Multivariate General Linear Models (MGLM) on Riemannian Manifolds with Applications to Statistical Analysis of Diffusion Weighted Images

Hyunwoo J. Kim Nagesh Adluru Maxwell D. Collins Moo K. Chung
Barbara B. Bendlin Sterling C. Johnson Richard J. Davidson Vikas Singh
University of Wisconsin-Madison
http://pages.cs.wisc.edu/~hwkim/projects/riem-mglm

Abstract

Linear regression is a parametric model which is ubiquitous in scientific analysis. The classical setup where the observations and responses, i.e., (x_i, y_i) pairs, are Euclidean is well studied. The setting where y_i is manifold valued is a topic of much interest, motivated by applications in shape analysis, topic modeling, and medical imaging. Recent work gives strategies for max-margin classifiers, principal components analysis, and dictionary learning on certain types of manifolds. For parametric regression specifically, results within the last year provide mechanisms to regress one real-valued parameter, $x_i \in \boldsymbol{R}$, against a manifold-valued variable, $y_i \in \mathcal{M}$. We seek to substantially extend the operating range of such methods by deriving schemes for multivariate multiple linear regression — a manifold-valued dependent variable against multiple independent variables, i.e., $f : \boldsymbol{R}^n \to \mathcal{M}$. Our variational algorithm efficiently solves for multiple geodesic bases on the manifold concurrently via gradient updates. This allows us to answer questions such as: what is the relationship of the measurement at voxel y to disease when conditioned on age and gender. We show applications to statistical analysis of diffusion weighted images, which give rise to regression tasks on the manifold $GL(n)/O(n)$ for diffusion tensor images (DTI) and the Hilbert unit sphere for orientation distribution functions (ODF) from high angular resolution acquisition. The companion open-source code is available on nitrc.org/projects/riem_mglm.

1. Introduction

Regression is ubiquitous in scientific analysis to identify how a dependent variable, $y \in \mathcal{Y}$ relates to an independent variable, $x \in \mathcal{X}$. Here, we are provided training data in the form, $(x_i, y_i)_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$, and seek to find the best model that explains these observations, given an appropriate loss function. The classical setting typically makes an assumption on the geometric structure of the data by capturing the notion of distance between points a and b by the expression, $\left(\sum_{j=1\ldots n}(a^j - b^j)^2\right)^{\frac{1}{2}}$, which holds whenever the data are real vectors. In the Euclidean setting, a simple parametric model, $y_i = \alpha + \beta x_i + \epsilon_i$, suffices to identify the linear relationship between scalar-valued $x_i \in \mathcal{X}$ and the dependent (i.e., response) variable $y_i \in \mathcal{Y}$ with error ϵ_i. The least squares estimate is,

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{N} \|y_i - \alpha - x_i\beta\|^2, \tag{1}$$

and the closed form solution to (1) is obtained as,

$$\hat{\beta} = \frac{\text{Cov}[x,y]}{\text{Var}[x]} = \frac{\mathbb{E}\left[(x-\bar{x})(y-\bar{y})\right]}{\mathbb{E}\left[(x-\bar{x})^2\right]}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \tag{2}$$

If x and y are multivariates, one can easily replace the multiplication and division with an outer product of vectors and matrix inversion respectively, and obtain an analytical solution.

Despite the simplicity of the above solution, it is not universally applicable. Occasionally, the variables that constitute the data are not in $\mathbf{R}^n$ and imposing the Euclidean topology forcibly can lead to poor estimation. For example, parametric families such as Dirichlet or multinomial cannot be treated as Euclidean spaces [18]. Similarly, in shape analysis, arbitrary linear combinations of shapes do not constitute a 'valid' shape — instead, feasible solutions are only those that correspond to points on a *shape* manifold. Covariance matrices in machine learning and diffusion tensors in neuroimaging are positive definite matrices. Operations that are simple in the Euclidean space such as taking the arithmetic mean cannot be applied directly to such data because of the absence of an additive structure. Manifold valued variables arise frequently in data analysis. While the relationship between statistics and Riemannian geometry is well known, in the last two decades, one can see a concerted effort to move beyond rigorous asymptotic results, and use these ideas to perform inference on real-world derived datasets in a more principled way. This has led to a

developing body of practical methods to operate on differentiable manifolds which are smooth and locally Euclidean and endowed with a metric given as a *local* inner product between tangent vectors.

Related work. There is a mature body of work in statistics dating back to the seminal work of Rao [21] and Efron [7] showing how the asympotic properties of a statistical model relate to the curvature of a corresponding manifold. But since the 1990s, such concepts have been leveraged towards characterizing problems in statistical learning and vision. On the machine learning front, Amari [1] showed how to use differential geometry and Fisher information principles for neural networks, whereas Jaakkola [12] derived hybrid schemes combining properties of generative and discriminative methods based on geometric principles. There are numerous learning theoretic results [17] examining geometric insights into the behavior of AdaBoost and large margin classifiers [18]. Independently, research in vision/medical imaging has studied manifold regression for shape analysis. These results include regression on a group of diffeomorphisms for brain atrophy models [20], semi-parametric regression [23], computing equivalence classes of curves [16], comparing images [27], and models for spherical regression [22]. Just within the last few years, this effort has provided new solutions for various vision problems: online covariance matrix updates for video tracking [2], spatial clustering for high angular resolution diffusion data [11], Principal Geodesic Analysis [9], dimensionality reduction [10], Nadaraya-Watson kernel regression [4] and dictionary learning for diffusion magnetic resonance images [24].

While these advances are notable, efficient methods that extend standard linear regression to the Riemannian manifold setting have appeared in the vision literature only recently. Last year, Fletcher [8] proposed an elegant parametric formulation called *geodesic regression* that regresses a manifold valued variable, $y \in \mathcal{Y}$ against a real valued variable, $x \in \mathcal{X}$. Soon thereafter, [6] adapted the ideas for regressing a dependent variable y which lies on a unit Hilbert sphere against reals, $x \in \mathcal{X}$. These solutions provide a rigorous framework for approaching important problems including those studied in [4]. For instance, [4] calculated anatomical shape change as a function of a random variable x in a population, i.e., y lies on a shape manifold and x corresponds to age (or disease). With [8], one may derive linear relationships for rotations, shapes, and other manifolds — a substantial step forward. Unfortunately, the models in their current form do not directly address various other interesting questions that come up routinely in practice. Consider the following motivating example: the shape of an anatomical region may not only be affected by disease but *also* by age and gender. These may be "nuisance" parameters [26]. In such cases, the desired statistical model must explicitly *control* for such confounding variables, otherwise will lead to false positives *and* false negatives (Type 1, Type 2 errors). In other words, the relationship between shape (or some manifold valued y) and disease, must be *conditioned* on age. This necessitates extending general linear models (GLM) to the manifold setting, which is precisely the goal of this paper. Our **contributions** are: **(i)** a novel framework for MGLM for manifold valued y (for certain manifolds) and *multiple* independent random variables, x; **(ii)** a gradient descent scheme for multiple linear regression on manifolds such as $P(n)$ and **(iii)** an approximate Log-Euclidean scheme with accompanying analysis. Jointly, these ideas expand the operating range of existing methods enabling various hypotheses testing studies with multiple confounding variables. We show applications to statistical analysis of diffusion weighted magnetic resonance images, which give rise to the manifold of positive definite matrices, $P(n)$, and the unit Hilbert sphere, S^∞, depending on the acquisition type. Our experiments are accompanied by an open source library, where few alternatives exist today.

2. Preliminaries

We briefly summarize certain basic concepts and notations [5] that the remainder of the presentation will utilize.

Riemannian manifolds. A *differentiable manifold* [5] 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)** $\cup_i(U_i) = \mathcal{M}$; **(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. In other words, a differentiable manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure. The tangent space at a point p on the manifold, $T_p\mathcal{M}$, is a vector space that consists of the tangent vectors of *all* possible curves passing through p. The Tangent bundle of $\mathcal{M}$, i.e., $T\mathcal{M}$, is the disjoint union of tangent spaces at all points of $\mathcal{M}$, $T\mathcal{M} = \coprod_{p \in \mathcal{M}} T_p\mathcal{M}$. The tangent bundle is equipped with a natural *projection map* $\pi : T\mathcal{M} \to \mathcal{M}$ [19].

A Riemannian manifold is equipped with a smoothly varying inner product. The family of inner products on all tangent spaces is known as the *Riemannian metric* of the manifold. We can define various geometric measures on the manifold such as the angle between two curves or the length of a curve. The *geodesic distance* between two points on $\mathcal{M}$ is the length of the shortest *geodesic* curve connecting the two points, analogous to straight lines in $\mathbf{R}^n$ [13]. The geodesic curve from y_i to y_j can be parameterized by a tangent vector in the tangent space at y_i with an exponential map $\mathrm{Exp}(y_i, \cdot) : T_{y_i}\mathcal{M} \to \mathcal{M}$. The inverse of the exponential map is the logarithm 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. Separate

from the above notation, matrix exponential (and logarithm) are simply given as $\exp(\cdot)$ (and $\log(\cdot)$).

Intrinsic mean. Let $d(\cdot,\cdot)$ define the distance between two points. The intrinsic (or Karcher) mean is the minimizer to

$$\bar{y} = \arg\min_{y\in\mathcal{M}} \sum_{i=1}^{N} w_i d(y, y_i)^2, \tag{3}$$

which may be an arithmetic, geometric or harmonic mean depending on $d(\cdot,\cdot)$. On manifolds, the Karcher mean with distance $d(y_i, y_j) = \|\text{Log}_{y_i} y_j\|$ satisfies $\sum_{i=1}^{N} \text{Log}_{\bar{y}} y_i = 0$. This identity means that $\bar{y}$ is a local minimum which has a zero norm gradient [15], i.e., the sum of all tangent vectors corresponding to geodesic curves from mean $\bar{y}$ to all points y_i is zero in the tangent space $T_{\bar{y}}\mathcal{M}$. On manifolds, the existence and uniqueness of the Karcher mean is not guaranteed unless we assume, for uniqueness, that the data is in a small neighborhood (see pseudocode in the extended version).

Parallel transport. Let $\mathcal{M}$ be a differentiable manifold with an affine connection ∇ and I be an open interval. Let $c : I \to \mathcal{M}$ be a differentiable curve in $\mathcal{M}$ and let V_0 be a tangent vector in $T_{c(t_0)}\mathcal{M}$, where $t_0 \in I$. Then, there exists a unique parallel vector field V along c, such that $V(t_0) = V_0$. Here, $V(t)$ is called the *parallel transport* of $V(t_0)$ along c.

2.1. Geodesic regression: The basic formulation

The basic geodesic regression model in [8] for extending linear regression to the Riemannian manifold setting is,

$$y = \text{Exp}(\text{Exp}(p, xv), \epsilon) \tag{4}$$

where ϵ is the error (a tangent vector), $x \in \mathbf{R}$ and $y \in \mathcal{M}$ are the independent and dependent variables resp., $p \in \mathcal{M}$ is a 'base' point on the manifold, and $v \in T_p\mathcal{M}$ is a tangent vector at p. For consistency with Euclidean space, we use m for the dimensionality of $T_p\mathcal{M}$. Comparing (4) and Table 1, observe that p and v correspond to the intercept α and the slope β in (1). Given N pairs of the form (x_i, y_i), [8] solves for $(p, v) \in T\mathcal{M}$ to fit *one* geodesic curve to the data,

$$E(p, v) := \frac{1}{2}\sum_{i=1}^{N} d(\text{Exp}(p, x_i v), y_i)^2, \tag{5}$$

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.

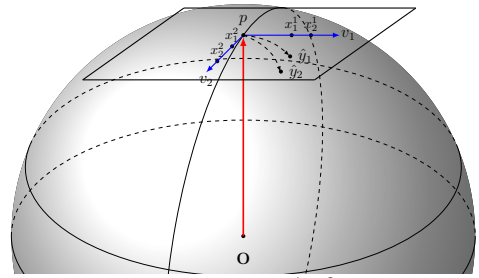

Figure 1: MGLM on manifolds. v^1, v^2 are tangent vectors. Each entry of independent variables $(x^1, x^2) \in \mathbf{R}^2$, is multiplied by v_1 and v_2 respectively in $T_p\mathcal{M}$. Here, x_i^j denotes j-th entry of the i-th instance.

providing the estimate $\hat{y}_i = \text{Exp}(p, x_i v)$. Here, $d(a, b) = \sqrt{\langle \text{Log}(a, b), \text{Log}(a, b)\rangle_a}$ is the geodesic distance between two points on $\mathcal{M}$. Rewriting (5) in the form of a minimization problem using the definition of $d(\cdot,\cdot)$, we obtain

$$\min_{(p,v)\in T\mathcal{M}} \quad \frac{1}{2}\sum_i \langle \text{Log}(\hat{y}_i, y_i), \text{Log}(\hat{y}_i, y_i)\rangle_{\hat{y}_i} \tag{6}$$

To minimize E, one first needs to specify $\nabla_p E(p, v)$ and $\nabla_v E(p, v)$. This requires computing derivatives of the exponential map with respect to p and v. The gradients are derived in terms of *Jacobi fields* (which are solutions to a second order equation subject to certain initial conditions [8]) or via introducing small perturbations [6]. To express this in a computable form, we need to find the so-called *adjoint derivative*. If $d_p\text{Exp}(p, v)$ is the derivative of the exponential map with respect to p, its *adjoint derivative* is $\langle d_p\text{Exp}(p, v)u, w\rangle = \langle u, d_p\text{Exp}(p, v)^\dagger w\rangle$, where w is a tangent vector. Putting these pieces together, the gradient descent scheme [8] can optimize (6) in a numerically stable manner and obtain the estimates of p and v.

Can we extend this idea to multiple linear regression? Given the precise form of the scheme above, it is natural to ask if a similar idea will work for the MGLM model. It turns out that there are certain conceptual and technical difficulties in attempting such an extension. Observe that geodesic regression in (5) works on a scalar independent variable in $\mathbf{R}$ (and thereby, a single geodesic). For multiple linear regression, one must invariably make use of a subspace instead. Shortly, we will see that this issue can be resolved by identifying multiple geodesic curves which correspond to 'basis' vectors in Euclidean space, see Fig. 1. It is similar to the concept of geodesic submanifold in [9]. Let us now look at some of the technical difficulties by writing down the form of the gradients in geodesic regression.

$$\underbrace{-\sum_{i=1}^{N} d_p\text{Exp}(p, x_i v)^\dagger \varpi_{y_i}}_{\nabla_p E(p,v)}, \text{ and } \underbrace{-\sum_{i=1}^{N} x_i d_v\text{Exp}(p, x_i v)^\dagger \varpi_{y_i}}_{\nabla_v E(p,v)}$$

where $\varpi_{y_i} = \text{Log}(\text{Exp}(p, x_i v), y_i)$ and $d_p\text{Exp}(p, x_i v)^\dagger$ is the adjoint derivative [8]. The derivative of the exponen-

tial map, $d\text{Exp}(p, v)$, is obtained by Jacobi fields *along* a geodesic curve parameterized by a tangent vector v. Here, this idea works because the prediction is a single geodesic curve. In multiple linear regression, predictions do not correspond to one geodesic curve; so, expressing the corresponding tangent vector (essential for Jacobi fields), is problematic. Next, a key property of the adjoint derivative above is that the result of applying the operator, $d(p, v)^\dagger$, on w should lie in a tangent space. But for manifolds like $GL(n)/O(n)$, the tangent space corresponds to symmetric matrices [2]. This requires designing a special adjoint operator which guarantees this property, which is not trivial.

3. General linear model (GLM) on manifolds

Let x and y be vectors in $\mathbf{R}^n$ and $\mathbf{R}^m$ respectively. The multivariate multilinear model in Euclidean space is.

$$y = \alpha + \beta^1 x^1 + \beta^2 x^2 + \ldots + \beta^n x^n + \epsilon \tag{7}$$

where α, β^i and the error ϵ are in $\mathbf{R}^m$ and $x = [x^1 \ldots x^n]^T$ are the independent variables. Just as (7) uses a vector of independent variables, multivariate multilinear geodesic regression uses a geodesic basis which corresponds to multiple tangent vectors, one for each independent random variable. Let $Vx_i := \sum_{j=1}^n v^j x_i^j$. Our formulation with multiple geodesic bases is

$$\min_{p \in \mathcal{M}, \forall j, v^j \in T_p\mathcal{M}} \quad \frac{1}{2} \sum_{i=1}^N d(\text{Exp}(p, Vx_i), y_i)^2, \tag{8}$$

3.1. Variational method for gradient descent

High level summary and idea. To address the technical issues pertaining to the adjoint derivatives, we will attempt to obtain a similar effect to that operator, but via different means. First, observe that in [8], to enable summing up the individual $d\text{Exp}(p, x_i v)^\dagger \text{Log}(\text{Exp}(p, x_i v), y_i)$'s which gives the gradient, ∇E, a necessary condition is that these terms should lie in $T_p\mathcal{M}$. Here, $\text{Exp}(p, x_i v)$ gives the predicted $\hat{y}_i$ for y_i, and so $\text{Log}(\text{Exp}(p, x_i v), y_i)$ is the error and must lie in $T_{\text{Exp}(p, x_i v)}\mathcal{M}$, i.e., $T_{\hat{y}_i}\mathcal{M}$. By this argument, $d\text{Exp}(p, x_i v)^\dagger$ actually plays a role of parallel transport to bring each error $\text{Log}(\text{Exp}(p, x_i v), y_i)$ from $T_{\hat{y}_i}\mathcal{M}$ to $T_p\mathcal{M}$. Since we hope to avoid constructing a special adjoint operator, we will instead perform parallel transport explicitly and derive the appropriate gradient terms, as outlined below.

Let us consider a slight variation of the objective function in (8). Let $\Gamma_{p \to q}(w)$ be a parallel transport of w from $T_p\mathcal{M}$ to $T_q\mathcal{M}$. Recall that parallel transport does *not* change the norm of tangent vectors, so the measurement of an error vector remains accurate. This ensures that the following modification preserves equivalence

$$E(p, v) = \frac{1}{2} \sum_i \langle \text{Log}(\hat{y}_i, y_i), \text{Log}(\hat{y}_i, y_i) \rangle_{\hat{y}_i} \tag{9}$$

$$= \frac{1}{2} \sum_i \langle \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i), \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i) \rangle_p, \tag{10}$$

where $\hat{y}_i = \text{Exp}(p, \sum_j x_i^j v^j)$. Comparing (9) and (10), we see that in (9), the inner product occurs at each tangent space $T_{\hat{y}_i}\mathcal{M}$, whereas (10) expresses all inner products in a tangent space $T_p\mathcal{M}$, *after* applying a parallel transport. Of course, when we have a single independent variable — we can either use the adjoint derivative operator as in [8] or the parallel transport strategy; we have verified that empirically, the results match. But the latter option is convenient for handling tangent spaces for multiple independent variables.

For an object u on a manifold, let $u^\wr$ denote the corresponding object in tangent space of u at $T_p\mathcal{M}$. To derive our gradient expression, we will use the model in the tangent space as a temporary placeholder, to keep notations simple. Let us first define a few useful terms. Below, E is the error from (5) and $E^\wr$ gives the Log-Euclidean error in $T_p\mathcal{M}$. Let $p^\wr := \text{Log}(p, p)$ and $y_i^\wr := \text{Log}(p, y_i)$. We are searching for not only tangent vectors but also the tangent space itself. To do so, by introducing a zero-norm tangent vector $p^\wr$ which corresponds to the origin of tangent space $T_p\mathcal{M}$, the direction to move the origin is obtained. So, the estimate, $\hat{y}_i{}^\wr$ is $\text{Log}(p, \hat{y}_i) = \text{Log}(p, \text{Exp}(p, Vx_i)) = Vx_i + p^\wr$, where $V = [v_1 \ldots v_n]$ is a m-by-n matrix, v^j is the m-dimensional tangent vector. The model in tangent space $T_p\mathcal{M}$ is given as

$$\min_{p^\wr, v} E^\wr(p^\wr, v) := \min_{p^\wr, v} \frac{1}{2} \sum_i \|(\sum_j v^j x_i^j + p^\wr) - y_i{}^\wr\|^2 \tag{11}$$

Its gradient is expressed as

$$\nabla_{p^\wr} E^\wr = \sum_i (\hat{y}_i{}^\wr - y_i^\wr), \quad \nabla_{v^j} E^\wr = \sum_i x_i^j (\hat{y}_i{}^\wr - y_i^\wr), \tag{12}$$

where $\hat{y}_i{}^\wr = \sum_j v^j x_i^j + p^\wr$. Note that this gradient is the 'approximate' gradient in the linearized (tangent) space. Of course, we are actually interested in minimizing the parallel transported error on the manifold. So, we will substitute the parallel transported form for the linearized expression, $(\hat{y}_i{}^\wr - y_i^\wr)$ in (12) above and obtain,

$$\nabla_p E \approx -\sum_i \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i), \quad \nabla_{v^j} E \approx -\sum_i x_i^j \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i)$$

where $\hat{y}_i = \text{Exp}(p, Vx_i)$. Separate from the intuitive reasoning above, one may take derivatives of (10) directly and verify that if a constant (and a constant times x_i^j) suffices for a first order approximation of $\frac{\partial}{\partial p} \Gamma_{\hat{y}_i \to p} \text{Log}(\text{Exp}(p, Vx_i), y_i)$ and $\frac{\partial}{\partial v^j} \Gamma_{\hat{y}_i \to p} \text{Log}(\text{Exp}(p, Vx_i), y_i)$ respectively, the form above approximates the gradient well. Otherwise, in a pathological case, we still make progress in a descent direction but the convergence rate will be worse (assuming the approximant is chosen to be small enough).

Remarks. Consider the Euclidean setting where x_i and x_j are large. The optimal intercept, p^*, will be far from y_i and y_j which is not a problem since we can explicitly solve for any value for the intercept. But parametric models for Riemannian manifolds are based on the assumption that data are distributed in a sufficiently small neighborhood where exponential and logarithm maps are well-defined. In addition, x should *not* have "large" entries (relative to the variance of x) otherwise p^* might be too far from the data and there is no well-defined exponential map to represent $y = \text{Exp}(p^*, Vx)$. Thus, we may explicitly solve for a parameter to translate the x variables, $y = \text{Exp}(p, V(x-b))$ where $b \in \mathbf{R}^n$. However, it may lead to many local minima. A simple fix to this problem is to first "center" the x variables which makes the optimization scheme stable (see pseudocode in Algo. 1).

4. Log-Euclidean geodesic regression

In this section, we outline an approximate algorithm that is simpler and offers more flexibility in analysis at the cost of a few empirically derived assumptions. To motivate the formulation, let us take a manifold perspective of (2): we see that analytical solutions can be obtained using the difference of each point from its mean both in $\mathcal{X}$ and $\mathcal{Y}$ space — that is, the quantities $\overrightarrow{\bar{x}x_i}$ and $\overrightarrow{\bar{y}y_i}$ calculated in the tangent space, $T_{p^*}\mathcal{M}$. Note that in (2), β corresponds to tangent vectors and α corresponds to p^*. Our scheme in (8) explicitly searches for p^*, but in experiments, we found that frequently, p^* turns out to be quite close to $\bar{y}$. This observation yields a heuristic where rather than solve for p, we operate entirely in $T_{\bar{y}}\mathcal{M}$. With this assumption, using the Karcher mean as $\bar{y}$ in (2) and the Log-Euclidean distance as a substitute for $\overrightarrow{\bar{x}x_i}$ and $\overrightarrow{\bar{y}y_i}$, we can derive a faster approximate procedure. This scheme has the additional benefit that it allows analyzing multiple manifold-valued *independent* variables in a GLM setting too, if desired.

The Log-Euclidean MGLM estimates a linear relationship between centered variables $\{x_i^\wr\}_{i=1}^N$ and $\{y_i^\wr\}_{i=1}^N$ where $x_i^\wr = x_i - \bar{x}$ and $y_i^\wr = \text{Log}(\bar{y}, y_i)$, where the number of tangent vectors we estimate is exactly equal to the number of independent variables, x. We do *not* use an affine model here because that makes drawing an analogy with our full MGLM formulation problematic. For example, an affine model in the Log-Euclidean setting will entail learning $(n+1)$ tangent vectors — for the n different x-variables plus one for the intercept. But even for $n = 1$, the estimates derived will no longer be valid geodesic curves since they do not pass through $p^\wr \in T_p\mathcal{M}$ (because of the intercept). By taking $\bar{y}$ as p preserves these geodesic properties.

Our basic procedure estimates the set of vectors $V = [v_1 \dots v_n]$ in tangent space $T_{\bar{y}}\mathcal{M}$ and p a point on $\mathcal{M}$ using the relation $Y^\wr = VX^\wr$. $Y^\wr \equiv [y_1^\wr \dots y_N^\wr]$ and $X^\wr \equiv [x_1^\wr \dots x_N^\wr]$ are respectively the mean centered data. p^* is given by the Karcher mean $\bar{y}$. V^* is given by the least squares estimation with respect to the Log-Euclidean metric and can be computed using the closed form solution, $Y^\wr X^{\wr T}\left(X^\wr X^{\wr T}\right)^{-1}$.

The following analysis shows that under some conditions, heuristically substituting $\bar{y}$ for p^* is justifiable beyond empirical arguments alone. In particular, if the y observations come from *some* geodesic curve, then all of the data can be parameterized by one tangent vector in the tangent space at $\bar{y}$. More specifically, we show that a Karcher mean exists on a geodesic curve. So, if the Karcher mean is unique, then the Karcher mean must lie on the geodesic curve. By the definition of exponential map and since the data are in sufficiently small neighborhood, it becomes possible to parameterize the observations by $\text{Exp}(\bar{y}, vx)$.

Prop. 1 shows the existence of the Karcher mean on a geodesic curve when the data lies on the unique geodesic curve, Ω, between two points.

Proposition 1. *Let $Y = \{y_1, \dots, y_N\}$ be a subset of a manifold $\mathcal{M}$. Suppose that Y is in a sufficiently small open cover $\mathcal{B}$ such that the exponential and logarithm maps are bijections. Suppose that all $y \in Y$ are on a curve Ω that is the unique geodesic curve between some y_i and y_j in Y. Then there exists $\bar{y}$ in Ω such that $\sum_{y \in Y} Log_{\bar{y}} y = 0$ (the first order condition for Karcher mean).*

Proof. Let $v \in T_{y_i}\mathcal{M}$ be the tangent vector $v = \text{Log}_{y_i} y_j$. Since all points of Y are a subset of a geodesic curve Ω between y_i, and y_j, for each $y_k \in Y$, there exists an $x_k \in [0, 1]$ such that $y_k = \text{Exp}(y_i, vx_k)$. So, let $\bar{x} = \frac{1}{N}\sum_{k=1}^N x_k$ and $\bar{y} = \text{Exp}(y_i, v\bar{x})$. Then, $\bar{y}$ satisfies $\sum_k \text{Log}_{\bar{y}} y_k = 0$ and it is in Ω since the arithmetic mean $\bar{x}$ is in $[0, 1]$. □

With this result in hand, we next show that the data is parameterizable in terms of V.

Proposition 2. *If $\bar{y}$ is the unique Karcher mean of $Y \subset \Omega$, and it is obtained in $\mathcal{B}$, then $\bar{y} \in \Omega$. Further, for some*

Algorithm 1 MGLM on manifolds

Input: $x_1, \dots, x_N \in \mathbf{R}^n$, $y_1, \dots, y_N \in \mathcal{M}$
Output: $p \in \mathcal{M}$, $v^1, \dots, v^n \in T_p\mathcal{M}$,

Initialize $p, v, \alpha, \alpha_{max}$ and center x
while termination condition **do**
  $p_{new} = \text{Exp}(p, -\alpha\nabla_p E)$
  $V_{new} = \Gamma_{p \to p_{new}}(V - \alpha\nabla_V E)$
  if $E(p_{new}, V_{new}) < E(p, V)$ **then**
    $V \leftarrow V_{new}$ and $P \leftarrow P_{new}$
    $\alpha = \min(2\alpha, \alpha_{max})$
  else
    $\alpha = \alpha/2$
  end if
end while

$v \in T_{\bar{y}}\mathcal{M}$ and each y, there exists $x \in \boldsymbol{R}$ such that $y = Exp(\bar{y}, vx)$.

Proof. If $\bar{y}$ is a unique Karcher mean of Y on $\mathcal{M}$ and it is obtained in a sufficiently small neighborhood $\mathcal{B} \subset \mathcal{M}$ of data, then $\sum_{k=1}^{N} \text{Log}_{\bar{y}} y_k = 0$ holds by the first order optimality condition of (3). Uniqueness of $\bar{y}$ and Prop. 1 implies that $\bar{y}$ is in Ω. In a small neighborhood, by definition of exponential map, there must exist an appropriate x. □

5. Experiments

Here, we show the application of our models for statistical analysis of diffusion weighted imaging (DWI) data. On synthetic simulations and real data from two distinct neuroimaging studies, the experiments evaluate whether (and to what extent) general linear model (GLM) analysis on diffusion weighted images in neuroimaging can benefit from **(a)** the ability to deal with manifold valued data **(b)** allowing multiple explanatory (including nuisance) variables.

5.1. Diffusion weighted imaging

DWI is a magnetic resonance (MR) imaging modality that non-invasively maps and characterizes the microstructural properties and macroscopic organization of the white matter regions in the brain [14]. This is achieved by sensitizing the MR signal to the diffusion of the water molecules (protons). Specifically, the diffusion of the protons causes signal attenuation proportional to the apparent diffusion coefficient (D_0) and the MR acquisition parameter known as the b-value. The b-value represents properties like the magnitude, duration and shape of the magnetic field gradients.

Diffusion tensor imaging (DTI). By measuring the attenuated signal in at least six different directions, one can estimate the diffusion pattern of the protons in the three orthogonal Cartesian directions using a symmetric positive definite covariance matrix, the diffusion tensor. From the diffusion tensor, we can obtain maps of the most commonly used scalar measure of diffusion anisotropy at each voxel called fractional anisotropy (FA) which is the normalized standard deviation of the eigen values of the tensor in $[0, 1]$.

Higher angular resolution diffusion imaging (HARDI). When DWI data is acquired in significantly more (≥ 30) number of diffusion directions and with large enough b-values, the data can be modeled by a spherical Fourier transform which yields the so-called orientation distribution functions (ODFs) [25]. ODFs provide much higher resolution of diffusion directions compared to DTI. We can extract scalar measures at each voxel such as generalized fractional anisotropy (GFA) [25] from the ODFs which can then be used in analysis.

5.2. Statistical analysis on DWI

Most neuroimaging studies acquire DWI data to perform follow-up statistical analysis. Assume that the images are are already registered to a common template. The scientific question may be to identify which regions of the brain *vary* across two groups of subjects: diseased and healthy. This can be answered by performing a hypothesis test at each voxel over the entire brain, and reporting the statistically significant ones as different across groups. Separately, one may want to identify regions which have a strong relationship with disease status. Independent of the specific setting, the classical analysis makes use a scalar-valued summary measure at each voxel: FA for DTI or GFA for HARDI. But this simplification, which makes the differential signal harder to detect, can be avoided. In DTI, the diffusion tensors lie on a $P(3)$ manifold (i.e., the quotient space $GL(3)/O(3)$). In HARDI, the square root parameterization of the ODF lies on a unit Hilbert sphere (S^{∞}) [3], which in practice, is expressed as a l-th order spherical harmonics (we use $l = 4$ implying the S^{14} setting). With the appropriate statistical models in hand, we may regress the manifold data directly against one or more independent variables.

5.3. Synthetic setting

We first artificially generate ODF and DTI data via a generative multiple linear model. We then estimate using our MGLM framework and the model in [6] (certain adjustments to [6] were needed for the $P(3)$ manifold). The results in Fig. 2 give strong evidence that when the characteristics of the data depends on multiple independent variables (e.g., disease *and* age), MGLM significantly outperforms linear geodesic regression (SLGR) which regresses $y \in \mathcal{Y}$ against $x \in \mathbf{R}$. In Fig. 2, GR1 and GR2 refer to the estimates from SLGR using variables x^1 and x^2 individually. MGLM is able to estimate the true signal far more accurately compared to either GR1 and GR2. Fig. 3 shows the quantitative results (error) of regression using four independent variables as a function of sample sizes. As expected, we see that the fit improves significantly with MGLM.

5.4. Neuroimaging data evaluations

We now present experiments using DWI data from two real neuroimaging studies. The first study investigates the neuroplasticity effects of meditation practice (e.g., for emotional well-being) on white matter. Meditators were trained in Buddhist meditation techniques, which lead to emotion regulation and attention control. An example scientific question here may be: what is the relationship between the number of years of meditation training on white matter when conditioned on age? Here, diffusion images in 48 non-collinear diffusion encoding directions were acquired, which after a sequence of pre-processing steps, provide the ODF representations for 23 long-term meditators (LTM)

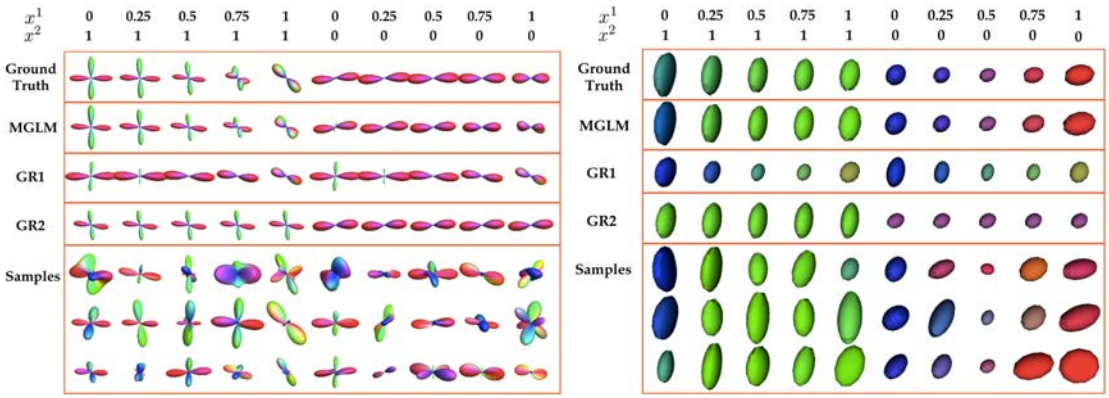

Figure 2: MGLM and SLGR results using synthesized ODF and DTI. First two rows give the values for x^1 and x^2 of the generative model.

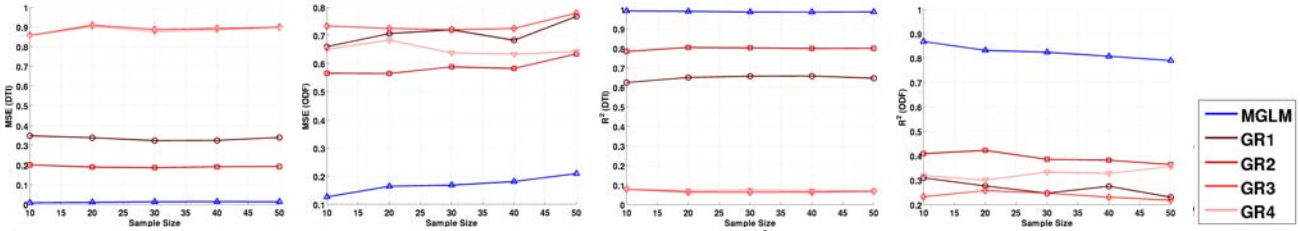

Figure 3: Plots showing the effect of sample size on mean squared error (MSE) and R^2 for the MGLM as well as SLGR using the individual variables. GR$\{1, \cdots, 4\}$ refer to the estimates from SLGR using the individual variables $x^{\{1,\cdots,4\}}$ individually.

and 26 wait-list control (WLC) subjects. In the second study, we investigate the effect of a genotype risk factor (i.e., APOE4 positive or negative) in Alzheimer's disease (AD) on white matter integrity in the brain. A scientific question here may be: what is the effect of age on white matter when we control for genotype and gender? Here, 40 encoding directions were acquired and images were obtained after pre-processing. The dataset covers 343 subjects (123 with APOE4+ and 220 with APOE4-).

GLM results. We estimate the following model at each voxel for both studies,

$$y = \text{Exp}(p, v^1\text{Group} + v^2\text{Gender} + v^3\text{Age}), \qquad (13)$$

where $y \in P(3)$, Group $\in$ {APOE4+, APOE4-} for AD-risk study and $y \in S^{14}$, Group $\in$ {LTM, WLC} for meditation study. Due to limited space, we only summarize results of ODF regression on the meditation data here (see results from AD-risk study in the extended version of the paper).

As a baseline, we present regression results using FA as the measure of interest. We note that regressing y against *one* independent variable as in Fig. 3 is some possible baseline but because it is restricted, it cannot fit the full model in (13). Therefore, FA is a better baseline for comparisons. The null hypothesis, H_0, here is that the linear combination of 'group', 'gender' and 'age' has no effect on the y measurement. Therefore, (13) serves as the "full" model and the intercept alone serves as the nested model. Then, an F-statistic can yield voxel-wise p-value maps when we regress on FA. However, F-statistics are not applicable to ODF regression. So, to obtain p-values, we use $20,000$ permutations to characterize the Null distribution of the R^2-fit. Then, the unpermuted R^2 is used to calculate the p-values. Comparing the two p-value maps (FA vs. ODF) shows which procedure is successfully picking up more differential signal in a statistically sound manner.

Fig. 4 shows the p-value maps, for FA and ODF based regression. We can observe the improved statistical sensitivity using the MGLM framework. Although neuroscientific interpretation of such results is outside the scope of this paper, the results from MGLM are more plausible since the statistically significant voxels follow white matter curvature more closely. This is expected because ODFs provide much higher resolution orientation information compared to DTI. We must point out that for this dataset, due to small sample size, MGLM or GLM do not pass the false discovery rate (FDR) threshold at $q = 0.05$. Nonetheless, the histograms of p values in Fig. 5 show that our MGLM is superior to univariate analysis on FA, even when the FA images are smoothed prior to testing (to increase normality). In summary, the experiments provide clear evidence that a full MGLM procedure on the actual manifold data ($P(3)$ or S^{14}), offers significant and consistent improvements in sensitivity for statistical analysis in neuroimaging.

6. Conclusions

This paper extends multivariate general linear model (MGLM) to the manifold setting. Such an extension allows regressing a manifold valued dependent variable $y \in \mathcal{M}$ against multiple independent variables, $x \in \mathcal{X}$, a property which extends the applicability of existing meth-

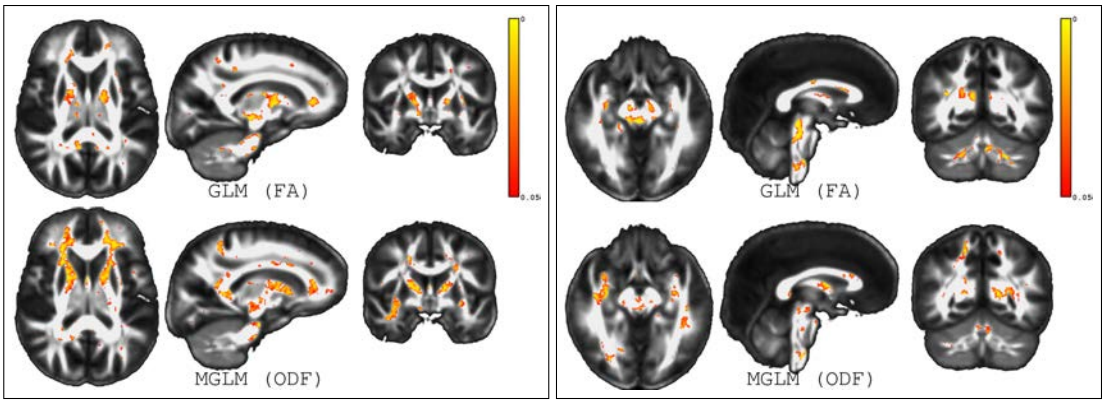

Figure 4: (Left) p-value maps obtained using GLM (top) and MGLM (bottom) frameworks. GLM was performed using smoothed FA images while MGLM used ODF images. Voxels that satisfy $p \leq 0.05$ are spatially more contiguous when performing MGLM. (Right) The thresholded p-value maps showing that the spatial extents in the brain stem (axial-left, sagittal-middle) and cerebellar (coronal-right) white matter are reduced when using MGLM.

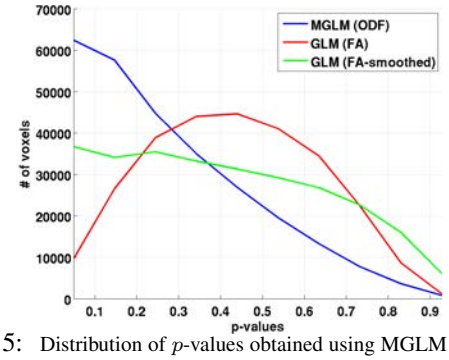

Figure 5: Distribution of p-values obtained using MGLM using ODFs and GLM using both smoothed and unsmoothed FA images.

ods, and will allow practitioners to easily regress voxel measurements in diffusion weighted imaging against clinical variables, while controlling for nuisance parameters, thereby obtaining results which better reflect hypotheses under study. The experiments give strong evidence of the improvements we may expect over traditional alternatives. The paper is accompanied by an open source codebase, which will enable easy deployment in practice.

Acknowledgments: This work was supported in part by NIH R01 AG040396; R01 AG037639; R01 AG027161; R01 AG021155; NSF CAREER award 1252725; NSF RI 1116584; Wisconsin Partnership Fund; UW ADRC P50 AG033514; UW ICTR 1UL1RR025011; a VA Merit Review Grant I01CX000165; NCCAM P01 AT004952-04 and the Waisman Core grant P30 HD003352-45. Collins was supported by a CIBM fellowship (NLM 5T15LM007359). The contents do not represent views of the Dept. of Veterans Affairs or the United States Government. We are grateful to Anqi Qiu (National University of Singapore) for explaining various mathematical details of ODF construction with code examples, which were useful in setting up the experimental analysis.

References

[1] S.-I. Amari. Information geometry of the EM and EM algorithms for neural networks. *Neural networks*, 8(9):1379–1408, 1995.

[2] G. Cheng and B. C. Vemuri. A novel dynamic system in the space of SPD matrices with applications to appearance tracking. *SIAM J. on Imag. Sci.*, 6(1):592–615, 2013.

[3] J. Cheng, A. Ghosh, T. Jiang, et al. A Riemannian framework for orientation distribution function computing. In *MICCAI*. 2009.

[4] B. C. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi. Population shape regression from random design data. *IJCV*, 90(2):255–266, 2010.

[5] M. P. Do Carmo. *Riemannian geometry*. Springer, 1992.

[6] J. Du, A. Goh, S. Kushnarev, and A. Qiu. Geodesic regression on ODFs with its application to an aging study. *NeuroImage*, 2013.

[7] B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. of Stat.*, 1975.

[8] P. T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *IJCV*, pages 1–15, 2012.

[9] P. T. Fletcher, C. Lu, et al. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005, 2004.

[10] A. Goh. *Estimation & processing of ODFs for HARDI*. PhD thesis, 2010.

[11] A. Goh and R. Vidal. Unsupervised Riemannian clustering of PDFs. In *MLKDD*, pages 377–392. 2008.

[12] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *NIPS*, pages 487–493, 1999.

[13] S. Jayasumana et al. Kernel methods on the Riemannian manifold of SPD matrices. In *CVPR*, pages 1–8, 2013.

[14] D. Jones, A. Simmons, et al. Non -invasive assessment of axonal fiber connectivity in the human brain via DT-MRI. *MRM*, 1999.

[15] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. pure and applied math.*, 30(5):509–541, 1977.

[16] S. Kurtek, E. Klassen, Z. Ding, et al. A novel Riemannian framework for shape analysis of 3D objects. In *CVPR*, 2010.

[17] J. D. Lafferty. The density manifold and configuration space quantization. *Trans. American Math. Soc.*, 305(2):699–741, 1988.

[18] G. Lebanon. *Riemannian geometry and statistical machine learning*. PhD thesis, 2005.

[19] J. M. Lee. *Introduction to smooth manifolds*. Springer, 2012.

[20] M. I. Miller. Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms. *NeuroImage*, 23:S19–S33, 2004.

[21] C. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37(3):81–91, 1945.

[22] M. Rosenthal, W. Wu, E. Klassen, and A. Srivastava. Spherical regression models using projective linear transformations. *Journal of American Statistical Association*, 2013.

[23] X. Shi, M. Styner, et al. Intrinsic regression models for manifold-valued data. pages 192–199, 2009.

[24] J. Sun, Y. Xie, et al. Dictionary learning on the manifold of square root densities and application to reconstruction of diffusion propagator fields. In *IPMI*, pages 619–631, 2013.

[25] D. Tuch. Q-ball imaging. *MRM*, 52(6):1358–1372, 2004.

[26] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

[27] Z. Zhang, E. Klassen, A. Srivastava, et al. Blurring-invariant Riemannian metrics for comparing signals and images. In *ICCV*, 2011.

Multivariate General Linear Models (MGLM) on Riemannian Manifolds with Applications to Statistical Analysis of Diffusion Weighted Images (supplement)

Hyunwoo J. Kim, Nagesh Adluru, Maxwell D. Collins, Moo K. Chung,
Barbara B. Bendlin, Sterling C. Johnson, Richard J. Davidson, Vikas Singh
University of Wisconsin-Madison
http://pages.cs.wisc.edu/~hwkim/projects/riem-mglm

1. Summary

We provide additional technical details for implementation and experimental results of our framework which were not included in the main paper due to limited space.

2. Additional technical details

In this section, we review the essential objects and operators needed for implementing MGLM on the two manifolds described in the main paper namely (1) the Hilbert unit sphere for orientation distribution functions (ODFs), and (2) the quotient space $GL(3)/O(3)$ for diffusion tensor images (DTI).

2.1. Karcher mean

Computing Karcher (or intrinsic) mean of data on manifolds forms one of the basic implementation steps in our framework and is given by the solution to the following (Eq. (3) in the main text),

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

where the w_i is weight of data $y_i \in \mathcal{M}$. Karcher mean serves as p^* in the Log-Euclidean framework for MGLM (section 4 in the main text). Karcher mean is also used in the R^2 statistic and measures the goodness of fit of linear models (Fig. 3 in the main text). The R^2 statistic is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \text{dist}(\hat{y}_i, y_i)^2}{\sum_{i=1}^{N} \text{dist}(\bar{y}, y_i)^2} \tag{1}$$

Karcher mean is obtained by Algorithm 1, where α denotes the step size ($\alpha = 1$ was used).

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 1: Karcher mean on manifolds

2.2. Manifolds for diffusion weighted imaging

For diffusion weighted imaging (DWI), we discuss models for DTI and ODFs. DTI data leads to the quotient space, $GL(n)/O(n)$, whereas square root parameterized ODFs lead to a regression problem on the Hilbert unit sphere. Technically, coefficients of spherical harmonics with a finite number of bases form a unit sphere. We provide all relevant tools for MGLM on those manifolds.

An inner product is the first step in defining Riemannian manifolds. To manipulate data, we also need the exponential map, logarithm map, and parallel transport. For prediction, the exponential map is used as $\bar{y} = \text{Exp}(p, Vx)$. Recall that we use the following gradients in our main paper.

$$\nabla_p E \approx -\sum_i \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i), \tag{2}$$

$$\nabla_{v^j} E \approx -\sum_i x_i^j \Gamma_{\hat{y}_i \to p} \text{Log}(\hat{y}_i, y_i), \tag{3}$$

Here, we provide details specific to each manifold.

2.3. Unit sphere

Unit sphere is a common manifold, and the first example demonstrated here. Let $S^n = \{p \in \mathbf{R}^{n+1} | p^T p = 1\}$ be the unit sphere in $\mathbf{R}^{n+1}$. Its tangent space is then,

$$T_p S^n = \{v \in \mathbf{R}^{n+1} | p^T v = 0\}$$

The inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is given by

$$\langle u, v\rangle_p = u^T v, \tag{4}$$

and the norm at p is

$$\|v\|_p = \sqrt{\langle u, v\rangle_p} \tag{5}$$

Since the inner product and the norm of tangent vectors are independent of p, we omit p in the following expressions. The main operators we need for MGLM on the unit sphere are,

$$\begin{aligned}
\mathrm{Exp}(p, v) &= p\cos(\|v\|) + \frac{v}{\|v\|}\sin(\|v\|) \\
\mathrm{Log}(p, q) &= \frac{(I - pp^T)q}{\sqrt{1-(p^Tq)^2}}\,\mathrm{acos}(p^Tq) \\
\Gamma_{p\to q}(w) &= -p\sin(\|v\|)v^Tw + \frac{v}{\|v\|}\cos(\|v\|)v^Tw \\
&\quad + (I - \frac{vv^T}{\|v\|^2})w, \text{where } v = \mathrm{Log}(p, q)
\end{aligned} \tag{6}$$

2.4. Orientation distribution functions

High Angular Resolution Diffusion Imaging (HARDI) is typically represented as a Orientation Distribution Function (ODF) at each voxel. Using the square root parameterization, ODFs form a unit Hilbert sphere [2],

$$\Psi = \{\psi : \mathbb{S}^2 \to \mathbb{R}^+ | \forall s \in \mathbb{S}^2, \psi(s) \geq 0; \int_{s\in\mathbb{S}^2} \psi^2(s)ds = 1\} \tag{7}$$

The inner product is given by

$$\langle \xi_i, \xi_j\rangle_{\psi_i} = \int_{s\in\mathbb{S}^2} \xi_i(s)\xi_j(s)ds, \tag{8}$$

and coincides with the Fisher-Rao metric [7]. All other operations needed for MGLM for this setting are the same as those shown for the unit sphere above.

2.5. Diffusion tensors

Diffusion tensors are symmetric positive definite matrices at each voxel. Let $P(n)$ be a manifold for symmetric positive definite matrices of size $n \times n$. This forms a quotient space $GL(n)/O(n)$, where $GL(n)$ denotes the general linear group (the group of $(n\times n)$ nonsingular matrices) and $O(n)$ is the orthogonal group (the group of $(n \times n)$ orthogonal matrices). The inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is given by [1]

$$\langle u, v\rangle_p = \mathrm{tr}(p^{-1/2}up^{-1}vp^{-1/2}) \tag{9}$$

This plays the role of the Fisher-Rao metric in the statistical model of multivariate distributions [6]. Here, $T_p\mathcal{M}$ is a tangent space at p (which is a vector space) is the space of symmetric matrices of dimension $(n+1)n/2$. The geodesic distance is,

$$d(p, q)^2 = \mathrm{tr}(\log^2(p^{-1/2}qp^{-1/2}))$$

The exponential map and logarithm map are,

$$\begin{aligned}
\mathrm{Exp}(p, v) &= p^{1/2}\exp(p^{-1/2}vp^{-1/2})p^{1/2} \\
\mathrm{Log}(p, q) &= p^{1/2}\log(p^{-1/2}qp^{-1/2})p^{1/2}
\end{aligned} \tag{10}$$

Let p, q be in $P(n)$ and a tangent vector $w \in T_p\mathcal{M}$, the tangent vector in $T_q\mathcal{M}$ which is the parallel transport of w along the shortest geodesic from p to q is given by [4]

$$\begin{aligned}
&\Gamma_{p\to q}(w) = p^{1/2}rp^{-1/2}wp^{-1/2}rp^{1/2} \\
&\quad \text{where } r = \exp\left(p^{-1/2}\frac{v}{2}p^{-1/2}\right) \text{ and } v = \mathrm{Log}(p, q)
\end{aligned} \tag{11}$$

3. Additional experiments

We provide additional GLM results DTI and HARDI. As mentioned in the main text, our data comes from two distinct neuroimaging studies. The first study investigates the neuroplasiticity effects of meditation practice on white matter. The second study is focused on Alzheimer's disease risk factors on white matter in the brain.

DWI acquisition and processing. The data was acquired using a diffusion-weighted, spin-echo, single-shot, echo planar imaging radio-frequency (RF) pulse sequence. For the meditation study, diffusion data in 48 non-collinear diffusion encoding directions with diffusion weighting factor of $b = 1000$ s·mm^{-2} and eight non-diffusion weighted ($b = 0$) images was acquired. For each of the 23 long-term meditators (LTM) and 26 wait-list control (WLC) subjects, ODFs were estimated and the square-root parameterization was obtained via linear spherical harmonic transform [5]. For the AD-risk study, images with 40 encoding directions at $b = 1300$ s·mm^{-2} and eight $b = 0$ images were acquired for each of the 343 subjects (123 with APOE4+ and 220 with APOE4-, where APOE denotes the Apolipoprotein E genotype). The diffusion tensors were estimated from DWI data from both the studies, using non-linear estimation using the Camino library [3]. The images for both studies were normalized spatially before statistical analysis using DTI-TK [8].

GLM results. We estimate the following models at each voxel for both studies,

$$\begin{aligned}
\mathrm{GLM}_{\mathrm{Full}}: &\quad y = \mathrm{Exp}(p, v^1\mathrm{Group} + v^2\mathrm{Gender} + v^3\mathrm{Age}), \\
\mathrm{GLM}_{\mathrm{Age}}: &\quad y = \mathrm{Exp}(p, v^2\mathrm{Gender} + v^3\mathrm{Age}), \\
\mathrm{GLM}_{\mathrm{Group}}: &\quad y = \mathrm{Exp}(p, v^1\mathrm{Group} + v^2\mathrm{Gender}),
\end{aligned}$$

where $y \in P(3)$, Group $\in$ {APOE4+, APOE4-} for AD-risk study and $y \in S^{14}$, Group $\in$ {LTM, WLC} for meditation study.

The uncorrected p-value maps and histograms for GLM_{Full} for the meditation study were presented in Figs. 4 and 5 respectively in the main text. Here we will present the following,

1. p-value maps and histograms for GLM_{Full} for the AD-risk study, presented in Fig. 2. We would like to recall that the p-values in this case are computed using the 20000 permutations to characterize the Null distribution of the R^2 values.

2. p-value maps and histograms for specific effect of Age and Group. Here, we compute p-values by simulating the Null distribution of the F ratio statistic again using 20000 permutations. The F ratio statistic is defined for a pair of nested GLMs as,

$$F = \frac{\frac{\text{RSS}_1 - \text{RSS}_2}{p_2 - p_1}}{\frac{\text{RSS}_2}{N - p_2}}, \tag{12}$$

   where $\text{RSS}_j = \sum_{i=1}^{N} \text{dist}(\hat{y}_{ij}, y_i)^2$, $\hat{y}_{ij}$s are estimated using GLM_j and p_j is the number of independent parameters in GLM_j. For obtaining the effect of Age and Group, RSS_1 is obtained using GLM_{Age} and $\text{GLM}_{\text{Group}}$, respectively. RSS_2 is obtained using GLM_{Full} in both the cases. The corresponding maps and histograms are shown in Figs. 3 and 5 respectively.

Main observations. In Fig. 2, we can observe that the thresholded regions are spatially more contiguous when using MGLM in the sagittal (middle) and coronal (right) slices compared to those obtained by running exactly the same model on the univariate (smoothed) FA images (the unsmoothed FA results are much worse). Note that a smoothing procedure on the DTI data (i.e., the tensors) further improves the results but the purpose here is to show that even the unsmoothed DTI (with MGLM) yields comparable (or better) results, as can be noticed in the histogram in the right panel in Fig. 2. We see the same behavior in Fig. 3. The left panel shows p-values obtained using the effect of age while the right panel shows the effect of group variable i.e., APOE4+ vs. APOE4-. Note that existing literature provides little guidance on algorithms for running such GLM models on the DTI data directly, as we show here. In the case of age effect, MGLM provides more spatially contiguous regions but in the case of APOE4 effect FA smoothed and MGLM are comparable, as can also be noted in Fig. 5. Figs. 4 and 6 show that MGLM using ODF provides more signal compared to GLM using (smoothed) FA both when trying to detect the effect of age as well as the group.

References

[1] G. Cheng and B. C. Vemuri. A novel dynamic system in the space of SPD matrices with applications to appearance tracking. *SIAM J. on Imag. Sci.*, 6(1):592–615, 2013.

[2] J. Cheng, A. Ghosh, T. Jiang, and R. Deriche. A Riemannian framework for orientation distribution function computing. In *MICCAI*, pages 911–918. 2009.

[3] P. Cook, Y. Bai, S. Nedjati-Gilani, K. K. Seunarine, M. G. Hall, G. J. Parker, and D. C. Alexander. Camino: Open-source diffusion-MRI reconstruction and processing. In *ISMRM*, page 2759, 2006.

[4] R. Ferreira, J. Xavier, et al. Newton method for Riemannian centroid computation in naturally reductive homogeneous spaces. In *ICASSP*, volume 3, 2006.

[5] A. Goh, C. Lenglet, P. M. Thompson, and R. Vidal. A nonparametric Riemannian framework for processing HARDI and its applications to ODF-based morphometry. *NeuroImage*, 56(3):1181–1201, 2011.

[6] D. Petz. Means of positive matrices: Geometry and a conjecture. In *Annales Mathematicae et Informaticae*, volume 32, pages 129–139, 2005.

[7] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *CVPR*, 2007.

[8] H. Zhang, P. Yushkevich, D. Alexander, and J. Gee. Deformable registration of diffusion tensor MR images with explicit orientation optimization. *Med. Img. Analysis*, 10:764–785, 2006.

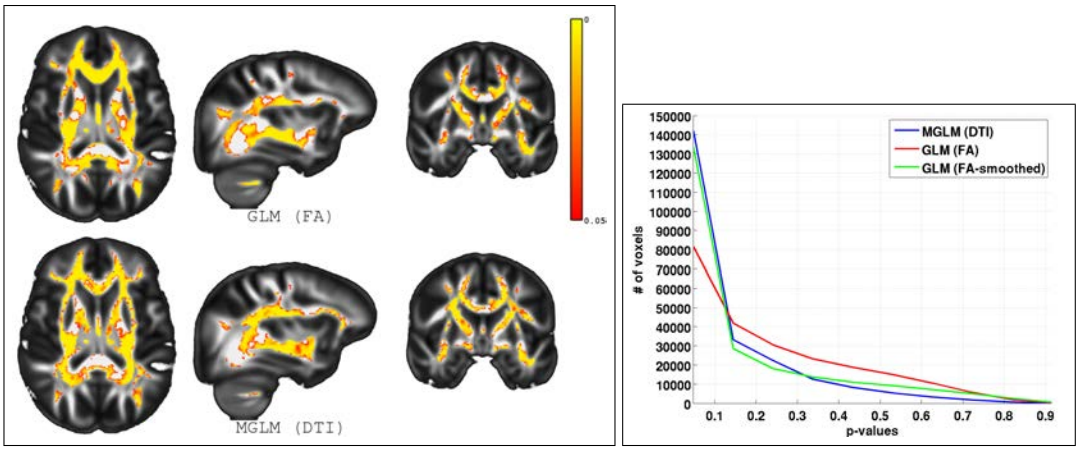

Figure 2: p-value maps obtained using FA (smoothed) based GLM as well as DTI based MGLM frameworks. Left panel shows thresholded ($p \leq 0.05$) p-values obtained using the full model. We can observe that the thresholded regions are spatially more contiguous when using MGLM. We can notice that more clearly in the sagittal (middle) and coronal (right) slices. Right panel shows distribution of p-values obtained using MGLM using DTI and GLM using both smoothed and unsmoothed FA images.

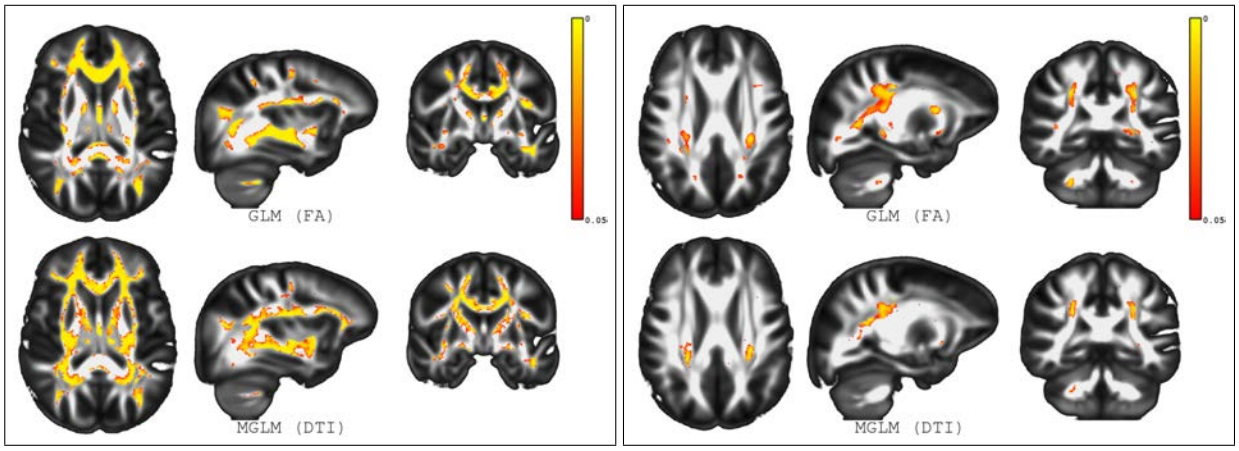

Figure 3: p-value maps obtained using FA (smoothed) based GLM as well as DTI based MGLM frameworks. Left panel shows p-values obtained using the effect of age while the right panel shows the effect of group variable i.e. APOE4+ vs. APOE4-.

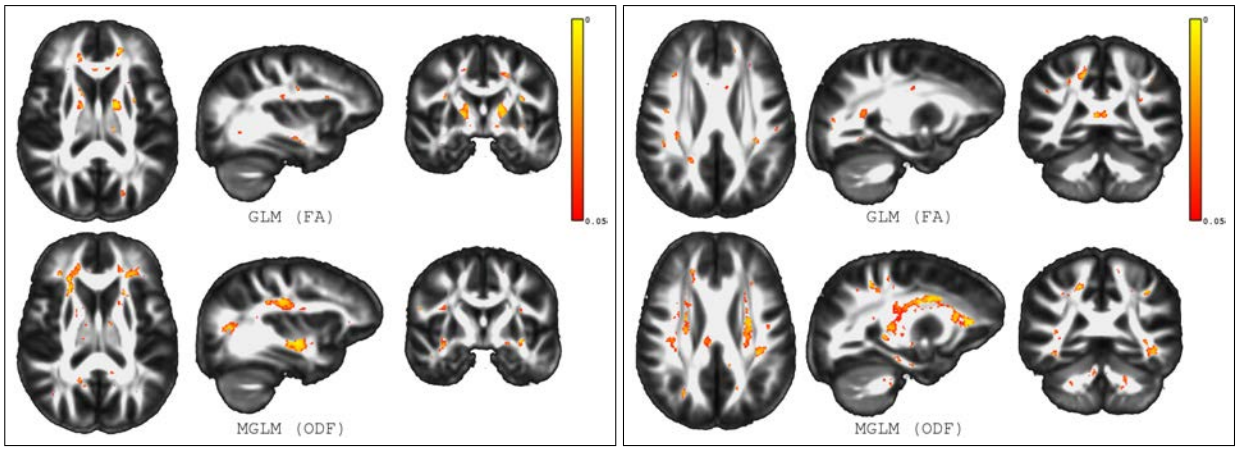

Figure 4: p-value maps obtained using FA (smoothed) based GLM as well as ODF based MGLM frameworks. Left panel shows the effect of age while the right panel shows the effect of group variable i.e. LTM vs. WLC.

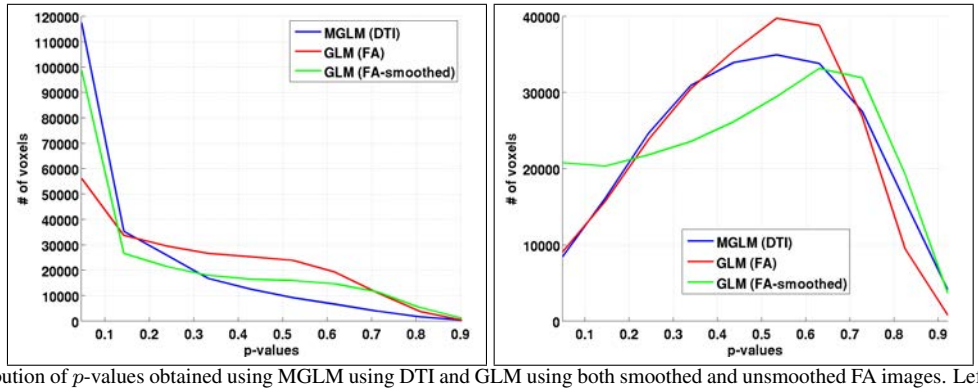

Figure 5: Distribution of p-values obtained using MGLM using DTI and GLM using both smoothed and unsmoothed FA images. Left: Age effect. Right: APOE4 effect. As discussed in Fig. 3, in case of APOE4 effect the smoothed FA based GLM and DTI based MGLM perform comparably.

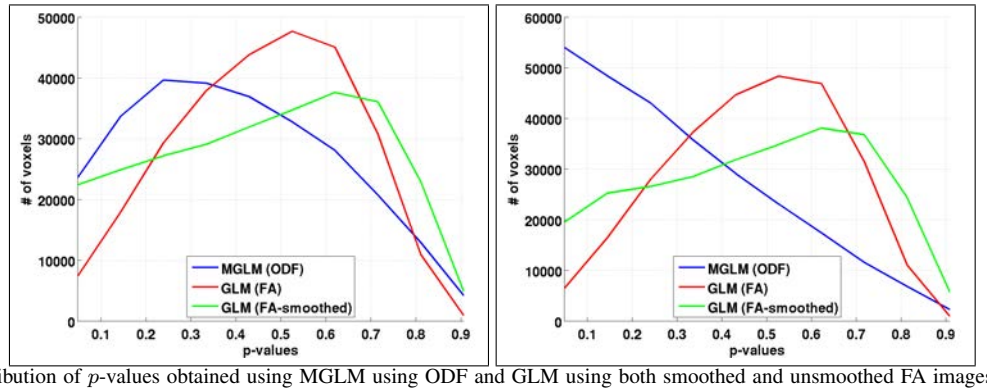

Figure 6: Distribution of p-values obtained using MGLM using ODF and GLM using both smoothed and unsmoothed FA images. Left: Age effect. Right: Group effect (LTM vs. WLC).