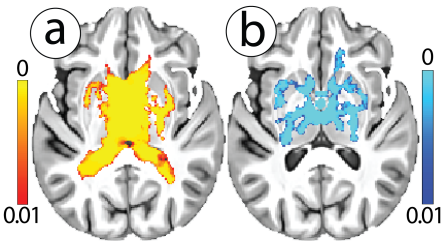

Figure 3. Results of Cramér's test showing voxels that are different between middle and old age groups ($p < 0.01$) from (a) CDTs and (b) det(J).

RNLMMs on longitudinal CDTs. We now present results using our Riemannian nonlinear mixed effects models (RNLMM) using subject specific transformation functions $\psi_i(x_{[\![ij]\!]})$ (6). Here, $x_{[\![ij]\!]} \in \mathbf{R}$ is used to represent the age of each subject at the previous visit and $y_{[\![ij]\!]} \in \mathcal{M}$, (the CDT image calculated from scans at two points). For these results, we used data from subjects who had at least three visits. We estimated our model at each voxel in the brain ($1.3M+$) using a total of $N = 228$ participants that had at least two CDT images. The maps for acceleration (α_i), spatial shift (U_i) and time shift (τ_i) for each of the subjects offer unique advantages. For instance, these maps are *not offered* by standard linear mixed effects models where only a subject specific slope or intercept is used as the random-effects (independently noted in [37]). Fig. 4 shows four representative subject-specific acceleration maps. The regions where *this specific individual* has a faster (slower) aging (or disease progression) compared to the population average rate are colored in yellow (and blue) color-scales respectively. These RNLMM maps can be used to perform additional "downstream" statistical tests using parametric tests. Here, we cover two specific examples. In Fig. 4, we show the kind of results our model can offer at the **individual level**. Fig. 4(a)-(d) shows four results, each pertaining to a different participant in the study. Fig. 4(a)-(b) show maps for

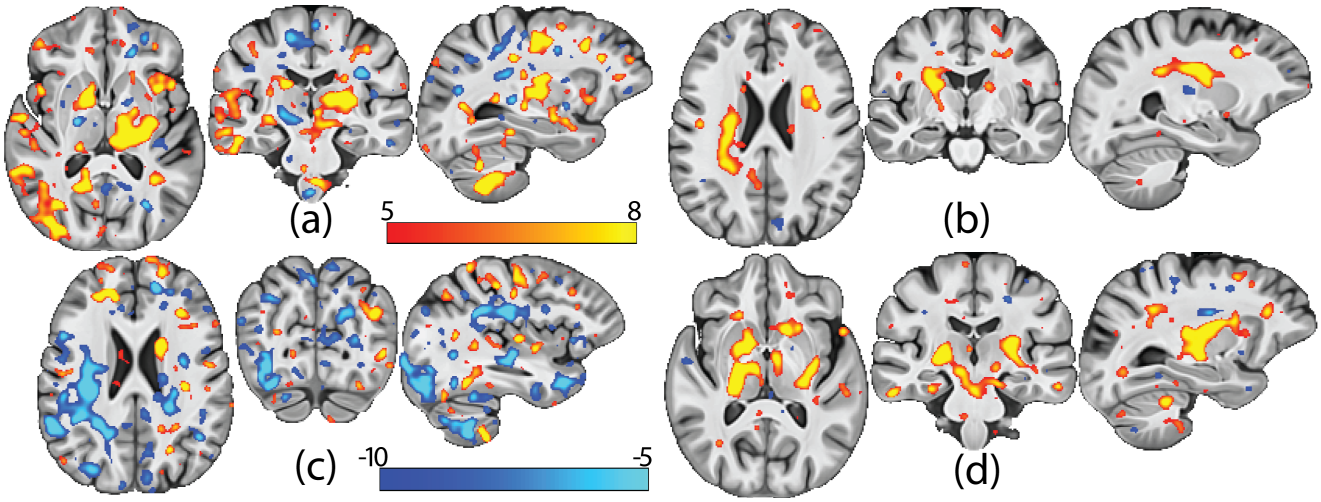

Figure 4. Representative acceleration (α_i) maps derived from our RNLMM. (a) Female, APOE-. (b) Female, APOE+. (c) Male, APOE-. (d) Male, APOE+. The male with no APOE risk shows slower progression (more blue regions) compared to the population average.

two females, whereas Fig. 4(c)-(d) show examples of two males. The color indicates the brain deformation over time (captured via acceleration), for this specific person, relative to the population. We see that a representative male (with no APOE risk) shows a slower acceleration rate (blue regions) compared to the population. Few other models in the literature can provide such *personalized assessment*.

Of course, such acceleration and spatial shift maps can also be used for **group level** analysis. We present results of Hotelling-T^2 tests on the group-wise U_i maps using the following two stratification variables: (a) males and females and (b) individuals with/without AD risk (due to APOE) [52, 53]. This enables us to identify longitudinal spatial shifts (deviations from population base points B) between these groups, shown in Fig. 5 for the gender and APOE stratification variables. The extended version provides several more analyses including synthetic experiments to show the behavior of RNLMMs in (10) and (11).

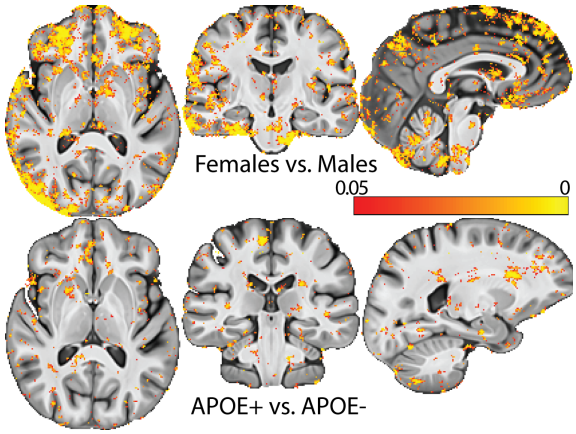

Figure 5. P-value maps of group differences in random effects (U_i). Top: Gender differences. Bottom: APOE group {APOE+, APOE-} differences. Gender differences can be effectively captured by our RNLMM.

7. Conclusion

This paper extends nonlinear mixed effects models to the setting where the responses lie on curved spaces such as the manifold of symmetric positive definite (SPD) matrices. By treating the subject-wise "non-linear warps" between consecutive time points as a field of Cauchy deformation tensors (CDT), we show how our model can facilitate longitudinal analysis that respects the geometry of such data. While the existing body of work dealing with regression models on manifold-valued data is inherently restricted to cross-sectional studies, the proposed mixed effects formulation significantly expands the operating range of the types of analyses we can conduct. For instance, the "random" effects in the construction parameterized by acceleration and spatial and time shifts offer interesting advantages. Not only can these quantities be directly used for downstream models but they also offer interpretability at the *level of individual subjects* — as an example, when conditioned on (or controlled for) race, sex and education, we can ask if a specific person's onset time of brain atrophy or *rate of atrophy*, at the level of individual voxels, deviates from the group. This capability is not currently available otherwise.

Acknowledgments. This research was supported in part by NIH grants AG040396, AG021155, BRAIN Initiative R01-EB022883, UW ADRC AG033514, UW ICTR 1UL1RR025011, Waisman IDDRC U54-HD090256, UW CPCP AI117924, and NSF grants: CAREER award 1252725 (VS), IIS 1525431 (BCV). We like to thank Andrew Schoen for providing modifications of ANTs software for execution on Condor.

References

[1] X. Pennec, P. Fillard, *et al.*, "A Riemannian framework for tensor computing," *IJCV*, vol. 66, no. 1, pp. 41–66, 2006. 1

[2] A. Srivastava, I. Jermyn, and S. Joshi, "Riemannian analysis of probability density functions with applications in vision," in *CVPR*, pp. 1–8, 2007. 1

[3] Y. Wang, H. Salehian, *et al.*, "Tracking on the product manifold of shape and orientation for tractography from diffusion mri," in *CVPR*, pp. 3051–3056, 2014. 1

[4] F. Porikli, "Regression on lie groups and its application to affine motion tracking," in *Algorithmic Advances in Riemannian Geometry and Applications*, pp. 173–185, 2016. 1

[5] O. Tuzel, F. Porikli, *et al.*, "Human detection via classification on riemannian manifolds," in *CVPR*, pp. 1–8, 2007. 1

[6] T. S. Cohen and M. Welling, "Harmonic exponential families on manifolds," in *ICML*, 2015. 1

[7] G. Lebanon *et al.*, *Riemannian geometry and statistical machine learning*. Carnegie Mellon University, Language Technologies Institute, School of Computer Science, 2005. 1, 2

[8] X. Li and G. S. Chirikjian, "Lie-theoretic multi-robot localization," in *Riemannian Computing in Computer Vision*, pp. 165–186, Springer, 2016. 1

[9] S. Joshi, B. Davis, *et al.*, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. 23, pp. S151–S160, 2004. 1

[10] A. Goh, C. Lenglet, *et al.*, "A nonparametric Riemannian framework for processing high angular resolution diffusion images HARDI," in *CVPR*, pp. 2496–2503, 2009. 1, 2

[11] H. E. Çetingül, M. J. Wright, *et al.*, "Segmentation of high angular resolution diffusion mri using sparse Riemannian manifold clustering," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 301–317, 2014. 1

[12] A. Goh and R. Vidal, "Unsupervised Riemannian clustering of probability density functions," in *ECML/KDD*, pp. 377–392, 2008. 1

[13] H. E. Cetingul and R. Vidal, "Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds," in *CVPR*, pp. 1896–1902, 2009. 1

[14] A. Goh and R. Vidal, "Clustering and dimensionality reduction on Riemannian manifolds," in *CVPR*, pp. 1–7, 2008. 1

[15] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," in *NIPS*, pp. 55–63, 2011. 1

[16] A. Srivastava, P. K. Turaga, *et al.*, *Riemannian Computing in Computer Vision*. Springer, 2016. 2

[17] P. T. Fletcher, "Geodesic regression and the theory of least squares on Riemannian manifolds," *IJCV*, vol. 105, no. 2, pp. 171–185, 2013. 2, 5

[18] E. Cornea, H. Zhu, *et al.*, "Regression models on Riemannian symmetric spaces," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. 2, 5, 6

[19] H. J. Kim, N. Adluru, *et al.*, "Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images," in *CVPR*, 2014. 2, 6

[20] N. Lepore, C. Brun, *et al.*, "Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors," *IEEE transactions on medical imaging*, vol. 27, no. 1, pp. 129–141, 2008. 2

[21] M. Banerjee, R. Chakraborty, *et al.*, "A nonlinear regression technique for manifold valued data with applications to medical image analysis," in *CVPR*, pp. 4424–4432, 2016. 2, 5

[22] H. Zhu, Y. Chen, *et al.*, "Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging," *Journal of the American Statistical Association*, vol. 104, no. 487, 2009. 2

[23] X. Shi, M. Styner, *et al.*, "Intrinsic regression models for manifold-valued data," in *MICCAI*, pp. 192–199, 2009. 2

[24] Y. Xie, B. C. Vemuri, *et al.*, "Statistical analysis of tensor fields," in *MICCAI*, pp. 682–689, 2010. 2

[25] S. Jayasumana, R. Hartley, *et al.*, "Kernel methods on the Riemannian manifold of symmetric positive definite matrices," in *CVPR*, pp. 73–80, 2013. 2

[26] Y. Xie, J. Ho, *et al.*, "On a nonlinear generalization of sparse coding and dictionary learning," in *JMLR*, vol. 28, pp. 1480–1488, 2013. 2

[27] P. Li, Q. Wang, *et al.*, "Log-Euclidean kernels for sparse representation and dictionary learning," in *ICCV*, pp. 1601–1608, 2013. 2

[28] P. T. Fletcher, C. Lu, *et al.*, "Principal geodesic analysis for the study of nonlinear statistics of shape," *Medical Imaging*, vol. 23, no. 8, pp. 995–1005, 2004. 2

[29] S. Huckemann, T. Hotz, *et al.*, "Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions," *Statistica Sinica*, vol. 20, pp. 1–100, 2010. 2

[30] S. Sommer, F. Lauze, *et al.*, "Optimization over geodesics for exact principal geodesic analysis," *Advances in Computational Mathematics*, pp. 1–31. 2

[31] H. J. Kim, N. Adluru, *et al.*, "Canonical correlation analysis on Riemannian manifolds and its applications," in *ECCV*, 2014. 2

[32] S. Sommer, "Horizontal dimensionality reduction and iterated frame bundle development," in *Geometric Science of Information*, pp. 76–83, 2013. 2

[33] B. C. Davis, P. T. Fletcher, *et al.*, "Population shape regression from random design data," in *ICCV*, pp. 1–7, IEEE, 2007. 2

[34] N. Singh, J. Hinkle, *et al.*, "A hierarchical geodesic model for diffeomorphic longitudinal shape analysis," in *IPMI*, pp. 560–571, 2013. 2

[35] M. Niethammer, Y. Huang, *et al.*, "Geodesic regression for image time-series," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–662, 2011. 2

[36] J. Hinkle, P. T. Fletcher, *et al.*, "Intrinsic polynomials for regression on Riemannian manifolds," *Journal of Mathematical Imaging and Vision*, vol. 50, no. 1-2, pp. 32–52, 2014. 2

[37] J.-B. Schiratti, S. Allassonniere, *et al.*, "Learning spatiotemporal trajectories from manifold-valued longitudinal data," in *NIPS*, pp. 2404–2412, 2015. 2, 6, 7

[38] M. P. do Carmo Valero, *Riemannian geometry*. 1992. 3

[39] J. M. Lee, "Smooth manifolds," in *Introduction to Smooth Manifolds*, pp. 1–29, 2003. 3

[40] N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," *Biometrics*, pp. 963–974, 1982. 3, 4

[41] B. B. Avants, C. L. Epstein, *et al.*, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008. 4

[42] A. Klein, J. Andersson, *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009. 4

[43] S. Durrleman, X. Pennec, *et al.*, "Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data," *IJCV*, vol. 103, no. 1, pp. 22–59, 2013. 5

[44] M. J. Lindstrom and D. M. Bates, "Nonlinear mixed effects models for repeated measures data," *Biometrics*, pp. 673–687, 1990. 5, 6

[45] H. J. Kim, J. Xu, B. C. Vemuri, and V. Singh, "Manifold-valued Dirichlet processes," in *ICML*, 2015. 5

[46] J. C. Pinheiro and E. C. Chao, "Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models," *Journal of Computational and Graphical Statistics*, 2012. 6

[47] C. Meza, F. Jaffrézic, *et al.*, "REML estimation of variance parameters in nonlinear mixed effects models using the saem algorithm," *Biometrical Journal*, vol. 49, no. 6, pp. 876–888, 2007. 6

[48] M. Moakher and P. G. Batchelor, "Symmetric positive-definite matrices: From geometry to applications and visualization," in *Visualization and Processing of Tensor Fields*, pp. 285–298, Springer, 2006. 6

[49] J. Ho, G. Cheng, *et al.*, "Recursive karcher expectation estimators and geometric law of large numbers.," in *AISTATS*, pp. 325–332, 2013. 6

[50] H. J. Kim, N. Adluru, B. B. Bendlin, S. C. Johnson, B. C. Vemuri, and V. Singh, "Canonical correlation analysis on SPD(n) manifolds," in *Riemannian Computing in Computer Vision*, pp. 69–100, Springer, 2016. 6

[51] L. Baringhaus and C. Franz, "On a new multivariate two-sample test," *Journal of multivariate analysis*, vol. 88, no. 1, pp. 190–206, 2004. 7

[52] M.-X. Tang, Y. Stern, *et al.*, "The apoe 4 allele and the risk of alzheimer disease among african americans, whites, and hispanics," *Jama*, vol. 279, no. 10, pp. 751–755, 1998. 8

[53] E. Corder, A. Saunders, *et al.*, "Gene dose of apolipoprotein E type 4 allele and the risk of alzheimers disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993. 8

Riemannian Nonlinear Mixed Effects Models: Analyzing Longitudinal Deformations in Neuroimaging (supplement)

Hyunwoo J. Kim[†] Nagesh Adluru[†] Heemanshu Suri[†]
Baba C. Vemuri[§] Sterling C. Johnson[†] Vikas Singh[†]
[†]University of Wisconsin-Madison [§]University of Florida

1. Introduction

We provide additional technical details, discussion and experimental results which were omitted in the main paper due to limited space. Specifically we present, 1) estimation of Euclidean mixed effects models, 2) estimation algorithms for RNLMMs, 3) simulations for comparison between mixed effects models and MGLMs, 4) unbiased estimation of longitudinal Cauchy deformation tensors and 5) additional backgrounds of differential geometry and geometry of SYM(n) and SPD(n).

2. Euclidean mixed effects models

We briefly discussed how to estimate mixed effects models in the Euclidean space. For linear mixed effects models in the Euclidean space, multiple numerical techniques have been proposed such as EM algorithms, Newton-Raphson methods, and MCMC. Particularly, EM algorithms treat random effects $\boldsymbol{u}$ as unobserved hidden variables and can naturally handle missing data.

As we discussed in the main paper, a linear mixed effects model is given as

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \sigma_\epsilon^2 I, \tag{1}$$

where $\boldsymbol{u} \sim \mathcal{N}(0, \tilde{\Sigma})$ and $\tilde{\Sigma} = \text{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_N) = \Sigma \otimes I$ (when $\Sigma_i = \Sigma,\ \forall i$), and $Z = \text{diag}(Z_1, Z_2, \ldots, Z_N)$.

Especially, when the variances are known, regression coefficients $\boldsymbol{\beta}$ (fixed effects) and random effects $\boldsymbol{u}$ can be estimated by a closed form solution. When $\tilde{\Sigma}$ and σ_ϵ^2 are known (or can be estimated), the mixed effects model can be rewritten as Henderson's mixed model equations (MME),

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma_\epsilon^2\tilde{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} X'\boldsymbol{y} \\ Z'\boldsymbol{y} \end{pmatrix}.$$

Thus when the variances are known, the *generalized least squares estimation* can be used as

$$\hat{\boldsymbol{\beta}} = (X^T S^{-1} X)^{-1} X^T S^{-1} \boldsymbol{y} \tag{2}$$

$$\hat{\boldsymbol{u}} = \tilde{\Sigma} Z^T S^{-1} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \tag{3}$$

where $S := \sigma_\epsilon^2 I + Z\tilde{D}Z^T$.

Applying the Gauss-Markov theorem, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ are the best linear unbiased estimates (BLUE) and predictors (BLUP), respectively [10]. Depending on the applications, the structure of covariance matrix may be differently specified as a function of some parameters, e.g., $\Sigma(\theta)$ rather than estimating the whole matrix [3].

3. Riemannian mixed effects models

Straight forward extension of linear mixed effects models on manifolds can be given by

$$y = \text{Exp}(\text{Exp}(B, Vx + Uz), \epsilon), \tag{4}$$

where $x \in \mathbf{R}^p$, $y \in \mathcal{M}$, $V \in T_B\mathcal{M}^p$ and $U \in T_B\mathcal{M}^q$, $y \in \mathcal{M}$, and $z \in \mathbf{R}^q$. In this case, approximation by linearization in the tangent space as B is simply reduced to Euclidean linear mixed effects models with tangent vectors. One problem of (4) is that all predictions are made in a tangent space. So depending on the type of manifolds, the injectivity radius might be limited. Also, exponentiating $Vx + Uz$ jointly is often less stable than exponentiating them independently.

So, in this paper, we mainly explore

$$y_{[\![ij]\!]} = \text{Exp}(\text{Exp}(\text{Exp}(B, U_i\mathbf{z}_{[\![ij]\!]}), V\boldsymbol{x}_{[\![ij]\!]}), \epsilon). \tag{5}$$

As discussed in the main paper, (5) is different from

$$y_{[\![ij]\!]} = \text{Exp}(\text{Exp}(\text{Exp}(B, V\boldsymbol{x}_{[\![ij]\!]}), U_i\mathbf{z}_{[\![ij]\!]}), \epsilon). \tag{6}$$

Exponential maps and associativity. We provide an example to show that the order of exponential maps changes models (i.e. Eqs. (5) and (6) are not equivalent). In other words, analog of associativity of addition does not hold on manifolds. For simplicity, we demonstrate this using the sphere. Let A, B, and C be the north pole, a point on the equator, and the 90° rotated B along equator. $V_1 = \text{Log}(A, B)$ $V_2 = \text{Log}(B, C)$. Then one can easily check that

$$\text{Exp}(\text{Exp}(A, V_1), V_2) = C \tag{7}$$

$$\text{Exp}(\text{Exp}(A, \Gamma_{B\to A}V_2), \Gamma_{A\to C}V_1) = B \tag{8}$$

3.1. Estimation algorithms and their derivations

In the main paper, we provided algorithms with high-level descriptions. Here we provide additional details of algorithms and their derivations.

First, we find the common longitudinal change direction.

Once the main change direction is fixed, we estimate U_i or B_i. As [13], one may limit the random effects to the orthogonal to $\boldsymbol{\eta}$.

$$U_i^* = \underset{U_i \perp \Gamma_{I\to B}(\eta)}{\operatorname{argmin}} \ \mathrm{d}(\bar{y}_i, \mathrm{Exp}(B, U_i))^2. \tag{9}$$

But in our experiments, this orthogonality constraint yields poorer goodness of fit of models and introduces random bias. So we did not use the orthogonality constraints.

The algorithm to get the longitudinal change direction given as

Algorithm 2 Calculate longitudinal change direction

1: Calculate the population Fréchet mean $\bar{\boldsymbol{y}}$ of response.
2: Calculate the Fréchet mean for each subject $\bar{\boldsymbol{y}}_i$.
3: Solve $\boldsymbol{y}^\wr_{[\![ij]\!]} = \Gamma_{\bar{\boldsymbol{y}}_i \to I}\mathrm{Log}(\bar{y}_i, \boldsymbol{y}_{[\![ij]\!]})$.
4: Solve $\boldsymbol{x}^\wr_{[\![ij]\!]} = \boldsymbol{x}_{[\![ij]\!]} - \bar{\boldsymbol{x}}_i$, where $\bar{\boldsymbol{x}}_i = \mathbb{E}_j[\boldsymbol{x}_{[\![ij]\!]}]$.
5: Collect $X^\wr = [\boldsymbol{x}^\wr_1; \ldots, \boldsymbol{x}^\wr_N]$, and $Y^\wr = [\boldsymbol{y}^\wr_1, \ldots, \boldsymbol{y}^\wr_N]$.
6: Calculate longitudinal change direction $\boldsymbol{\eta}$ by least squares estimation, $\boldsymbol{\eta}$=$((X^\wr)^T X^\wr)^{-1}((X^\wr)^T Y) \in T_I\mathcal{M}$.

In the main paper, we describe Alg. 3 to approximate the parameter estimation for the model (6). Here, we provide details of the derivations of Step 7 and Step 8.

Algorithm 3 Riemannian mixed effects models with $\psi_i(x)$

1: Calculate the Fréchet mean $\bar{y} \in \mathcal{M}$ of population.
2: Calculate the Fréchet mean for each subject $\bar{y}_i \in \mathcal{M}$.
3: Main longitudinal change direction η by algorithm (2).
4: Calculate subject-specfic base points (random effects) $B_i = \mathrm{Exp}(B, U_i^*)$, where $U_i^* = \mathrm{argmin}_{U_i}\, \mathrm{d}(\bar{y}_i, \mathrm{Exp}(B, U_i))^2 + \lambda_{U_i}\|U_i\|_B^2$.
5: $y^\wr_{ij} = \Gamma_{B_i \to I}\mathrm{Log}(B_i, y_{ij})$.
6: **while** until convergence **do**
7: Calculate the common speed of change $V = c\boldsymbol{\eta}$ and common time intercept $t_0 = b/c$ with fixed all other variables by

$$\begin{bmatrix} \sum_{ij} q_i^T q_i & \sum_{ij} p_{ij}^T q_i \\ \sum_{ij} p_{ij}^T q_i & \sum_{ij} p_{ij}^T p_{ij} \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{ij} q_i^T y^\wr_{ij} \\ \sum_{ij} p_{ij}^T y^\wr_{ij} \end{bmatrix},$$

where $b := t_0 c$, $q_i := \eta(1-\alpha_i)$,$p_{ij} := \eta(\alpha_i x_{ij} - \alpha_i \tau_i)$.

8: Given V, t_0, calculate the subject-specific acceleration α_i, and time-shift τ_i by generalized least square estimation with the priors for α_i and $\tau_i = d_i/\alpha_i$

$$\begin{bmatrix} \sum_j W_{ij}^T W_{ij} & -\sum_j W_{ij}^T V \\ \sum_j W_{ij}^T V & -\sum_j V^T V \end{bmatrix} \begin{bmatrix} \alpha_i \\ d_i \end{bmatrix} = \begin{bmatrix} \sum_j \Upsilon_{ij}^T W_{ij} \\ \sum_j \Upsilon_{ij}^T V \end{bmatrix},$$

where $\Upsilon_{ij} := y^\wr_{ij} - Vt_0$, $W_{ij} := V(X_{ij} - t_0)$ and $d_i = \alpha_i\tau_i$.
9: **end while**

Derivation of step 7.

$$\begin{aligned} y_{ij} &= \mathrm{Exp}(B_i, \Gamma_{I\to B_i}(c\eta)(\alpha_i(x_{ij} - \tau_i - t_0) + t_0)) \\ \mathrm{Log}(B_i, y_{ij}) &= \Gamma_{I\to B_i}(c\eta)(\alpha_i(x_{ij} - \tau_i - t_0) + t_0) \\ y^\wr_{ij} &= c\eta(\alpha_i x_{ij} - \alpha_i\tau_i - \alpha_i t_0 + t_0), \text{ by } \Gamma_{B_i\to I} \\ &= \eta(\alpha_i x_{ij} - \alpha_i\tau_i)c + \eta(1-\alpha_i)t_0 c \end{aligned}$$

where $b := t_0 c$. Let $q_i := \eta(1-\alpha_i)$
and $p_{ij} := \eta(\alpha_i x_{ij} - \alpha_i\tau_i)$.

$$\begin{aligned} &\underset{b,c}{\operatorname{argmin}} \sum_{ij} (y^\wr_{ij} - p_{ij}c - q_i b)^T (y^\wr_{ij} - p_{ij}c - q_i b) \\ &= \underset{b,c}{\operatorname{argmin}} \sum_{ij} (y^\wr_{ij})^T y^\wr_{ij} + c^2 p_{ij}^T p_{ij} + b^2 q_i^T q_i - 2c(y^\wr_{ij})^T p_{ij} \\ &\qquad + 2bcp_{ij}^T q_i - 2bq_i^T y^\wr_{ij} \end{aligned}$$

Take the partial derivatives

$$\frac{\partial}{\partial b} = -\sum_{ij} q_i^T y^\wr_{ij} + c\sum_{ij} p_{ij}^T q_i + b\sum_{ij} q_i^T q_i = 0 \tag{10}$$

$$\frac{\partial}{\partial c} = -\sum_{ij} p_{ij}^T y^\wr_{ij} + c\sum_{ij} p_{ij}^T p_{ij} + b\sum_{ij} p_{ij}^T q_i = 0 \tag{11}$$

So the system of equations from the KKT condition is the following.

$$\begin{bmatrix} \sum_{ij} q_i^T q_i & \sum_{ij} p_{ij}^T q_i \\ \sum_{ij} p_{ij}^T q_i & \sum_{ij} p_{ij}^T p_{ij} \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{ij} q_i^T y^\wr_{ij} \\ \sum_{ij} p_{ij}^T y^\wr_{ij} \end{bmatrix}. \tag{12}$$

Derivation for step 8. Given $V := c\eta$, and t_0,

$$\begin{aligned} y_{ij} &= \mathrm{Exp}(B_i, \Gamma_{I\to B_i}(V)(\alpha_i(X_{ij} - \tau_i - t_0) + t_0)) \\ y^\wr_{ij} &= V(\alpha_i(X_{ij} - \tau_i - t_0) + t_0), \text{the same trick} \\ y^\wr_{ij} &= V(\alpha_i X_{ij} - \alpha_i\tau_i - \alpha_i t_0 + t_0) && (13) \\ y^\wr_{ij} - Vt_0 &= V\alpha_i X_{ij} - V\alpha_i\tau_i - V\alpha_i t_0 && (14) \\ \Upsilon_{ij} &= W_{ij}\alpha_i - Vd_i && (15) \\ & && (16) \end{aligned}$$

where $\Upsilon_{ij} := y^\wr_{ij} - Vt_0$, $W_{ij} := V(X_{ij} - t_0)$ and $d_i = \alpha_i\tau_i$

$$\underset{j}{\operatorname{argmin}} \sum (\Upsilon_{ij} - W_{ij}\alpha_i + Vd_i)^T (\Upsilon_{ij} - W_{ij}\alpha_i + Vd_i) \tag{17}$$

$$= \underset{j}{\operatorname{argmin}} \sum \Upsilon_{ij}^T\Upsilon_{ij} + W_{ij}^T W_{ij}\alpha_i^2 + V^T V d_i^2 \tag{18}$$

$$- 2\Upsilon_{ij}^T W_{ij}\alpha_i + 2\Upsilon_{ij}^T V d_i - 2W_{ij}^T V\alpha_i d_i \tag{19}$$

Take the derivatives

$$\frac{\partial}{\partial \alpha_i} = \sum \alpha_i W_{ij}^T W_{ij} - \Upsilon_{ij} W_{ij} - W_{ij}^T V d_i \quad (20)$$

$$\frac{\partial}{\partial d_i} = \sum V^T V d_i + \Upsilon_{ij}^T V - W_{ij}^T V \alpha_i \quad (21)$$

The system of equations is given as

$$\begin{bmatrix} \sum_j W_{ij}^T W_{ij} & -\sum_j W_{ij}^T V \\ \sum_j W_{ij}^T V & -\sum_j V^T V \end{bmatrix} \begin{bmatrix} \alpha_i \\ d_i \end{bmatrix} = \begin{bmatrix} \sum_j \Upsilon_{ij}^T W_{ij} \\ \sum_j \Upsilon_{ij}^T V \end{bmatrix}. \quad (22)$$

Alg. 3 utilizes noninformative priors. With minor changes, we can easily incorporate normal distribution priors, It turns out that if the response y has a generalized normal distributed noise on manifolds, as in the Euclidean space, the least-squares estimator (first term in (23)) is the same as the maximum likelihood estimator [5].

$$\min_{c,t_0,\{B_i,U_i,\alpha_i,\tau_i\}} \sum_{ij} \mathrm{d}(y_{[\![ij]\!]}, \mathrm{Exp}(B_i, \Gamma_{I\to B_i}(c\eta)(\psi_i(\boldsymbol{x}_{[\![ij]\!]}))))^2 + \lambda_u \sum_i \|U_i\|_B^2 + \lambda_\alpha \sum_i \|\alpha_i - 1\|^2 + \lambda_\tau \sum_i \|\tau_i\|^2, \quad (23)$$

where $\psi_i(\boldsymbol{x}_{[\![ij]\!]}) := \alpha_i(\boldsymbol{x}_{[\![ij]\!]} - \tau_i - t_0) + t_0$; $B_i = \mathrm{Exp}(B, U_i)$.

Remarks. Notice the regularizers in (23) comprise of the last three terms. This is based on the fact that MLE estimation of linear regression with a normal distributional prior for the coefficients is equivalent to the ridge regression estimate [7]. For inference with priors, Step 4 and 7-8 in Alg. 3 need to be substituted with ridge regression.

4. Additional details on experiments

4.1. Simulations

In this section, we demonstrate the difference between our Riemannian nonlinear mixed effects model in (24) and MGLMs.

$$y_{[\![ij]\!]} = \mathrm{Exp}(\mathrm{Exp}(B_i(r), \Gamma_{B\to B_i(r)}(V)(\boldsymbol{x}_{[\![ij]\!]} - \tau_i(r))), \epsilon). \quad (24)$$

The model in (24) is reduced to MGLM [2] as $r \to 0$ since all the subjects have the same intercept and same slope V, see the trend in Fig. 1,2,3.

4.2. Cauchy deformation tensors (CDTs)

Recall that Cauchy deformation tensors are derived from a nonlinear deformation $\Phi(\mathtt{vox})$ for voxels (spatial locations) $\mathtt{vox} \in \Omega$ for each image (rather, for each $(\mathcal{I}_{i,j+1}, \mathcal{I}_{i,j})$ pair) is given as

$$\Phi : \mathcal{I}_{i,j+1} \to \mathcal{I}_{i,j} \quad (25)$$

$$\Phi(\mathtt{vox} + d\mathtt{vox}) = \Phi(\mathtt{vox}) + J(\mathtt{vox}) d\mathtt{vox} + \mathcal{O}(d\mathtt{vox}^2),$$

where $J(\mathtt{vox})$ denotes the Jacobian of the deformations at position $\mathtt{vox}$. A nice property of CDTs is that it preserves the determinant of $J(\mathtt{vox})$, since $\det(J(\mathtt{vox})) > 0$. So, a CDT representation introduced in the main, nicely symmetrizes $J(\mathtt{vox})$ without affecting the volumetric change information, i.e., $\det(J) = \det(\sqrt{J^T J})$. To prove this there are few assumptions which are commonly made by registraion algorithms to get a nonlinear deformation. In neuroimaging applications, registration algorithms generally assume that deformations are *diffeomorphic* and *orientation preserving*. A *diffeomorphism* requires a Jacoian matrix of deformation to be invertible. Orientation preserv-

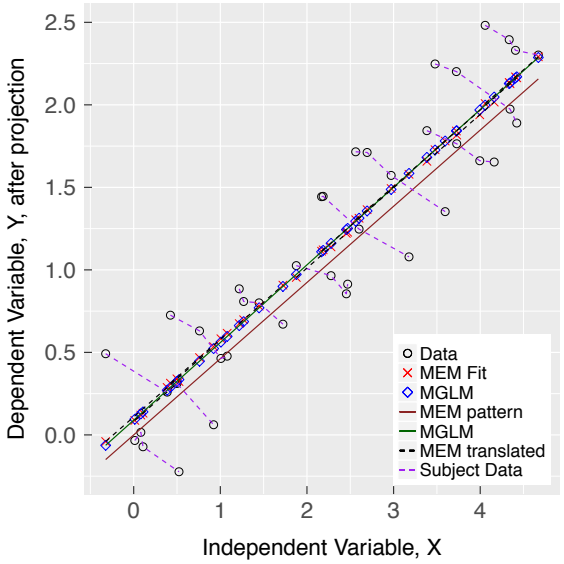

Figure 1. Mixed Effects Model on manifolds ($r = 0.1$)

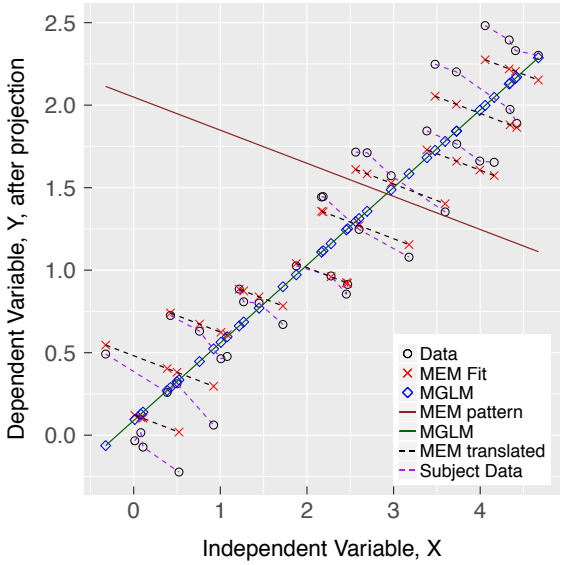

Figure 2. Mixed Effects Model on manifolds ($r = 0.9$)

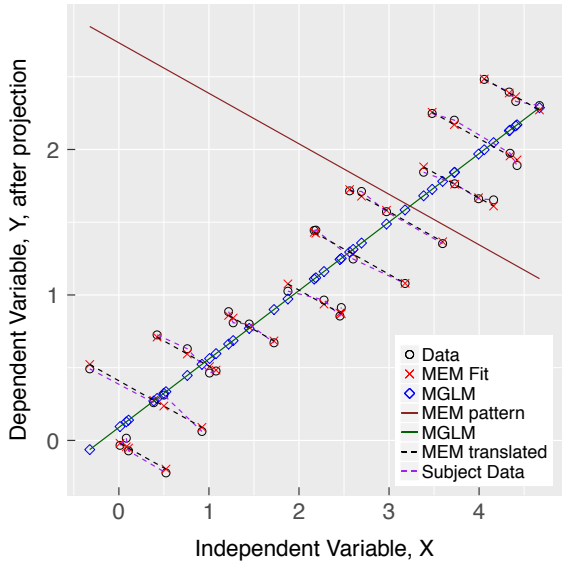

Figure 3. Mixed effects model on manifolds ($r = 1$). When the variability over subjects is large than variability of a subject over time, the linear model (MGLM in blue) captures the overall relationship between x and y. In this example, the correlation captured by MGLM [2] is an opposite direction to the trajectory of each subject. Each subject is measured four times and the trajectory correctly captured by the mixed effects model (in gray dot lines). The mixed effects model control the variability between subjects as random effects and captured the common longitudinal pattern (in red). The common longitudinal change is translated to each subject (in gray dot lines).

ing implies that the determinants of Jacobian matrices are positive. Hence, the spatial gradient of deformation, (Jacobian $J(\mathtt{vox})$), forms a subgroup of general linear group, $\mathrm{GL}^+(n)$, which is a subgroup of invertible matrices with positive determinants, where n is the number of rows (or columns) of a matrix.

More explicitly each Jacobian matrix can be written as

$$J(\mathtt{vox}) = \mathcal{D}\Phi|_{\mathtt{vox}} = \begin{pmatrix} \partial_1\phi^1|_{\mathtt{vox}} & \partial_2\phi^1|_{\mathtt{vox}} & \partial_3\phi^1|_{\mathtt{vox}} \\ \partial_1\phi^2|_{\mathtt{vox}} & \partial_2\phi^2|_{\mathtt{vox}} & \partial_3\phi^2|_{\mathtt{vox}} \\ \partial_1\phi^3|_{\mathtt{vox}} & \partial_2\phi^3|_{\mathtt{vox}} & \partial_3\phi^3|_{\mathtt{vox}} \end{pmatrix}. \tag{26}$$

where $\mathcal{D}$ is the Jacobian operator (derivative of vector field) and $\partial_i\phi^j$ is a derivative along i and j component of Ω and Γ respectively.

Now, CDTs can be derived from $J(x)$ with matrix operation

$$\mathrm{CDT}(\mathtt{vox}) = \sqrt{J(\mathtt{vox})^T J(\mathtt{vox})}, \tag{27}$$

where $\sqrt{(\cdot)}$ is matrix square root. As mentioned above, one nice property of CDTs is that it preserves the determinant of $J(\mathtt{vox})$, since $\det(J(\mathtt{vox})) > 0$. So, CDT transformation nicely symmetrizes $J(\mathtt{vox})$ without changing information of volumetric changes, i.e., $\det(J) = \det(\sqrt{J^T J})$.

Lemma 1 $\det(J) = \det(\sqrt{J^T J})$

Proof: In general, the square root of a matrix can be multiple. Fortunately, positive (semi) definite matrix has a unique positive (semi) definite square root matrix [8]. Also the square root matrix of a symmetric positive (semi) definite matrix can be written as

$$X^{1/2} = VD^{1/2}V^T, \text{where } X = X^T, X \succeq 0, X = VDV^T \tag{28}$$

So, let $X = J^T J$. Then, since $X \succeq 0, X^T = X$, we have $\det(\sqrt{J^T J}) = \det(X^{1/2}) = \prod_i \sqrt{D_{ii}} = \sqrt{\prod_i D_{ii}} = \sqrt{\det(X)} = \sqrt{\det(J^T)\det(J)} = \sqrt{\det(J)\det(J)} = \det(J)$. □

4.3. Unbiased estimation of CDTs

We present experimental details on deriving the deformation fields and then the CDT images which capture the subject-wise longitudinal changes. These CDT images will be in a least biased coordinate system to allow for voxel-wise analysis of morphometric longitudinal changes. We first estimate an unbiased global template space as shown in Fig. 4. While estimating an unbiased atlas (coordinate system) is well investigated in cross-sectional imaging studies, there are fewer validation studies for the 3D+time regime. The additional bias which we must restrict in a longitudinal study is the interpolation asymmetry that can arise when selecting only one of the time points as a temporal representative in generating the population/study level coordinate system. Based on the current best practices, we first estimate a subject-specific average that is temporally unbiased. The subject-specific averages are then used to generate an unbiased population level average template space (Fig. 4). Each of the curved black lines represents a combination of affine and non-linear diffeomorphic transformations. These transformations and the spatial averages are estimated iteratively until convergence using ANTS [1, 15] based implementation.

So far, we have only described constructing a common coordinate system. N ext we describe how the longitudinal deformation fields are generated in the global coordinate system for voxel-wise analyses. This step turns out to be non-trivial for CDT images. Most existing publicly available pipelines such as SPM, Freesurfer, AFNI, FSL do not generate such results and instead provide scalar/univariate images representing longitudinal magnitude of Jacobians or divergences of the deformation fields. Note that the widely used Freesurfer processing streams allows subject-level longitudinal analysis of features such as rate of change of thickness on the cortical surfaces but not the morphological changes themselves. While such summary measures of structures are relevant, in the case of preclinical AD, they are complementary to morphological changes captured at the *voxel-level*.

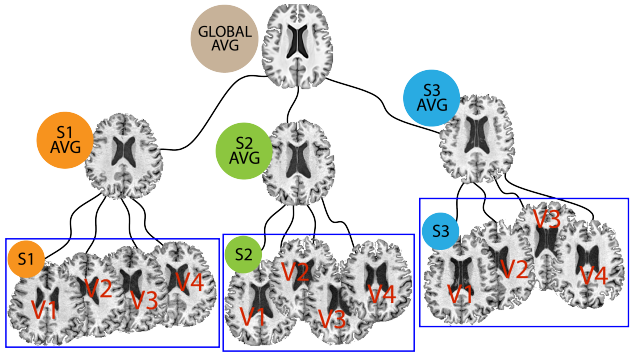

Figure 4. Schematic for generating least biased global coordinate system for the longitudinally acquired imaging data. Visits V1-V4 are averaged first which are then used to estimate the global average.

Fortunately based on the work in gravitation theory Lorenzi and Pennec recently developed a computationally efficient framework for parallel transport of stationary velocity fields along other stationary velocity fields (SVFs) [12]. Using this framework we can obtain *longitudinal* deformations in a global coordinate system. We first register the set of longitudinal images (using rigid transformations - rotations and translations) from all visits (V_i) to the global average (GA) estimated as described in the previous paragraph i.e. $V_i \mapsto_R GA, \quad \forall i$. We thus have V_i^R and GA in the same global coordinate system. Now the key non-linear symmetric diffeomorphic deformations are generated using [11]. Images are registered pairwise between consecutive visits, i.e. $V_{i+1}^R \hookrightarrow_{SVF} V_i^R$) resulting in a stationary velocity field (SVF) ($\mathcal{V}^{SVF}_{(i+1)\mapsto i}$) representing longitudinal progression between visits $i+1$ and i. The individual visit images V_i^R are non-linearly registered to GA ($V_i^R \hookrightarrow_{SVF} GA$) resulting in a "subject-to-template" SVF, $\mathcal{V}_i^{SVF}$. The $\mathcal{V}^{SVF}_{(i+1)\mapsto i}$ are then parallel transported in the direction of $\mathcal{V}_i^{SVF}$ resulting in SVFs that represents longitudinal progression in the global coordinate system. Cauchy deformation tensor fields are constructed from Jacobians of these final transported vector fields.

4.4. Data

Fig. 5 shows the age distributions of the subjects analyzed in this project. We can see that most subjects are between the age of 50 and 70. A key important feature is the actual intervals between consecutive visits for all the subjects. These intervals determine how much of longitudinal progression one can capture in a study. If the intervals are too narrow for an age range the progression features might be too weak in detecting signal and on the other hand if the intervals are too broad then various confounding factors can hinder specificity and interpretability of findings. Fig. 6 shows the distribution of age intervals in years. We can

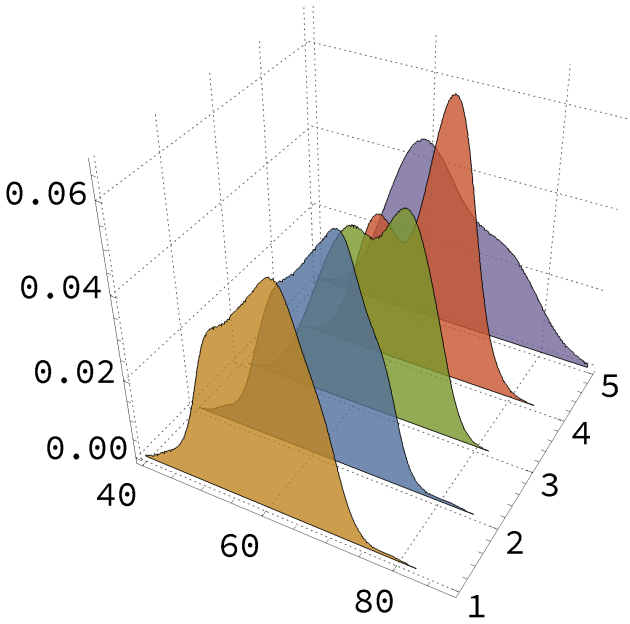

Figure 5. Longitudinal evolution of the age distributions of the subjects from the study for five visits.

notice that except for a few whose intervals are around 5-6 years of age most consecutive visits are about 2 years apart thus reducing various confounding factors involved in aging.

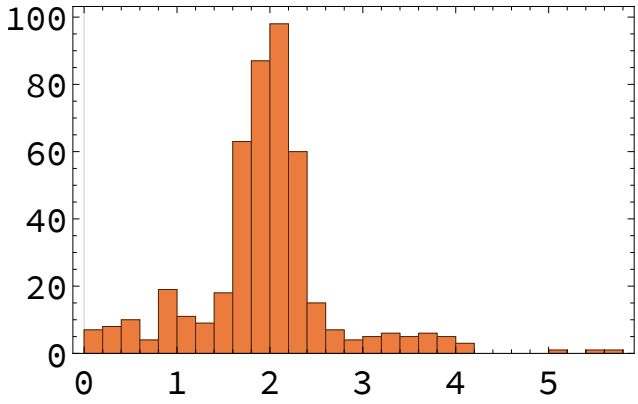

Figure 6. Distribution of intervals (in age years) between consecutive visits for the 100 individuals analyzed in our experiments.

Our RNLMMs can be used to derive various progression based features that can be used to investigate subjects on different cognitive spectrum. Fig. 7 shows the spread of the subjects on the progression coordinate space. This progression space is spanned by the number of fast (> 5) and slow (< -5) progressing voxels derived from our acceleration maps. The subjects are grouped based on consensus conference scores [6] and are labeled as "cognitive decliners" if during any of the visits they show a decrease in cognition scores. Although most decliners and non-decliners are not

separable in the progression coordinate space, we can detect special individuals whose decline is much more severe for example the individual highlighted in the red square in Fig. 7 who has higher cognitive decline compared to the others in the population.

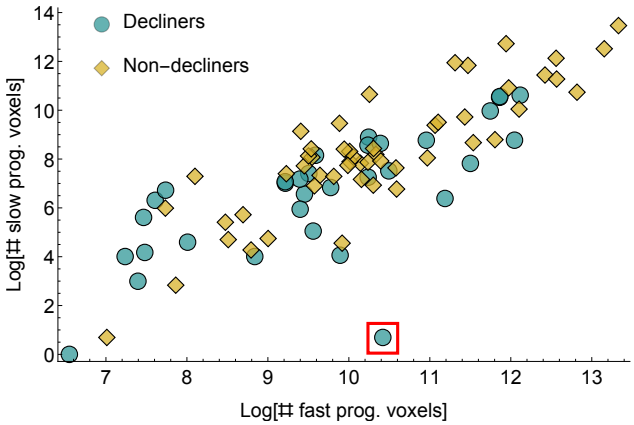

Figure 7. Scatter plots of the individuals in the progression coordinate space color-coded by cognitive decline status as derived from consensus conference score. Such coordinate space allows us to identify individuals such as the one highlighted in the red rectangle who deviate significantly from population in terms of cognitive or clinical covariates. We would like to highlight that not all coordinate spaces are able to identify such clinically meaningful deviations.

5. Differential geometry background

We briefly introduce notions that we used in the main paper. For more details, we refer the reader to [4, 9]. We first mention that we use slightly different notations from standard notations in differential geometry text books. Since here we use multiply nested equations, the standard notation $\mathrm{Exp}_B V$ may be harder to read. So instead we use $\mathrm{Exp}(B, V)$ in this paper. i.e., we use $\mathrm{Exp}(\mathrm{Exp}(B, V), \epsilon)$ instead of $\mathrm{Exp}_{\mathrm{Exp}_p V}\epsilon$ in this paper. Also to avoid the confusion by multiple different exponentiations, we use $\exp(\cdot)$ for the exponentiation of a real number or matrix, i.e, $\exp(X) := \sum \frac{1}{k!} X^k$, where $0! = 1$ and $X^0 = I$.

A *differentiable (smooth) manifold* of dimension n is a set $\mathcal{M}$ and a maximal family of *injective* mappings $\varphi_i : U_i \subset \mathbf{R}^n \to \mathcal{M}$ of open sets U_i of $\mathbf{R}^n$ into $\mathcal{M}$ such that: (1) $\cup_i \varphi_i(U_i) = \mathcal{M}$ (2) for any pair i, j with $\varphi_i(U_i) \cap \varphi_j(U_j) = W \neq \phi$, the sets $\varphi_i^{-1}(W)$ and $\varphi_j^{-1}(W)$ are open sets in $\mathbf{R}^n$ and the mappings $\varphi_j^{-1} \circ \varphi_i$ are differentiable, where $\circ$ denotes function composition. (3) The family $\{(U_i, \varphi_i)\}$ is maximal relative to the conditions (1) and (2).

Roughly speaking, a differentiable (smooth) manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure. A **Riemannian manifold** is a smooth manifold equipped with a smoothly varying metric (inner product), which is called **Riemannian metric**. The **tangent space** at $B \in \mathcal{M}$ (denoted by $T_B\mathcal{M}$) is the vector space, which comprises the tangent vectors of *all* possible curves passing through B. The **tangent bundle** $T\mathcal{M}$ of manifold $\mathcal{M}$ is the disjoint union of tangent spaces at all points of $\mathcal{M}$, i.e., $T\mathcal{M} = \coprod_{B\in\mathcal{M}} T_B\mathcal{M}$. The tangent bundle is equipped with a natural **projection map** $\pi : T\mathcal{M} \to \mathcal{M}$. Various geometric notions, e.g., the angle between two curves or the length of a curve, can be extended on the manifold. A **geodesic curve** on a Riemannian manifold is the locally (NOT globally) shortest (distance-minimizing) curve. In general, geodesic curves between two points on a manifold may not be unique. Geodesic curves are the analog of straight lines in Euclidean space, which is used to generalize linear models to Riemannian manifolds. On manifolds, using the length of the shortest geodesic curve connecting the two points, distance between two points is measured. It is called **geodesic distance** defined as the infimum of the length of all differentiable curves connecting the two points. Let γ be a continuously differentiable curve $\gamma : [a, b] \to \mathcal{M}$ between p and q in $\mathcal{M}$ and g be a metric tensor in Mc. Then, formally, the distance between p and q is defined as

$$\mathrm{d}(p, q) := \inf_{\gamma} \int_a^b \sqrt{g_\gamma(t)(\dot\gamma(t), \dot\gamma(t))} dt \qquad (29)$$

where $\gamma(a) = p$ and $\gamma(b) = q$.

An **exponential map** maps tangent vectors in a tangent space $T_B\mathcal{M}$ to $\mathcal{M}$ preserving the length of the geodesic curve. It is usually locally defined due to the existence and uniqueness of ordinary differential equation (ODE) for the map. Using the exponential map, the geodesic curve can be parameterized by a tangent vector, e.g., $\mathrm{Exp}(B, \cdot) : T_B\mathcal{M} \to \mathcal{M}$. The **logarithm map** is the inversion of exponential map, $\mathrm{Log}(y_i, \cdot) : \mathcal{M} \to T_{y_i}\mathcal{M}$. For completeness, Table 1 shows corresponding operations in the Euclidean space and Riemannian manifolds. In the main paper, for the readability when operations are multiply nested, exponential map and its inverse logarithm map are denoted by $\mathrm{Exp}(B, x)$ and $\mathrm{Log}(B, v)$ respectively, where $B, x \in \mathcal{M}$ and $v \in T_B\mathcal{M}$. They are usually denoted $\exp_B(x)$ and $\log_B(v)$ in most of differential geometry books.

The **intrinsic (or Karcher) mean** is the minimizer to

$$\bar y = \arg\min_{y\in\mathcal{M}} \sum_{i=1}^N d(y, y_i)^2, \qquad (30)$$

which may be an arithmetic, geometric or harmonic mean depending on distance metric $d(\cdot, \cdot)$. Usually, a local minimum to (30) is called Karcher mean and a global minimum is referred as a Fréchet mean.

Algorithm 1 : Karcher mean
Input: $y_1, \ldots, y_N \in \mathcal{M}$, α
Output: $\bar{y} \in \mathcal{M}$
$\bar{y}_0 = y_1$
while $\| \sum_{i=1}^{N} \text{Log}(\bar{y}_k, y_i) \| > \epsilon$ **do**
$\Delta\bar{y} = \frac{\alpha}{N} \sum_{i=1}^{N} \text{Log}(\bar{y}_k, y_i)$
$\bar{y}_{k+1} = \text{Exp}(\bar{y}_k, \Delta\bar{y})$
end while

Figure 8. Karcher mean on manifolds

This identity implies the first order necessary condition of (30), i.e., $\bar{y}$ is a local minimum with a zero norm gradient. In general, on manifolds, the existence and uniqueness of th.e Karcher mean is not guaranteed unless we assume, for uniqueness, that the data is in a small neighborhood.

Parallel transport is the analog of parallel translation in the Euclidean space. Parallel transport maps a tangent vector V along a differentiable curve $c(t)$ preserving the length of the tangent vector and angle between $\dot{c}(t)$ and V.

5.1. Geometry of SPD manifolds and priors

We provide basic properties of SPD(n). iThe SPD manifold is the space of symmetric positive definite matrices of size $n \times n$, which is a quotient space $GL(n)/O(n)$. iThe inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is defined as

$$\langle u, v \rangle_p = \text{tr}(p^{-1/2} u p^{-1} v p^{-1/2}) \tag{31}$$

The tangent space of SPD(n) is the space of symmtric matrices of dimension $(n+1)n/2$.

The exponential map and logarithm map are given as

$$\begin{aligned} \text{Exp}(p, v) &= p^{1/2} \exp(p^{-1/2} v p^{-1/2}) p^{1/2}, \\ \text{Log}(p, q) &= p^{1/2} \log(p^{-1/2} q p^{-1/2}) p^{1/2}. \end{aligned} \tag{32}$$

The geodesic distance is $d(p, q)^2 = \text{tr}(\log^2(p^{-1/2} q p^{-1/2}))$.

The parallel transport of $w := \text{Log}(p, q)$ along the geodesic from p to q is given by

$$\begin{aligned} &\Gamma_{p \to q}(w) = p^{1/2} r p^{-1/2} w p^{-1/2} r p^{1/2} \\ &\text{where } r = \exp\left(p^{-1/2} \frac{v}{2} p^{-1/2}\right) \text{ and } v = \text{Log}(p, q) \end{aligned} \tag{33}$$

Operation	Euclidean	Riemannian
Subtraction	$\overrightarrow{x_i x_j} = x_j - x_i$	$\overrightarrow{x_i x_j} = \text{Log}(x_i, x_j)$
Addition	$x_i + \overrightarrow{x_j x_k}$	$\text{Exp}(x_i, \overrightarrow{x_j x_k})$
Distance	$\|\overrightarrow{x_i x_j}\|$	$\|\text{Log}(x_i, x_j)\|_{x_i}$
Mean	$\sum_{i=1}^{n} \overrightarrow{\bar{x} x_i} = 0$	$\sum_{i=1}^{n} \text{Log}(\bar{x}, x_i) = 0$
Covariance	$\mathbb{E}\left[(x_i - \bar{x})(x_i - \bar{x})^T\right]$	$\mathbb{E}\left[\text{Log}(\bar{x}, x)\text{Log}(\bar{x}, x)^T\right]$

Table 1. Basic operations in Euclidean space and Riemannian manifolds.

5.2. Distributions for matrix-valued parameters

Symmetric matrices $U_i \in \text{Sym}(n)$ describe random effects in our model and they are tangent vectors on SPD(n). For U_i, we may use mulitple normal distributions. Let $q = n(n+1)/2$.

Normal distribution 1.

$$f(X) = (2\pi)^{q/2} \exp(-1/2 tr(X^2)), \tag{34}$$

where mean matrix is 0 and covariance matrix is I.

Normal distribution 2.

$$f(X; M, \Sigma) = \frac{1}{(2\pi)^{q/2} |\Sigma|^{(n+1)/2}} \exp\left(-\frac{1}{2}\text{tr}((X - M)\Sigma^{-1})\right)^2 \tag{35}$$

where M is a mean matrix and Σ is a covariance matrix. More distributions are discussed in [14]

References

[1] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008. 4

[2] E. Cornea, H. Zhu, et al. Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. 3, 4

[3] E. Demidenko. *Mixed models: theory and applications with R*. John Wiley & Sons, 2013. 1

[4] M. P. do Carmo Valero. *Riemannian geometry*. 1992. 6

[5] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *IJCV*, 105(2):171–185, 2013. 3

[6] N. L. Foster, K. Wilhelmsen, and othres. Frontotemporal dementia and Parkinsonism linked to chromosome 17: a consensus conference. *Annals of neurology*, 41(6):706–715, 1997. 5

[7] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 3

[8] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012. 4

[9] J. M. Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–29. 2003. 6

[10] M. J. Lindstrom and D. M. Bates. Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988. 1

[11] M. Lorenzi, N. Ayache, G. B. Frisoni, X. Pennec, A. D. N. I. (ADNI, et al. Lcc-demons: a robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage*, 81:470–483, 2013. 5

[12] M. Lorenzi and X. Pennec. Efficient parallel transport of deformations in time series of images: from schilds to pole ladder. *Journal of Mathematical Imaging and Vision*, 50(1-2):5–17, 2014. 5

[13] J.-B. Schiratti, S. Allassonniere, et al. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *NIPS*, pages 2404–2412, 2015. 2

[14] A. Schwartzman. *Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data*. PhD thesis, Stanford University, 2006. 7

[15] N. J. Tustison and B. B. Avants. Explicit b-spline regularization in diffeomorphic image registration. *Front. Neuroinform*, 7(39), 2013. 4