Canonical Correlation Analysis on Riemannian Manifolds and Its Applications

Hyunwoo J. Kim[1], Nagesh Adluru[1], Barbara B. Bendlin[1], Sterling C. Johnson[1], Baba C. Vemuri[2], and Vikas Singh[1]

[1] University of Wisconsin–Madison
[2] University of Florida
http://pages.cs.wisc.edu/~hwkim/projects/riem-cca

Abstract. Canonical correlation analysis (CCA) is a widely used statistical technique to capture correlations between two sets of multi-variate random variables and has found a multitude of applications in computer vision, medical imaging and machine learning. The classical formulation assumes that the data live in a pair of *vector spaces* which makes its use in certain important scientific domains problematic. For instance, the set of symmetric positive definite matrices (SPD), rotations and probability distributions, all belong to certain curved Riemannian manifolds where vector-space operations are in general not applicable. Analyzing the space of such data via the classical versions of inference models is rather sub-optimal. But perhaps more importantly, since the algorithms do not respect the underlying geometry of the data space, it is hard to provide statistical guarantees (if any) on the results. Using the space of SPD matrices as a concrete example, this paper gives a principled generalization of the well known CCA to the Riemannian setting. Our CCA algorithm operates on the product Riemannian manifold representing SPD matrix-valued fields to identify meaningful statistical relationships on the product Riemannian manifold. As a proof of principle, we present results on an Alzheimer's disease (AD) study where the analysis task involves identifying correlations across diffusion tensor images (DTI) and Cauchy deformation tensor fields derived from T1-weighted magnetic resonance (MR) images.

1 Introduction

Canonical correlation analysis (CCA) is a powerful statistical technique to extract linear components that capture correlations between two multi-variate random variables [15]. CCA provides an answer to the following question: suppose we are given data of the form, $(\boldsymbol{x}_i \in \mathcal{X}, \boldsymbol{y}_i \in \mathcal{Y})_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{x}_i \in \mathbf{R}^m$ and $\boldsymbol{y}_i \in \mathbf{R}^n$, find a model that explains *both* of these observations. More precisely, CCA provides an answer to this question by identifying a pair of directions where the projections (namely, u and v) of the random variables, $\boldsymbol{x}$ and $\boldsymbol{y}$ yield maximum correlation $\rho_{u,v} = \mathrm{COV}(u,v)/\sigma_u\sigma_v$. Here, $\mathrm{COV}(u,v)$ denotes the covariance function and $\sigma_\cdot$ gives the standard deviation. During the last decade, the CCA formulation has been broadly applied to various unsupervised learning

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 251–267, 2014.

problems in computer vision and machine learning including image retrieval [11], face/gait recognition [38], super-resolution [19] and action classification [24].

Beyond the applications described above, a number of works have recently investigated the use of CCA in analyzing neuroimaging data [3], which is a main focus of this paper. Here, for each participant in a clinical study, we acquire different types of images such as Magnetic Resonance (MRI), Computed Tomography (CT) and functional MRI. It is expected that each imaging modality captures a unique aspect of the underlying disease pathology. Therefore, given a group of N subjects and their corresponding brain images, we may want to identify strong relationships (e.g., anatomical/functional correlations) across different image types. When performed across different diseases, such an analysis will reveal insights into what is similar and what is different across diseases even when their symptomatic presentation may be similar. Alternatively, CCA may serve a feature extraction role. That is, the brain regions found to be strongly correlated can be used directly in downstream statistical analysis. In a study of a large number of subjects, rather than performing a hypothesis test on *all* brain voxels independently for each imaging modality, restricting the number of tests only to the set of 'relevant' voxels (found via CCA) is known to improve statistical power (since the False Discovery Rate correction will be less severe).

The classical version of CCA described above concurrently seeks two linear subspaces (straight lines) in *vector spaces* $\mathbf{R}^m$ and $\mathbf{R}^n$ for the two multi-variate random variables $\boldsymbol{x}$ and $\boldsymbol{y}$. The projection on to the straight line (linear subspace) is obtained by an inner product. This formulation is broadly applicable but encounters problems for manifold-valued data that are becoming increasingly important in present day research. For example, diffusion tensor magnetic resonance images (DTI) allow one to infer the diffusion tensor characterizing the anisotropy of water diffusion at each voxel in an image volume. This tensorial feature can be visualized as an ellipsoid and represented by a 3×3 symmetric positive definite (SPD) matrix at each voxel in the acquired image volume. Neither the individual SPD matrices nor the field of these SPD matrices lie in a vector space but instead are elements of a negatively curved Riemannian manifold where standard vector space operations are not valid. Hence, classical CCA is not applicable in this setting. For T1-weighted Magnetic resonance images (MRIs), we are frequently interested in analyzing not just the 3D intensity image on its own, but rather a quantity that captures the deformation field between each image and a *population template*. A registration between the image and the template yields the deformation field required to align the image pairs and the determinant of the Jacobian J of this deformation at each voxel is a commonly used feature that captures local volume changes [6,17]. Quantities such as the Cauchy deformation tensor defined as $\sqrt{J^T J}$ have been reported in literature for use in morphometric analysis [18]. The input to the statistical analysis is a 3D image of voxels, where each voxel corresponds to a matrix $\sqrt{J^T J} \succ 0$ (the Cauchy deformation tensor). Another example of manifold-valued fields is derived from high angular resolution diffusion images (HARDI) and can be used to compute the ensemble average propagators (EAPs) at each voxel of the given

HARDI data. The EAP is a probability density function that is related to the diffusion sensitized MR signal via the Fourier transform [5]. Since an EAP is a probability density function, by using a square root parameterization of this density function, it is possible to identify it with a point on the unit Hilbert Sphere. Once again, to perform any statistical analysis of these data derived features, we cannot apply standard vector-space operations since the unit Hilbert sphere is a positively curved manifold. When analyzing real brain imaging data, it is entirely possible that no meaningful correlations exist in the data. The key difficulty is that we do not know whether the experiment (i.e., inference) failed because there is in fact no statistically meaningful signal in the dataset or if the algorithms being used are sub-optimal.

Related Work. There are two somewhat distinct bodies of work that are related to and motivate this work. The first one relates to the extensive study of the classical CCA and its non-linear variants. These include various interesting results based on kernelization [1,4,12], neural networks [25,16], and deep architectures [2]. Most, if not all of these strategies extend CCA to arbitrary nonlinear spaces. However, this flexibility brings with it the associated issues of model selection (and thereby, regularization), controlling the complexity of the neural network structure, choosing an appropriate activation function and so on. It is an interesting question though not completely clear to us what type of a regularizer should be used if one were to explicitly impose a Riemannian structure on the objectives described in the works above. As opposed to regularization, the second line of work incorporates the specific geometry of the data directly within the estimation problem. Various statistical constructs have been generalized to Riemannian manifolds: these include regression [39,31], classification [36], kernel methods [21], margin-based and boosting classifiers [26], interpolation, convolution, filtering [10] and dictionary learning [14,27]. Among the most closely related are ideas related to projective dimensionality reduction methods. For instance, the generalization of Principal Components analysis (PCA) via the so-called Principal Geodesic Analysis (PGA) [9], Geodesic PCA [20], Exact PGA [33], Horizontal Dimension Reduction [32] with frame bundles, and an extension of PGA to the product space of Riemannian manifolds, namely, tensor fields [36]. It is important to note that except the non-parametric method of [34], most of these strategies focus on one rather than two sets of random variables (as is the case in CCA). Even in this setting, the first results on successful generalization of parametric regression models to Riemannian manifolds is relatively recent: geodesic regression [8,29] and polynomial regression [13] (note that the adaptive CCA formulation in [37] seems related to our work but is not designed for manifold-valued data).

This paper provides a parametric model between two different tensor fields on a Riemannian manifold, which is a significant step beyond these recent works. The CCA formulation we present requires the optimization of functions over either a single product manifold or a pair of product manifolds (of different dimensions) concurrently. The latter problem involving product manifolds of different dimensions will not be addressed in this paper. Note that in general, on manifolds the projection operation does not have a nice closed form solution. So, we need to perform projections via an optimization scheme on the two manifolds and find the best pair of geodesic subspaces. We provide a precise solution to this problem. To our knowledge, this is the first extension of CCA to Riemannian manifolds. Our approach has two advantages relative to other non-linear extensions of CCA. The first advantage is that no model selection is required. Also our method incorporates the known geometry of data space. Our **key contributions** are: **a)** A principled generalization of CCA for Riemannian manifolds; **b)** First, a numerical optimization scheme for identifying the subspaces and later, single path algorithms with approximate projections (both these ideas may be applicable beyond the CCA formulation). **c)** Providing experimental evidence how the Riemannian CCA formulation expands the operating range of statistical analysis of neuroimaging data.

2 Canonical Correlation in Euclidean Space

First, we will briefly review the classical CCA in Euclidean space to motivate the rest of our presentation. Recall that Pearson's product-moment correlation coefficient is a quantity to measure the relationship of two random variables, $x \in \mathbf{R}$ and $y \in \mathbf{R}$. For one dimensional random variables,

$$\rho_{x,y} = \frac{\mathrm{COV}(x,y)}{\sigma_x \sigma_y} = \frac{\mathbb{E}[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2}\sqrt{\sum_{i=1}^{N}(y_i-\mu_y)^2}} \quad (1)$$

For high dimensional data, $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, we cannot however perform a direct calculation as above. So, we need to project each set of variables on to a special axis in each space $\mathcal{X}$ and $\mathcal{Y}$. CCA generalizes the concept of correlation to random vectors (potentially of different dimensions). It is convenient to think of CCA as a measure of correlation between two multivariate data based on the *best* projection which maximizes their mutual correlation.

Canonical Correlation for $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ is given by

$$\max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \mathrm{corr}(\pi_{\boldsymbol{w}_x}(\boldsymbol{x}), \pi_{\boldsymbol{w}_y}(\boldsymbol{y})) = \max_{\boldsymbol{w}_x,\boldsymbol{w}_y} \frac{\sum_{i=1}^{N} \boldsymbol{w}_x^T(\boldsymbol{x}_i-\boldsymbol{\mu}_x)\boldsymbol{w}_y^T(\boldsymbol{y}_i-\boldsymbol{\mu}_y)}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{w}_x^T(\boldsymbol{x}_i-\boldsymbol{\mu}_x))^2}\sqrt{\sum_{i=1}^{N}\left(\boldsymbol{w}_y^T(\boldsymbol{y}_i-\boldsymbol{\mu}_y)\right)^2}} \quad (2)$$

where $\pi_{\boldsymbol{w}_x}(\boldsymbol{x}) := \arg\min_{t\in\mathbb{R}} \mathrm{d}(t\boldsymbol{w}_x, \boldsymbol{x})^2$. We will call $\pi_{\boldsymbol{w}_x}(\boldsymbol{x})$ the *projection coefficient* for $\boldsymbol{x}$ (similarly for $\boldsymbol{y}$). Define $S_{\boldsymbol{w}_x}$ as the subspace which is the span of $\boldsymbol{w}_x$. The projection of $\boldsymbol{x}$ on to $S_{\boldsymbol{w}_x}$ is given by $\Pi_{S_{\boldsymbol{w}_x}}(\boldsymbol{x})$. We can then verify that the relationship between the projection and the projection coefficient is,

$$\Pi_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}) := \arg\min_{x' \in S_{\boldsymbol{w}_x}} \mathrm{d}(\boldsymbol{x},\boldsymbol{x}')^2 = \frac{\boldsymbol{w}_x^T\boldsymbol{x}}{\|\boldsymbol{w}_x\|}\frac{\boldsymbol{w}_x}{\|\boldsymbol{w}_x\|} = \frac{\boldsymbol{w}_x^T\boldsymbol{x}}{\|\boldsymbol{w}_x\|^2}\boldsymbol{w}_x = \pi_{\boldsymbol{w}_x}(\boldsymbol{x})\boldsymbol{w}_x \quad (3)$$

In the Euclidean space, $\Pi_{S_{\boldsymbol{w}_x}}(\boldsymbol{x})$ has a closed form solution. In fact, it is obtained by an inner product, $\boldsymbol{w}_x^T\boldsymbol{x}$. Hence, by replacing the projection coefficient $\pi_{\boldsymbol{w}_x}(\boldsymbol{x})$ with $\boldsymbol{w}_x^T\boldsymbol{x}/\|\boldsymbol{w}_x\|^2$ and after a simple calculation, one obtains the form in (2). Without loss of generality, assume that $\boldsymbol{x}, \boldsymbol{y}$ are centered. Then the optimization problem can be written as,

$$\max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \boldsymbol{w}_x^T X^T Y \boldsymbol{w}_y \text{ subject to } \boldsymbol{w}_x^T X^T X \boldsymbol{w}_x = \boldsymbol{w}_y^T Y^T Y \boldsymbol{w}_y = 1 \tag{4}$$

where $\boldsymbol{x}, \boldsymbol{w}_x \in \mathbb{R}^m$, $\boldsymbol{y}, \boldsymbol{w}_y \in \mathbb{R}^n$, $X = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_N]^T$ and $Y = [\boldsymbol{y}_1 \ldots \boldsymbol{y}_N]^T$. The only difference here is that we remove the denominator. Instead, we have two equality constraints (note that correlation is scale-invariant).

3 Mathematical Preliminaries

We now briefly summarize certain basic concepts [7] which we will use later.

Riemannian Manifolds. A *differentiable manifold* [7] of dimension n is a set $\mathcal{M}$ and a 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. In other words, a differentiable manifold $\mathcal{M}$ is a topological space that is locally similar to an 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.

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. 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 $\mathbf{R}^n$. The geodesic curve from x_i to x_j can be parameterized by a tangent vector in the tangent space 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, $\text{Log}(y_i, \cdot) : \mathcal{M} \to T_{y_i}\mathcal{M}$. Separate from these notations, matrix exponential (and logarithm) are given as $\exp(\cdot)$ (and $\log(\cdot)$).

Intrinsic Mean. Let $\mathrm{d}(\cdot, \cdot)$ define the geodesic distance between two points. The intrinsic (or Karcher) mean of a set of points $\{x_i\}$ with non-negative weights $\{w_i\}$ is the minimizer of,

$$\bar{y} = \arg\min_{y \in \mathcal{M}} \sum_{i=1}^{N} w_i \mathrm{d}(y, y_i)^2, \tag{5}$$

which may be an arithmetic, geometric or harmonic mean depending on $\mathrm{d}(\cdot, \cdot)$.

On manifolds, the Karcher mean with distance $\mathrm{d}(y_i, y_j) = \|\text{Log}_{y_i} y_j\|$ is, $\sum_{i=1}^{N} \text{Log}_{\bar{y}} y_i = 0$. This identity implies that $\bar{y}$ is a local minimum which has a zero norm gradient [22], 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.

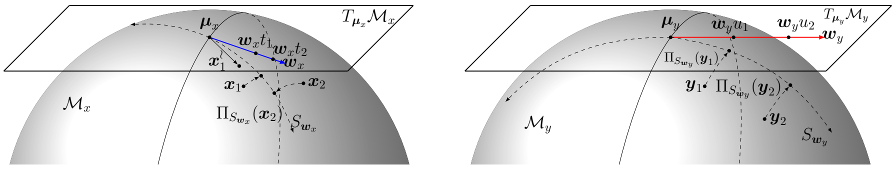

Fig. 1. CCA on Riemannian manifolds. CCA searches geodesic submanifolds (subspaces), S_{w_x} and S_{w_y} at the Karcher mean of data on each manifold. Correlation between projected points $\{\mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}_i)\}_{i=1}^N$ and $\{\mathit{\Pi}_{S_{\boldsymbol{w}_y}}(\boldsymbol{y}_i)\}_{i=1}^N$ is equivalent to the correlation between *projection coefficients* $\{t_i\}_{i=1}^N$ and $\{u_i\}_{i=1}^N$. Although $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to the same manifold we show them in different plots for ease of explanation.

Geodesically Convex. A subset C of $\mathcal{M}$ is said to be a *geodesically convex set* if there is a minimizing geodesic curve in C between *any* two points in C. This assumption is commonly used [8] and essential to ensure that the Riemannian operations such as the exponential and logarithm maps are well-defined.

4 A Model for CCA on Riemannian Manifolds

We now present a step by step derivation of our Riemannian CCA model. Classical CCA finds the mean of each data modality. Then, it maximizes correlation between projected data on each subspace at the mean. Similarly, CCA on manifolds must first compute the intrinsic mean (i.e., Karcher mean) of each data set. It must then identify a 'generalized' version of a subspace at each Karcher mean to maximize the correlation of projected data. The generalized form of a subspace on Riemannian manifolds has been studied in the literature [33,26,20,9]. The so-called *geodesic submanifold* [9,36,23] which has been used for geodesic regression serves our purpose well and is defined as $S = \mathrm{Exp}(\boldsymbol{\mu}, \mathrm{span}(\{\boldsymbol{v}_i\}) \cap U)$, where $U \subset T_{\boldsymbol{\mu}}\mathcal{M}$, and $\boldsymbol{v}_i \in T_{\boldsymbol{\mu}}\mathcal{M}$ [9]. When S has only one tangent vector $\boldsymbol{v}$, then the geodesic submanifold is simply a geodesic curve, see Figure 1.

We can now proceed to formulate the precise form of projection on to a geodesic submanifold. Recall that when given a point, its projection on a set is the closest point in the set. So, the projection on to a geodesic submanifold (S) must be a function satisfying this behavior. This is given by,

$$\mathit{\Pi}_S(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}' \in S} \mathrm{d}(\boldsymbol{x}, \boldsymbol{x}')^2 \tag{6}$$

In Euclidean space, the projection on a convex set (e.g., subspace) is unique. It is also unique on some manifolds under special conditions, e.g., quaternion sphere [30]. However, the uniqueness of the projection on geodesic submanifolds in general conditions cannot be ensured. Like other methods, we assume that given the specific manifold and the data, the projection is well-posed.

Finally, the correlation of points (*after* projection) can be measured by the distance from the mean to the projected points. To be specific, the projection on a geodesic submanifold corresponding to $\boldsymbol{w}_x$ in classical CCA is given by

$$\mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}) \quad := \arg \min_{\boldsymbol{x}' \in S_{\boldsymbol{w}_x}} \|\mathrm{Log}(\boldsymbol{x}, \boldsymbol{x}')\|^2_{\boldsymbol{x}} \tag{7}$$

$S_{\boldsymbol{w}_x} := \mathrm{Exp}(\boldsymbol{\mu}_x, \mathrm{span}\{\boldsymbol{w}_x\} \cap U)$ where $\boldsymbol{w}_x$ is a basis tangent vector and $U \subset T_{\boldsymbol{\mu}_x}\mathcal{M}_x$ is a small neighborhood of $\boldsymbol{\mu}_x$. The expression for *projection coefficients* can now be given as

$$t_i = \pi_{\boldsymbol{w}_x}(\boldsymbol{x}_i) \quad := \arg \min_{t'_i \in (-\epsilon, \epsilon)} \|\mathrm{Log}(\mathrm{Exp}(\boldsymbol{\mu}_x, t'_i \boldsymbol{w}_x), \boldsymbol{x}_i)\|^2_{\boldsymbol{\mu}_x} \tag{8}$$

where $\boldsymbol{x}_i, \boldsymbol{\mu}_x \in \mathcal{M}_x$, $\boldsymbol{w}_x \in T_{\boldsymbol{\mu}_x}\mathcal{M}_x, t_i \in \mathbf{R}$. The term, $u_i = \pi_{\boldsymbol{w}_y}(\boldsymbol{y})$ is defined analogously. t_i is a real value to obtain the point $\mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}) = \mathrm{Exp}(\boldsymbol{\mu}_x, t_i \boldsymbol{w}_x)$. As mentioned above, $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to the same manifold. Note that we are dealing with a single manifold, however, we use two different notations $\mathcal{M}_x$, and $\mathcal{M}_y$ to show that they are differently distributed for ease of discussion.

Notice that we have $\mathrm{d}(\boldsymbol{\mu}_x, \mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}_i)) = \|\mathrm{Log}(\boldsymbol{\mu}_x, \mathrm{Exp}(\boldsymbol{\mu}_x, \boldsymbol{w}_x t_i))\|_{\boldsymbol{\mu}_x} = t_i \|\boldsymbol{w}_x\|_{\boldsymbol{\mu}_x}$. By inspection, this shows that the projection coefficient is proportional to the length of the geodesic curve from the base point $\boldsymbol{\mu}_x$ to the projection of $\boldsymbol{x}$, $\mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x})$. Correlation is scale invariant, as expected. Therefore, the correlation between projected points $\{\mathit{\Pi}_{S_{\boldsymbol{w}_x}}(\boldsymbol{x}_i)\}_{i=1}^N$ and $\{\mathit{\Pi}_{S_{\boldsymbol{w}_y}}(\boldsymbol{y}_i)\}_{i=1}^N$ reduces to the correlation between the quantities that serve as projection coefficients here, $\{t_i\}_{i=1}^N$ and $\{u_i\}_{i=1}^N$.

Putting these pieces together, we obtain our generalized formulation for CCA,

$$\rho_{\boldsymbol{x},\boldsymbol{y}} = \mathrm{corr}(\pi_{\boldsymbol{w}_x}(\boldsymbol{x}), \pi_{\boldsymbol{w}_y}(\boldsymbol{y})) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}} \frac{\sum_{i=1}^N (t_i - \bar{t})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^N (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^N (u_i - \bar{u})^2}} \tag{9}$$

where $t_i = \pi_{\boldsymbol{w}_x}(\boldsymbol{x}_i)$, $\boldsymbol{t} := \{t_i\}$, $u_i = \pi_{\boldsymbol{w}_y}(\boldsymbol{y}_i)$, $\boldsymbol{u} := \{u_i\}$, $\bar{t} = \frac{1}{N}\sum_{i=1}^N t_i$ and $\bar{u} = \frac{1}{N}\sum_{i=1}^N u_i$. Expanding out components in (9) further, it takes the form,

$$\begin{aligned} \rho_{\boldsymbol{x},\boldsymbol{y}} = & \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}} \frac{\sum_{i=1}^N (t_i - \bar{t})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^N (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^N (u_i - \bar{u})^2}} \\ s.t. \quad t_i & = \arg \min_{t_i \in (-\epsilon, \epsilon)} \|\mathrm{Log}(\mathrm{Exp}(\boldsymbol{\mu}_x, t_i \boldsymbol{w}_x), \boldsymbol{x}_i)\|^2, \forall i \in \{1, \ldots, N\} \\ u_i & = \arg \min_{u_i \in (-\epsilon, \epsilon)} \|\mathrm{Log}(\mathrm{Exp}(\boldsymbol{\mu}_y, u_i \boldsymbol{w}_y), \boldsymbol{y}_i)\|^2, \forall i \in \{1, \ldots, N\} \end{aligned} \tag{10}$$

Directly, we see that (10) is a multilevel optimization and solutions from nested sub-optimization problems may be needed to solve the higher level problem. It turns out that deriving the first order optimality conditions suggests a cleaner approach.

Define $f(\boldsymbol{t}, \boldsymbol{u}) := \frac{\sum_{i=1}^N (t_i - \bar{t})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^N (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^N (u_i - \bar{u})^2}}$, $g(t_i, \boldsymbol{w}_x) := \|\mathrm{Log}(\mathrm{Exp}(\boldsymbol{\mu}_x, t_i \boldsymbol{w}_x), \boldsymbol{x}_i)\|^2$, and $g(u_i, \boldsymbol{w}_y) := \|\mathrm{Log}(\mathrm{Exp}(\boldsymbol{\mu}_y, u_i \boldsymbol{w}_y), \boldsymbol{y}_i)\|^2$. Then, we may replace the equality constraints in (10) with optimality conditions rather than another optimization problem for each i. Using this idea, we have

$$\begin{aligned} \rho(\boldsymbol{w}_x, \boldsymbol{w}_y) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}} & f(\boldsymbol{t}, \boldsymbol{u}) \\ s.t.\ & \nabla_{t_i} g(t_i, \boldsymbol{w}_x) = 0, \nabla_{u_i} g(u_i, \boldsymbol{w}_y) = 0, \forall i \in \{1, \ldots, N\} \end{aligned} \tag{11}$$

5 Optimization Schemes

We present two different algorithms to solve the problem of computing CCA on Riemannian manifolds. The first algorithm is based on a numerical optimization for (11). We only summarize the main model here and provide all technical details in the extended version for space reasons. Subsequently, we present the second approach which is based on an approximation for a more efficient algorithm.

5.1 An Augmented Lagrangian Method

The augmented Lagrangian technique is a well known variation of the penalty method for constrained optimization problems. Given a constrained optimization problem $\max f(\boldsymbol{x})$ s.t. $c_i(\boldsymbol{x}) = 0, \forall i$, the augmented Lagrangian method solves a sequence of the following models while increasing ν_k.

$$\max f(\boldsymbol{x}) + \sum_i \lambda_i c_i(\boldsymbol{x}) - \nu^k \sum_i c_i(\boldsymbol{x})^2 \tag{12}$$

The augmented Lagrangian formulation for our CCA formulation is given by

$$\begin{aligned} \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}} \mathcal{L}_A(\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\lambda}^k; \nu^k) = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y, \boldsymbol{t}, \boldsymbol{u}} f(\boldsymbol{t}, \boldsymbol{u}) + \sum_i^N \lambda_{t_i}^k \nabla_{t_i} g(t_i, \boldsymbol{w}_x) + \\ \sum_i^N \lambda_{u_i}^k \nabla_{u_i} g(u_i, \boldsymbol{w}_y) - \frac{\nu^k}{2} \left(\sum_{i=1}^N \nabla_{t_i} g(t_i, \boldsymbol{w}_x)^2 + \nabla_{u_i} g(u_i, \boldsymbol{w}_y)^2 \right) \end{aligned} \tag{13}$$

The pseudocode for our algorithm is summarized in Algorithm 1.

Remarks. Note that for Algorithm 1, we need the second derivative of g, in particular, for $\frac{d^2}{dwdt} g$, $\frac{d^2}{dt^2} g$. The literature does not provide a great deal of guidance on second derivatives of functions involving Log$(\cdot)$ and Exp$(\cdot)$ maps on general Riemannian manifolds. However, depending on the manifold, it can be obtained analytically or numerically (see extended version of the paper).

Approximate strategies. It is clear that the core difficulty in deriving the algorithm above was the lack of a closed form solution to projections on to geodesic submanifolds. If however, an approximate form of the projection can lead to significant gains in computational efficiency with little sacrifice in accuracy, it is worthy of consideration. The simplest approximation is to use a Log-Euclidean model. But it is well known that the Log-Euclidean is reasonable for data that are tightly clustered on the manifold and not otherwise. Further, the Log-Euclidean metric lacks the important property of affine invariance. We can obtain a more

Algorithm 1. Riemannian CCA based on the Augmented Lagarangian method

```
$\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathcal{M}_x$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_N \in \mathcal{M}_y$
Given $\nu^0 > 0, \tau^0 > 0$, starting points $(\boldsymbol{w}_x^0, \boldsymbol{w}_y^0, \boldsymbol{t}^0, \boldsymbol{u}^0)$ and $\boldsymbol{\lambda}^0$
for $k = 0, 1, 2 \dots$ do
    Start at $(\boldsymbol{w}_x^k, \boldsymbol{w}_y^k, \boldsymbol{t}^k, \boldsymbol{u}^k)$
    Find an approximate minimizer $(\boldsymbol{w}_x^k, \boldsymbol{w}_y^k, \boldsymbol{t}^k, \boldsymbol{u}^k)$ of $\mathcal{L}_A(\cdot, \boldsymbol{\lambda}^k; \nu^k)$, and terminate
  when $\|\nabla \mathcal{L}_A(\boldsymbol{w}_x^k, \boldsymbol{w}_y^k, \boldsymbol{t}^k, \boldsymbol{u}^k, \boldsymbol{\lambda}^k; \nu^k)\| \leq \tau^k$
    if a convergence test for (11) is satisfied then
        Stop with approximate feasible solution
    end if
    $\lambda_{t_i}^{k+1} = \lambda_{t_i}^k - \nu^k \nabla_{t_i} g(t_i, \boldsymbol{w}_x), \forall i$
    $\lambda_{u_i}^{k+1} = \lambda_{u_i}^k - \nu^k \nabla_{u_i} g(u_i, \boldsymbol{w}_y), \forall i$
    Choose new penalty parameter $\nu^{k+1} \geq \nu^k$
    Set starting point for the next iteration
    Select tolerance $\tau^{k+1}$
end for
```

Algorithm 2. CCA with approximate projection

```
Input $X_1, \dots, X_N \in \mathcal{M}_y$, $Y_1, \dots, Y_N \in M_y$
Compute intrinsic mean $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y$ of $\{X_i\}$, $\{Y_i\}$
Compute $X_i^{\wr} = \mathrm{Log}(\boldsymbol{\mu}_x, X_i)$, $Y_i^{\wr} = \mathrm{Log}(\boldsymbol{\mu}_y, Y_i)$
Transform (using group action) $\{X_i^{\wr}\}$, $\{Y_i^{\wr}\}$ to the $T_I\mathcal{M}_x, T_I\mathcal{M}_y$
Perform CCA between $T_I\mathcal{M}_x, T_I\mathcal{M}_y$ and get axes $W_a \in T_I\mathcal{M}_x$, $W_b \in T_I\mathcal{M}_y$
Transform (using group action) $W_a, W_b$ to $T_{\boldsymbol{\mu}_x}\mathcal{M}_x$, $T_{\boldsymbol{\mu}_y}\mathcal{M}_y$
```

accurate projection using the submanifold expression given in [36]. The form of projection is,

$$\mathit{\Pi}_S(\boldsymbol{x}) \approx \mathrm{Exp}(\boldsymbol{\mu}, \sum_{i=1}^{d} \boldsymbol{v}_i \langle \boldsymbol{v}_i, \mathrm{Log}(\boldsymbol{\mu}, \boldsymbol{x}) \rangle_{\boldsymbol{\mu}}\ ) \tag{14}$$

where $\{\boldsymbol{v}_i\}$ are *orthonormal basis* at $T_{\boldsymbol{\mu}}\mathcal{M}$. The CCA algorithm with this approximation for the projection is summarized as Algorithm 2.

Finally, we provide a brief remark on one remaining issue. This relates to the question why we use group action rather than other transformations such as parallel transport. Observe that Algorithm 2 sends the data from the tangent space at the Karcher mean of the samples to the tangent space at Identity I. The purpose of the transformation is to put all samples at the Identity of the SPD manifold, to obtain a more accurate projection, which can be understood using (14). The projection and inner product depend on the anchor point μ. If μ is Identity, then there is no discrepancy between the Euclidean and the Riemannian inner products. Of course, one may use a parallel transport. However, group action may be substantially more efficient than parallel transport since the former does not require computing a geodesic curve (which is needed for parallel transport). Interestingly, it turns out that on SPD manifolds with a GL-invariant metric, parallel transport from an arbitrary point p to Identity I is *equivalent* to the transform using a group action. So, one can parallel transport tangent vectors from p to I using the group action more efficiently. The proof of Theorem 1 is available in the extended version.

Theorem 1. *On SPD manifold, let $\mathit{\Gamma}_{p\rightarrow I}(w)$ denote the parallel transport of $w \in T_p\mathcal{M}$ along the geodesic from $p \in \mathcal{M}$ to $I \in \mathcal{M}$. The parallel transport is equivalent to group action by $p^{-1/2}wp^{-T/2}$, where the inner product $\langle u, v\rangle_p = tr(p^{-1/2}up^{-1}vp^{-1/2})$.*

5.2 Extensions to the Product Riemannian Manifold

In the types of imaging datasets of interest in this paper, we seek to perform an analysis on an entire population of images (of multiple types). For such data, each image must be treated as a single entity, which necessitates extending the formulation above to a Riemannian product space.

Let us define a Riemannian metric on the product space $\boldsymbol{\mathcal{M}} = \mathcal{M}_1 \times \ldots \times \mathcal{M}_m$. A natural choice is the following idea from [36].

$$\langle \boldsymbol{X}_1, \boldsymbol{X}_2 \rangle_{\boldsymbol{P}} = \sum_{j=1}^{m} \langle X_1^j, X_2^j \rangle_{P^j} \tag{15}$$

where $\boldsymbol{X}_1 = \left(X_1^1, \ldots, X_1^m\right) \in \boldsymbol{\mathcal{M}}$, and $\boldsymbol{X}_2 = \left(X_2^1, \ldots, X_2^m\right) \in \boldsymbol{\mathcal{M}}$ and $\boldsymbol{P} = \left(P^1, \ldots, P^m\right) \in \boldsymbol{\mathcal{M}}$. Once we have the exponential and logarithm maps, CCA on a Riemannian product space can be directly performed by Algorithm 2. The exponential map $\mathrm{Exp}(\boldsymbol{P}, \boldsymbol{V})$ and logarithm map $\mathrm{Log}(\boldsymbol{P}, \boldsymbol{X})$ are given by

$$(\mathrm{Exp}(P^1, V^1), \ldots, \mathrm{Exp}(P^m, V^m)) \text{ and } (\mathrm{Log}(P^1, X^1), \ldots, \mathrm{Log}(P^m, X^m)) \tag{16}$$

respectively, where $\boldsymbol{V} = (V^1, \ldots, V^m) \in T_{\boldsymbol{P}}\boldsymbol{\mathcal{M}}$. The length of tangent vector is $\|\boldsymbol{V}\| = \sqrt{\|V^1\|_{P^1}^2 + \cdots + \|V^m\|_{P^m}^2}$, where $V^i \in T_{P^i}\mathcal{M}_i$. The geodesic distance between two points $\mathrm{d}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ on Riemannian product space is also measured by the length of tangent vector from one point to the other. So we have

$$\mathrm{d}(\boldsymbol{\mu}_x, \boldsymbol{X}) = \sqrt{\mathrm{d}(\mu_x^1, X^1)^2 + \cdots + \mathrm{d}(\mu_x^m, X^m)^2} \tag{17}$$

From our previous discussion of the relationship between *projection coefficients* and distance from the mean to points (after *projection*) in Section 4, we have $t_i = \mathrm{d}(\boldsymbol{\mu}_x, \mathit{\Pi}_{S_{\boldsymbol{W}_x}}(\boldsymbol{X}_i))/\|\boldsymbol{W}_x\|_{\boldsymbol{\mu}_x}$ and $t_i^j = \mathrm{d}(\mu_x^j, \mathit{\Pi}_{S_{W_x^j}}(X_i^j))/\|W_x^j\|_{\mu_x^j}$. By substitution, the *projection coefficients* on Riemannian product space are given by

$$t_i = \mathrm{d}(\boldsymbol{\mu}_x, \mathit{\Pi}_{S_{\boldsymbol{W}_x}}(\boldsymbol{X}_i))/\|\boldsymbol{W}_x\|_{\boldsymbol{\mu}_x} = \sqrt{\sum_{j}^{m} \left(t_i^j \left\|W_x^j\right\|_{\mu_x^j}\right)^2 / \sum_{j=1}^{m} \left\|W_x^j\right\|_{\mu_x^j}^2} \tag{18}$$

We can now mechanically substitute these “product space” versions of the terms in (18) to derive a CCA on Riemannian product space. The full model is provided in the extended version.

6 Experiments

6.1 CCA on SPD Manifolds

Diffusion tensors are symmetric positive definite matrices at each voxel in DTI. Let SPD(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 and $O(n)$ is the orthogonal group. The inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is given by $\langle u, v\rangle_p = \mathrm{tr}(p^{-1/2}up^{-1}vp^{-1/2})$. 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 $d(p,q)^2 = \mathrm{tr}(\log^2(p^{-1/2}qp^{-1/2}))$.

Here, the exponential map and logarithm map are defined as,

$$\mathrm{Exp}(p, v) = p^{1/2}\exp(p^{-1/2}vp^{-1/2})p^{1/2}, \;\; \mathrm{Log}(p, q) = p^{1/2}\log(p^{-1/2}qp^{-1/2})p^{1/2} \quad (19)$$

and the first derivative of g in equation (11) on SPD(n) is given by

$$\begin{aligned} \frac{d}{dt_i}g(t_i, \boldsymbol{w}_x) &= \frac{d}{dt_i}\|\mathrm{Log}(\mathrm{Exp}(\mu_x, t_iW_x), X_i)\|^2 = \frac{d}{dt_i}\mathrm{tr}[\log^2(X_i^{-1}S(t_i))] \\ &= 2\mathrm{tr}[\log(X_i^{-1}S(t_i))S(t_i)^{-1}\dot{S}(t_i)], \text{ according to Prop. 2.1 in [28]} \end{aligned} \quad (20)$$

where $S(t_i) = \mathrm{Exp}(\mu_x, t_iW_x) = \mu_x^{1/2}\exp^{t_iA}\mu_x^{1/2}$, and $\dot{S}(t_i) = \mu_x^{1/2}A\exp^{t_iA}\mu_x^{1/2}$ and $A = \mu_x^{-1/2}W_x\mu_x^{-1/2}$. The derivative of equality constraints, namely $\frac{d^2}{dWdt}g$, $\frac{d^2}{dt^2}g$ are calculated by numerical derivatives. Embedding the tangent vectors in the $n(n+1)/2$ dimensional space with orthonormal basis in the tangent space enables one to compute numerical differentiation. Details are provided in the extended paper.

6.2 Synthetic Experiments

In this section we provide experimental results using a synthetic dataset to evaluate the performance of Riemannian CCA. The samples are generated to be spread far apart on the manifold $\mathcal{M}(\equiv \mathrm{SPD}(3))$ so that the curvature of the manifold plays a key role in the maximization of the correlation function. In order to sample data from different regions of the manifold, we generate data around two well separated means $\mu_{x_1}, \mu_{x_2} \in \mathcal{X}$, $\mu_{y_1}, \mu_{y_2} \in \mathcal{Y}$ by perturbing the data randomly (see the extended version) in the corresponding tangent spaces. Fig. 2 shows the CCA results obtained by Riemannian and Euclidean methods. We can clearly see the improvements from the manifold approach by inspecting the correlation coefficients $\rho_{x,y}$ on the respective titles.

6.3 CCA for Multi-modal Risk Analysis

Motivation: We collected multi-modal magnetic resonance imaging (MRI) data to investigate the effects of risk for Alzheimer's disease (AD) on the white and gray matter in the brain. One of the central goals in analyzing this rich dataset

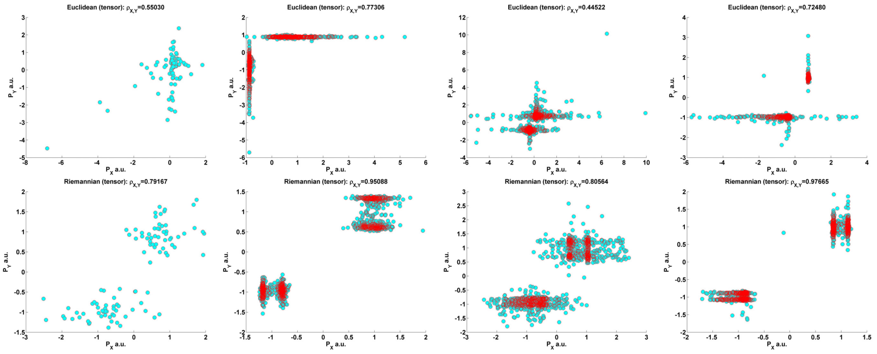

Fig. 2. Synthetic experiments showing the benefits of Riemannian CCA. The top row shows the projected data using the Euclidean CCA and the bottom using Riemannian CCA. P_X and P_Y denote the projected axes. Each column represents a synthetic experiment with a specific set of $\{\mu_{x_j}, \epsilon_{x_j}; \mu_{y_j}, \epsilon_{y_j}\}$. The first column presents results with 100 samples while the three columns on the right show with 1000 samples. The improvements in the correlation coefficients $\rho_{x,y}$ can be clearly seen from the corresponding titles.

is to find statistically significant AD risk ↔ brain relationships. We can adopt many different ways of modeling these relationships but a potentially useful way is to analyze multi modality imaging data simultaneously, using CCA.

Risk for AD is characterized by their familial history (FH) status as well as APOE genotype risk factor. In the current experiments, we include a subset of 343 subjects and first investigate the effects of age and gender in a multimodal fashion since these variables are also important factors in healthy aging.

Brain structure is characterized by diffusion weighted images (DWI) for white matter and T1-weighted (T1W) image data for the gray matter. DWI data provides us information about the microstructure of the white matter. We use diffusion tensor ($\mathcal{D} \in \text{SPD}(3)$) model to represent the diffusivity in the microstructure. T1W data can be used to obtain volumetric properties of the gray-matter. The volumetric information is obtained from Jacobian matrices (J) of the diffeomorphic mapping to a population specific template. These Jacobian matrices can be used to obtain the Cauchy deformation tensors which also belong to SPD(3).

Hippocampus and cingulum bundle (shown in Fig. 3) are two important regions in the brain. They are *a priori* believed to be significant in AD↔brain structure relationships, primarily due to the role of hippocampus in memory function and the projections of cingulum onto the hippocampus. However, detecting *risk*-brain relationships *before* the memory/cognitive function is impaired is difficult due to several factors (such as noise in the data, small sample and effect sizes, type I error due to multiple comparisons.). One approach to improve the statistical power in such a setting would be to perform tests on average properties in regions of interest (ROI) in the brain. This procedure reduces both noise and the number of comparisons/tests. However, taking averages will also dampen the signal of interest which is already weak in such pre-clinical studies.

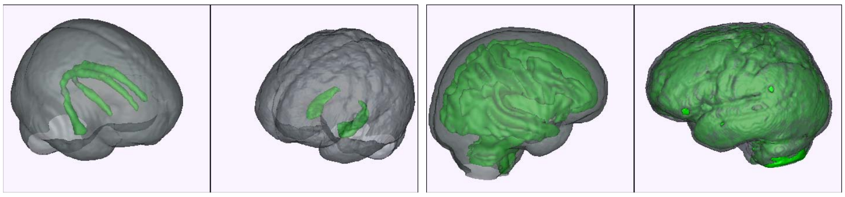

Fig. 3. Shown on the left are the bilateral cingulum bundles (green) inside a brain surface obtained from a population DTI template. Similarly on the right are the bilateral hippocampi. The gray and white matter ROIs are also shown on the right.

CCA can take the multi-modal information from the imaging data and project the voxels into a space where the signal of interest is likely to be stronger.

Experimental Design: The key multimodal linear relations we examine are

$$\begin{aligned} Y_{\text{DTI}} &= \beta_0 + \beta_1 \text{Gender} + \beta_2 X_{\text{T1W}} + \beta_3 X_{\text{T1W}} \cdot \text{Gender} + \varepsilon, \\ Y_{\text{DTI}} &= \beta_0' + \beta_1' \text{AgeGroup} + \beta_2' X_{\text{T1W}} + \beta_3' X_{\text{T1W}} \cdot \text{AgeGroup} + \varepsilon, \end{aligned}$$

where the AgeGroup is defined as a categorical variable with 0 (middle aged) if the age of the subject ≤ 65 and 1 (old) otherwise. The sample under investigation is between 43 and 75 years of age. The statistical tests ask if we can reject the Null hypotheses $\beta_3 = 0$ and $\beta_3' = 0$ using our data at $\alpha = 0.05$. We report the results from the following four sets of analyses: **(i)** Classical ROI-average analysis: This is a standard type of setting where the brain measurements in an ROI are averaged. Here $Y_{\text{DTI}} = \overline{\text{MD}}$ i.e., the average mean diffusivity in the cingulum bundle. $X_{\text{T1W}} = \overline{\log|J|}$ i.e., the average volumetric change (relative to the population template) in the hippocampus. **(ii)** Euclidean CCA using scalar measures (MD and $\log|J|$) in the ROIs: Here, the voxel data is projected using the classical CCA approach [35] i.e., $Y_{\text{DTI}} = \mathbf{w}_{\text{MD}}^T \text{MD}$ and $X_{\text{T1W}} = \mathbf{w}_{\log|J|}^T \log|J|$. **(iii)** Euclidean CCA using $\mathcal{D}$ and $\mathcal{J}$ in the ROIs: This setting is an improvement to the setting above in that the projections are performed using the full tensor data [35]. Here $Y_{\text{DTI}} = \mathbf{w}_{\mathcal{D}}^T \mathcal{D}$ and $X_{\text{T1W}} = \mathbf{w}_{\mathcal{J}}^T \mathcal{J}$. **(iv)** Riemannian CCA using $\mathcal{D}$ and $\mathcal{J}$ in the ROIs: Here $Y_{\text{DTI}} = \langle \mathbf{w}_{\mathcal{D}}, \mathcal{D} \rangle_{\mu_{\mathcal{D}}}$ and $X_{\text{T1W}} = \langle \mathbf{w}_{\mathcal{J}}, \mathcal{J} \rangle_{\mu_{\mathcal{J}}}$.

The findings are shown in Fig. 4. We can see that the performance of CCA using the full tensor information improves the statistical significance for both Euclidean and Riemannian approaches. The weight vectors in the different settings for both Euclidean and Riemannian CCA are shown in Fig. 5 top row. We would like to note that there are several different approaches of using the data from CCA and we performed experiments with full gray matter and white matter regions in the brain whose results are included in the extended version. We show the representative weight vectors (in Fig. 5 bottom row) obtained using the full brain analyses. Interestingly, the weight vectors are spatially cohesive even without enforcing any spatial constraints. What is even more remarkable is that the regions picked between the DTI and T1W modalities are complimentary in a biological sense. Specifically, when performing our CCA on the ROIs, although the cingulum bundle extends into the superior mid-brain regions the weights are

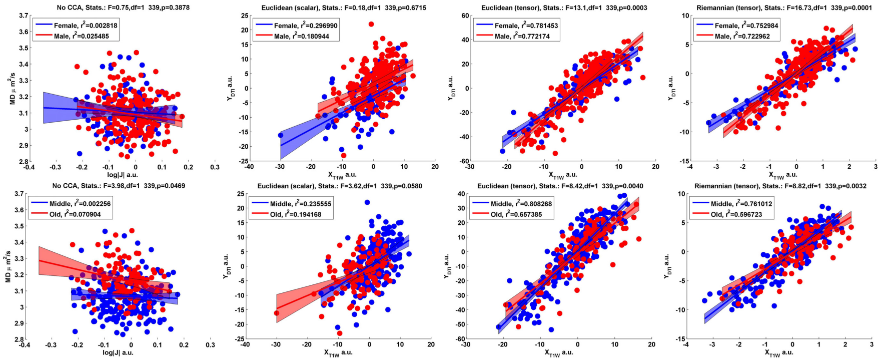

Fig. 4. Experimental evidence showing the improvements in statistical significance of finding the multi-modal risk-brain interaction effects. Top row shows the gender, volume and diffusivity interactions. Second row shows the interaction effects of the middle/old age groups.

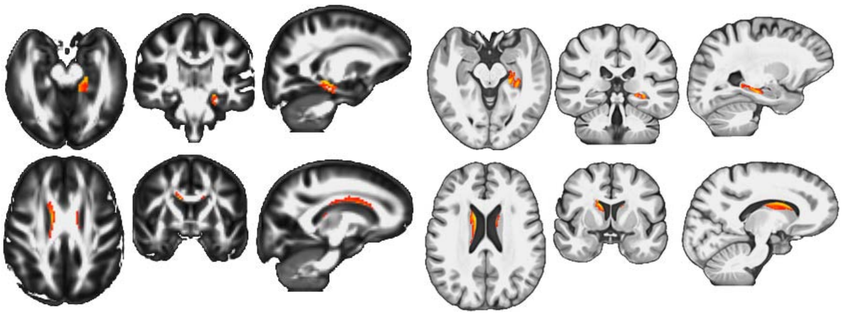

Fig. 5. Weight vectors (in red-yellow color) obtained from our Riemannian CCA approach. The weights are in arbitrary units. The top row is from applying Riemannian CCA on data from the cingulum and hippocampus ROIs (Fig. 3) while the bottom row is obtained using data from the entire white and gray matter regions of the brain. On the left (three columns) block we show the results in orthogonal view for DTI and on the right for T1W. The corresponding underlays are the population averages of the fractional anisotropy and T1W contrast images respectively.

non-zero in its hippocampal projections. In the case of entire white and gray matter regions, the volumetric difference (from the population template) in the inferior part of the corpus callosum seem to be highly cross-correlated to the diffusivity in the corpus callosum. Our CCA finds these projections without any a priori constraints in the optimization suggesting that performing CCA on the intrinsic nature of the data can reveal biologically meaningful patterns. Due to space constraints, we refer the interested reader to the extended version of the paper for additional details.

7 Conclusion

The classical CCA assumes that data live in a pair of vector spaces. However, many modern scientific disciplines require the analysis of data which belong to *curved* spaces where classical CCA is no longer applicable. Motivated by the properties of imaging data from neuroimaging studies, we generalize CCA to Riemannian manifolds. We employ differential geometry tools to extend operations in CCA to the manifold setting. Such a formulation results in a multi-level optimization problem. We derive solutions using the first order condition of projection and an augmented Lagrangian method. In addition, we also develop an efficient single path algorithm with approximate projections. Finally, we propose a generalization to the product space of SPD(n), namely, tensor fields allowing us to treat a full brain image as a point on the product manifold. On the experimental side, we presented neuroimaging findings using our proposed CCA on DTI and T1W imaging modalities on an Alzheimer's disease (AD) dataset focused on risk factors for this disease. Here, the proposed methods perform well and yield scientifically meaningful results. In closing, we note that our core optimization methods can be readily applied when maximizing correlation between data from two *different* types of Riemannian manifolds — this may open the doors to various other types of analysis not explicitly investigated in this paper.

Acknowledgments. This work was supported in part by NIH grants AG040396 (VS), AG037639 (BBB), AG021155 (SCJ), AG027161 (SCJ), NS066340 (BCV) and NSF CAREER award 1252725 (VS). Partial support was also provided by UW ADRC, UW ICTR, and Waisman Core grant P30 HD003352-45. The contents do not represent views of the Dept. of Veterans Affairs or the United States Government.

References

1. Akaho, S.: A kernel method for canonical correlation analysis. In: International Meeting on Psychometric Society (2001)
2. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML (2013)
3. Avants, B.B., Cook, P.A., Ungar, L., Gee, J.C., Grossman, M.: Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with SCCA. Neuroimage 50(3), 1004–1016 (2010)
4. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. The Journal of Machine Learning Research 3, 1–48 (2003)
5. Callaghan, P.T.: Principles of nuclear magnetic resonance microscopy. Oxford University Press (1991)
6. Chung, M., Worsley, K., Paus, T., Cherif, C., Collins, D., Giedd, J., Rapoport, J., Evans, A.: A unified statistical approach to deformation-based morphometry. NeuroImage 14(3), 595–606 (2001)
7. Do Carmo, M.P.: Riemannian geometry (1992)
8. Fletcher, P.T.: Geodesic regression and the theory of least squares on riemannian manifolds. International Journal of Computer Vision 105(2), 171–185 (2013)

9. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. Medical Imaging 23(8), 995–1005 (2004)
10. Goh, A., Lenglet, C., Thompson, P.M., Vidal, R.: A nonparametric Riemannian framework for processing high angular resolution diffusion images (HARDI), pp. 2496–2503 (2009)
11. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: CCA: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
12. Hardoon, D.R., et al.: Unsupervised analysis of fMRI data using kernel canonical correlation. NeuroImage 37(4), 1250–1259 (2007)
13. Hinkle, J., Fletcher, P.T., Joshi, S.: Intrinsic polynomials for regression on Riemannian manifolds. Journal of Mathematical Imaging and Vision, 1–21 (2014)
14. Ho, J., Xie, Y., Vemuri, B.: On a nonlinear generalization of sparse coding and dictionary learning. In: ICML, pp. 1480–1488 (2013)
15. Hotelling, H.: Relations between two sets of variates. Biometrika 28(3/4), 321–377 (1936)
16. Hsieh, W.W.: Nonlinear canonical correlation analysis by neural networks. Neural Networks 13(10), 1095–1105 (2000)
17. Hua, X., Gutman, B., et al.: Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. Neuroimage 57(1), 5–14 (2011)
18. Hua, X., Leow, A.D., Parikshak, N., Lee, S., Chiang, M.C., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M.: Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 43(3), 458–469 (2008)
19. Huang, H., He, H., Fan, X., Zhang, J.: Super-resolution of human face image using canonical correlation analysis. Pattern Recognition 43(7), 2532–2543 (2010)
20. Huckemann, S., Hotz, T., Munk, A.: Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. Statistica Sinica 20, 1–100 (2010)
21. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In: CVPR, pp. 73–80 (2013)
22. Karcher, H.: Riemannian center of mass and mollifier smoothing. Communications on Pure and Applied Mathematics 30(5), 509–541 (1977)
23. Kim, H.J., Adluru, N., Collins, M.D., Chung, M.K., Bendlin, B.B., Johnson, S.C., Davidson, R.J., Singh, V.: Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In: CVPR (2014)
24. Kim, T.K., Cipolla, R.: CCA of video volume tensors for action categorization and detection. PAMI 31(8), 1415–1428 (2009)
25. Lai, P.L., Fyfe, C.: A neural implementation of canonical correlation analysis. Neural Networks 12(10), 1391–1397 (1999)
26. Lebanon, G., et al.: Riemannian geometry and statistical machine learning. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science (2005)
27. Li, P., Wang, Q., Zuo, W., Zhang, L.: Log-Euclidean kernels for sparse representation and dictionary learning. In: ICCV, pp. 1601–1608 (2013)
28. Moakher, M.: A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM Journal on Matrix Analysis and Applications 26(3), 735–747 (2005)
29. Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: MICCAI, pp. 655–662 (2011)

30. Said, S., Courty, N., Le Bihan, N., Sangwine, S.J., et al.: Exact principal geodesic analysis for data on SO(3). In: Proceedings of the 15th European Signal Processing Conference, pp. 1700–1705 (2007)
31. Shi, X., Styner, M., Lieberman, J., Ibrahim, J.G., Lin, W., Zhu, H.: Intrinsic regression models for manifold-valued data. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 192–199. Springer, Heidelberg (2009)
32. Sommer, S.: Horizontal dimensionality reduction and iterated frame bundle development. In: Nielsen, F., Barbaresco, F. (eds.) GSI 2013. LNCS, vol. 8085, pp. 76–83. Springer, Heidelberg (2013)
33. Sommer, S., Lauze, F., Nielsen, M.: Optimization over geodesics for exact principal geodesic analysis. Advances in Computational Mathematics, 1–31
34. Steinke, F., Hein, M., Schölkopf, B.: Nonparametric regression between general Riemannian manifolds. SIAM Journal on Imaging Sciences 3(3), 527–563 (2010)
35. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3), 515–534 (2009)
36. Xie, Y., Vemuri, B.C., Ho, J.: Statistical analysis of tensor fields. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 682–689. Springer, Heidelberg (2010)
37. Yger, F., Berar, M., Gasso, G., Rakotomamonjy, A.: Adaptive canonical correlation analysis based on matrix manifolds. In: ICML (2012)
38. Yu, S., Tan, T., Huang, K., Jia, K., Wu, X.: A study on gait-based gender classification. IEEE Transactions on Image Processing 18(8), 1905–1910 (2009)
39. Zhu, H., Chen, Y., Ibrahim, J.G., Li, Y., Hall, C., Lin, W.: Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. Journal of the American Statistical Association 104(487) (2009)

Canonical Correlation Analysis on Riemannian Manifolds and its Applications (supplement)

Hyunwoo J. Kim[†] Nagesh Adluru[†] Barbara B. Bendlin[†]
Sterling C. Johnson[†] Baba C. Vemuri[§] Vikas Singh[†]
http://pages.cs.wisc.edu/ hwkim/projects/riem-cca

[†]University of Wisconsin–Madison [§]University of Florida

1 Summary

We provide additional experimental results and technical details for implementation which were not included in the main paper due to limited space. Also we discuss relationships between Log-Euclidean framework and group action framework for Riemannian CCA.

2 Parallel transport and Group action

Theorem 1. *On SPD manifold, let $\Gamma_{p\to I}(w)$ denote the parallel transport of $w \in T_p\mathcal{M}$ from $p \in \mathcal{M}$ to $I \in \mathcal{M}$. The parallel transport is equivalent to group action by $p^{-1/2}wp^{-T/2}$, where the inner product $\langle u, v\rangle_p = tr(p^{-1/2}up^{-1}vp^{-1/2})$.*

Proof. Parallel transport Γ from p to q is given by [1]

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

Let's transform the tangent vector w at $T_p\mathcal{M}$ to I by setting $q = I$.

$$\begin{aligned}\Gamma_{p\to I}(w) &= p^{1/2}rp^{-1/2}wp^{-1/2}rp^{1/2}\\ \text{where } r &= \exp(p^{-1/2}\frac{v}{2}p^{-1/2})\\ v &= \mathrm{Log}(p,I) = p^{1/2}\log(p^{-1/2}Ip^{-1/2})p^{1/2} = p^{1/2}\log(p^{-1})p^{1/2} \qquad \text{(a)}\end{aligned}$$

Then r is given as above

$$\begin{aligned}r &= \exp(p^{-1/2}\frac{v}{2}p^{-1/2})\\ &= \exp(p^{-1/2}p^{1/2}\log(p^{-1})p^{1/2}p^{-1/2}/2), \text{ by (a)}\\ &= \exp(\log(p^{-1})/2)\\ &= p^{-1/2} \qquad \text{(b)}\end{aligned}$$

$$\begin{aligned}\mathit{\Gamma}_{p\to I}(w) &= p^{1/2} r p^{-1/2} w p^{-1/2} r p^{1/2} \\ &= p^{1/2} p^{-1/2} p^{-1/2} w p^{-1/2} p^{-1/2} p^{1/2}, \text{ since } r = p^{-1/2} \text{ by (b)} \\ &= p^{-1/2} w p^{-1/2} \\ &= p^{-1/2} w p^{-T/2} \text{ since } p^{-1/2} \text{ is SPD}\end{aligned}$$

3 Extension to Riemannian product spaces

Let us define a Riemannian metric on the product space $\boldsymbol{\mathcal{M}} = \mathcal{M}_1 \times \mathcal{M}_2 \dots \mathcal{M}_m$ like the following [2].

$$\langle \boldsymbol{X}_1, \boldsymbol{X}_2 \rangle_{\boldsymbol{P}} = \sum_{j=1}^{m} \langle X_1^j, X_2^j \rangle_{P^j} \tag{1}$$

where $\boldsymbol{X}_1 = \left(X_1^1, \dots, X_1^m\right) \in \boldsymbol{\mathcal{M}}$, and $\boldsymbol{X}_2 = \left(X_2^1, \dots, X_2^m\right) \in \boldsymbol{\mathcal{M}}$ and $\boldsymbol{P} = \left(P^1, \dots, P^m\right) \in \boldsymbol{\mathcal{M}}$. The exponential map $\mathrm{Exp}(\boldsymbol{P}, \boldsymbol{V})$ and logarithm map $\mathrm{Log}(\boldsymbol{P}, \boldsymbol{X})$ are given by

$$(\mathrm{Exp}(P^1, V^1), \dots, \mathrm{Exp}(P^m, V^m)) \text{ and } (\mathrm{Log}(P^1, X^1), \dots, \mathrm{Log}(P^m, X^m)) \tag{2}$$

respectively, where $\boldsymbol{V} = (V^1, \dots, V^m) \in T_{\boldsymbol{P}}\boldsymbol{\mathcal{M}}$. The length of tangent vector is $\|\boldsymbol{V}\| = \sqrt{\|V^1\|_{P^1}^2 + \dots + \|V^m\|_{P^m}^2}$, where $V^i \in T_{P^i}\mathcal{M}_i$. The geodesic distance between two points $\mathrm{d}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ on Riemannian product space is also measured by the length of tangent vector from one point to the other. So we have

$$\mathrm{d}(\boldsymbol{\mu}_x, \boldsymbol{X}) = \sqrt{\mathrm{d}(\mu_x^1, X^1)^2 + \dots + \mathrm{d}(\mu_x^m, X^m)^2} \tag{3}$$

We have $t_i = \mathrm{d}(\boldsymbol{\mu}_x, \mathit{\Pi}_{S_{\boldsymbol{W}_x}}(\boldsymbol{X}_i))/\|\boldsymbol{W}_x\|_{\boldsymbol{\mu}_x}$ and $t_i^j = \mathrm{d}(\mu_x^j, \mathit{\Pi}_{S_{W_x^j}}(X_i^j))/\|W_x^j\|_{\mu_x^j}$. By substitution, the *projection coefficients* on Riemannian product space are given by

$$t_i = \mathrm{d}(\boldsymbol{\mu}_x, \mathit{\Pi}_{S_{\boldsymbol{W}_x}}(\boldsymbol{X}_i))/\|\boldsymbol{W}_x\|_{\boldsymbol{\mu}_x} = \sqrt{\sum_{j}^{m} \left(t_i^j \left\|W_x^j\right\|_{\mu_x^j}\right)^2 / \sum_{j=1}^{m} \left\|W_x^j\right\|_{\mu_x^j}^2} \tag{4}$$

We can now mechanically substitute these "product space" versions of the terms in (4) to derive a CCA on Riemannian product space.

Hence, the formulation of Riemannian CCA on Riemannian product space is given by

$$
\begin{aligned}
\rho(\boldsymbol{W}_a, \boldsymbol{W}_b) = \max_{\boldsymbol{W}_x, \boldsymbol{W}_y, \boldsymbol{t}, \boldsymbol{u}} & \frac{\sum_{i=1}^{N}(t_i - \bar{t})(u_i - \bar{u})}{\sqrt{\sum_{i=1}^{N}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{N}(u_i - \bar{u})^2}} \\
s.t. \quad t_i^j & = \arg \min_{t_i^j \in (-\epsilon, \epsilon)} \|\mathrm{Log}(\mathrm{Exp}(\mu_x^j, t_i^j W_x^j), X_i^j)\|^2, \forall i, \forall j \\
u_i^k & = \arg \min_{u_i^k \in (-\epsilon, \epsilon)} \|\mathrm{Log}(\mathrm{Exp}(\mu_y^k, u_i^k W_y^k), Y_i^k)\|^2, \forall i, \forall k \\
t_i & = \frac{\sqrt{\sum_{j=1}^{m}\left(t_i^j \|W_x^j\|_{\mu_x^j}\right)^2}}{\sqrt{\sum_{j=1}^{m}\|W_x^j\|_{\mu_x^j}^2}}, \quad u_i = \frac{\sqrt{\sum_{k=1}^{n}\left(u_i^k \|W_y^k\|_{\mu_y^k}\right)^2}}{\sqrt{\sum_{k=1}^{n}\|W_y^k\|_{\mu_y^k}^2}}, \forall i \\
\bar{t} & = \frac{1}{N}\sum_{i}^{N} t_i, \quad \bar{u} = \frac{1}{N}\sum_{i}^{N} u_i
\end{aligned}
\tag{5}
$$

where $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, m\}$, and $\forall k \in \{1, \ldots, n\}$. This can be optimized by similar constrained optimization algorithms as described in the main paper with relative small changes.

4 Implementation details

4.1 Synthetic data experiments

We create synthetic data to see different behaviours of Euclidean CCA and Riemannian CCA. Each data set has two clusters. Each cluster is perturbed randomly by Gaussian-like noise in each tangent space at cluster mean μ_{x_j} and μ_{y_j}, where $j \in \{1, 2\}$ is the index for cluster. The procedure on P_n is described in Algorithm 1.

Algorithm 1 Data synthesis

```
1: $\boldsymbol{\epsilon}' \in \mathbf{R}^{n(n+1)/2} \sim \mathcal{N}(0, \sigma I)$
2: $\boldsymbol{\epsilon}' \leftarrow \boldsymbol{\epsilon}' \min(\|\boldsymbol{\epsilon}'\|, c_1)/\|\boldsymbol{\epsilon}'\|$        ▷ $c_1$ is a parameter for a safeguard
3: $\epsilon_I \leftarrow \mathrm{mat}(\boldsymbol{\epsilon}')$,        ▷ tangent vector at $I$
4: Transform (using group action) $\epsilon_I$ to $T_\mu\mathcal{M}$
5: Perturb data $X \leftarrow \mathrm{Exp}(\mu, \epsilon_\mu)$, where $\epsilon_\mu \in T_\mu\mathcal{M}$
```

The algorithm 1 simulates truncated-Gaussian-like noise. The second step (safeguard) in the pseudocode ensures that the data lives in a reasonably small neighborhood. We define subroutines for mapping from $\mathbf{R}^{n(n+1)/2}$ to P_n or vice versa. For the sake of simplicity of discussion, we show subroutines for P_3.

$$\begin{aligned}\mathrm{vec}(S) &:= [s_{11}, \sqrt{2}s_{12}, \sqrt{2}s_{13}, s_{22}, \sqrt{2}s_{23}, s_{33}]^T, \text{where } S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ s_{31} & s_{32} & s_{33} \end{bmatrix} \\ \mathrm{mat}(\boldsymbol{v}) &:= \begin{bmatrix} v_1 & \frac{1}{\sqrt{2}}v_2 & \frac{1}{\sqrt{2}}v_3 \\ \frac{1}{\sqrt{2}}v_2 & v_4 & \frac{1}{\sqrt{2}}v_5 \\ \frac{1}{\sqrt{2}}v_3 & \frac{1}{\sqrt{2}}v_5 & v_6 \end{bmatrix}, \text{where } \boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_6 \end{bmatrix}^T \end{aligned} \tag{6}$$

'vec' embeds tangent vectors in $T_I\mathcal{M}$ into $\mathbf{R}^6$ preserving inner product. 'mat' is the inversion of 'vec'. By construction, we have $\langle S_1, S_2\rangle_I = \langle v_1, v_2\rangle$, where $v_i = \mathrm{vec}(S_i)$. In other words, the distance from base point/origin to each point is identical in two spaces by the construction. Using group actions and these subroutines, points can be mapped from an abitral tangent space $T_p\mathcal{M}$ to $\mathbf{R}^6$, or vice versa, where $p \in \mathcal{M}$. CCA algorithm with approximate projection, namely, Algorithm 2 in our main paper, can be implemented by these two subroutines with group actions. Also orthogonal basis at arbitrary tangent spaces can be easily found using this trick: 1. Pick a set of orthogonal basis in $\mathbf{R}^6$, 2. map them to $T_I\mathcal{M}$, 3. transform (using group action) from $T_I\mathcal{M}$ to $T_p\mathcal{M}$. Then the transformed tangent vectors are orthogonal basis in $T_p\mathcal{M}$. This is used for numerical differentiation of g in (11) in the main paper.

5 Comparison of Log-Euclidean framework and Group action framework for CCA on P_n

In this section, we compare two different approximations for Riemannian CCA. Projection on geodesic submanifolds can be approximated by inner product. In Log-Euclidean frameworks, data is treated as in Euclidean space once they are mapped onto tangent spaces. Hence Log-Euclidean CCA uses Euclidean inner product $\langle,\rangle$. However algorithm 2 with group actions in our main paper uses the inner product $\langle,\rangle_p$ induced by Riemannian metric. This gives better approximation for projection in general. Two inner products are distinct for P_n manifolds.

$$\begin{aligned}\langle W_x, X\rangle &= \mathrm{tr}(W_x^T X), \text{ Euclidean metric} \\ \langle W_x, X\rangle_P &= \mathrm{tr}(P^{-1/2} W_x P^{-1} X P^{-1/2}), \text{ Riemannian metric}\end{aligned} \tag{7}$$

5.1 Log-Euclidean framework for CCA

Without loss of generality, we assume that data $X_i^{\wr} := \mathrm{Log}(\mu_x, X_i)$, $Y_i^{\wr} := \mathrm{Log}(\mu_y, Y_i)$ are centered in each tangent space, since in vector space, projection on subspace after centering is identical to centering after projection. We will show that Log-Euclidean framework for CCA is equivalent to defining covariance matrix between two random variables on manifolds. First, we observe that in Log-Euclidean CCA, projection onto each axis is done by Euclidean inner product like the following.

$$\pi_{W_x}(X^{\wr}) \approx \langle W_x, X^{\wr}\rangle / \|W_x\| \propto \langle W_x, X^{\wr}\rangle = \mathrm{tr}(W_x X^{\wr}) \tag{8}$$

The following derivation shows the equivalence.

$$\begin{aligned}
\rho_{x,y} &= \max_{W_x,W_y} \mathrm{corr}(\pi_{W_x}(X), \pi_{W_y}(Y)) \\
&= \max_{W_x,W_y} \frac{\sum_{i=1}^N \langle W_x, X_i^{\wr}\rangle \langle W_y, Y_i^{\wr}\rangle}{\sqrt{\sum_{i=1}^N \langle W_x, X_i^{\wr}\rangle^2}\sqrt{\sum_{i=1}^N \langle W_y, Y_i^{\wr}\rangle^2}} \\
&= \max_{W_x,W_y} \frac{\sum_{i=1}^N \langle W_x, \mathrm{Log}(\mu_x, X_i)\rangle \langle W_y, \mathrm{Log}(\mu_y, Y_i)\rangle}{\sqrt{\sum_{i=1}^N \langle W_x, \mathrm{Log}(\mu_x, X_i)\rangle^2}\sqrt{\sum_{i=1}^N \langle W_y, \mathrm{Log}(\mu_y, Y_i)\rangle^2}} \\
&= \max_{W_x,W_y} \frac{W_x^T \mathrm{COV}(X,Y) W_y}{\sqrt{W_x^T \mathrm{COV}(X,X) W_x}\sqrt{W_y^T \mathrm{COV}(Y,Y) W_y}}
\end{aligned} \tag{9}$$

where $\mathrm{COV}(X,Y) = \mathbb{E}[\mathrm{Log}(\mu_x, X)\mathrm{Log}(\mu_y, Y)^T]$. $\mathrm{COV}(X,X)$, and $\mathrm{COV}(Y,Y)$ are defined analogously [3]. Therefore, to perform CCA in Log-Euclidean framework is equivalent to run classical CCA based on the covariance matrices above. The Log-Euclidean framework for CCA is summarized in Alg 2.

Algorithm 2 Log-Euclidean CCA

```
1: Calculate intrinsic mean of each space, $\mu_x \in M_1$, $\mu_y \in M_2$
2: Map points onto each tangent space $X^{\wr} = \mathrm{Log}(\mu_x, X)$, $Y^{\wr} = \mathrm{Log}(\mu_y, Y)$
3: Perform CCA between $X^{\wr}$ and $Y^{\wr}$
```

Now, we discuss algebraically about the group action framework for Riemannian CCA in algorithm 2 in our main paper. In group action framework, the projection coefficient is approximated by

$$\pi_{W_x}(X^{\wr}) \approx \langle W_x, X^{\wr}\rangle_{\mu_x} / \|W_x\|_{\mu_x} \propto \langle W_x, X^{\wr}\rangle_{\mu_x} = \mathrm{tr}(\mu_x^{-1} W_x \mu_x^{-1} X^{\wr}) \tag{10}$$

Recall that correlation is scale-invariant. Hence, the formulation of group action framework for CCA is given by

$$\begin{aligned}
\rho_{\boldsymbol{X},\boldsymbol{Y}} &= \max_{W_x,W_y} \mathrm{corr}(\pi_{W_x}(X), \pi_{W_y}(Y)) \\
&= \max_{W_x,W_y} \frac{\sum_{i=1}^N \langle W_x, X_i^{\wr}\rangle_{\mu_x} \langle W_y, Y_i^{\wr}\rangle_{\mu_y}}{\sqrt{\sum_{i=1}^N \langle W_x, X_i^{\wr}\rangle_{\mu_x}^2}\sqrt{\sum_{i=1}^N \langle W_y, Y_i^{\wr}\rangle_{\mu_y}^2}} \\
&= \max_{W_x,W_y} \frac{\sum_{i=1}^N \mathrm{tr}(\mu_x^{-1} W_x \mu_x^{-1} X_i^{\wr}) \mathrm{tr}(\mu_y^{-1} W_y \mu_y^{-1} Y_i^{\wr})}{\sqrt{\sum_{i=1}^N \mathrm{tr}(\mu_x^{-1} W_x \mu_x^{-1} X_i^{\wr})^2}\sqrt{\sum_{i=1}^N \mathrm{tr}(\mu_y^{-1} W_y \mu_y^{-1} Y_i^{\wr})^2}}
\end{aligned} \tag{11}$$

where $W_x \in T_{\mu_x}\mathcal{M}$ and $W_y \in T_{\mu_y}\mathcal{M}$. This optimization looks complicated. We observe that $\langle W_x, X\rangle_{\mu_x} = \langle W_a, A\rangle$, where $A_i = \mu_x^{-1}X_i^{\wr}$ and $W_a = \mu_x^{-1}W_x$. This substitution changes Riemannian metric $\langle , \rangle_\mu$ to Euclidean metric $\langle , \rangle$. Therefore once data is transformed by μ_x^{-1} and μ_y^{-1}, the optimization is equivalent to performing Euclidean CCA in transformed tangent spaces, $\mu_x^{-1}T_{\mu_x}\mathcal{M}_x$ and $\mu_y^{-1}T_{\mu_y}\mathcal{M}_y$. It is defined as $\mu_x^{-1}T_{\mu_x}\mathcal{M}_x := \{\mu_x^{-1}X^{\wr} | X^{\wr} \in T_{\mu_x}\mathcal{M}_x\}$.
Now CCA formulation is given by

$$\begin{aligned}\rho_{\boldsymbol{X},\boldsymbol{Y}} &= \rho_{\boldsymbol{A},\boldsymbol{B}} \\ &= \frac{\sum_{i=1}^N \operatorname{tr}(W_a A_i)\operatorname{tr}(W_b B_i)}{\sqrt{\sum_{i=1}^N \operatorname{tr}(W_a A_i)^2}\sqrt{\sum_{i=1}^N \operatorname{tr}(W_b B_i)^2}}\end{aligned} \tag{12}$$

where $A_i = \mu_x^{-1}X_i^{\wr}$, $B_i = \mu_y^{-1}Y_i^{\wr}$, $W_a = \mu_x^{-1}W_x$, and $W_b = \mu_y^{-1}W_y$. The result is equivalent to algorithm 2 in our main paper.

6 Differentiation of g for SPD

We discussed CCA algorithms with approximated projections. Due to closed form solutions to approximated projections, the proposed methods are single path algorithms. However with exact projection, iterative methods are needed in general. We discussed an iterative method in our main paper with derivative of g. More details on calculation of derivative of g are provided in this section.

6.1 First derivative of g for SPD

Now given P_n, the gradient of g with respect to t is obtained by the following proposition in [4].

Proposition 1. *Let $F(t)$ be a real matrix-valued function of the real variable t. We assume that, for all t in its domain, $F(t)$ is an invertible matrix which does not have eigenvalues on the closed negative real line. Then*

$$\frac{d}{dt}\operatorname{tr}[\log^2 F(t)] = 2\operatorname{tr}[\log F(t) F(t)^{-1}\frac{d}{dt}F(t)] \tag{13}$$

The derivation of $\frac{d}{dt_i}g(t_i, \boldsymbol{w}_x)$ is the following.

$$\begin{aligned}\frac{d}{dt_i}g(t_i, \boldsymbol{w}_x) &= \frac{d}{dt_i}\|\operatorname{Log}(\operatorname{Exp}(\mu_x, t_i W_x), X_i)\|^2 \\ &= \frac{d}{dt_i}\operatorname{tr}[\log^2(X_i^{-1}S(t_i))]\end{aligned} \tag{14}$$

where $S(t_i) = \operatorname{Exp}(\mu_x, t_i W_x) = \mu_x^{1/2}\exp^{t_i A}\mu_x^{1/2}$ and $A = \mu_x^{-1/2}W_x\mu_x^{-1/2}$.

In our formulation, $F(t) = X_i^{-1}S(t_i)$. Then we have $F(t)^{-1} = S(t_i)^{-1}X_i$ and $\frac{d}{dt}F(t) = X_i^{-1}\dot{S}(t_i)$. Hence derivative of g with respect to t_i is given by

$$\begin{aligned}\frac{d}{dt_i}g(t_i, \boldsymbol{w}_x) &= 2\mathrm{tr}[\log(X_i^{-1}S(t_i))S(t_i)^{-1}X_iX_i^{-1}\dot{S}(t_i)], \text{ according to proposition 1}\\ &= 2\mathrm{tr}[\log(X_i^{-1}S(t_i))S(t_i)^{-1}\dot{S}(t_i)]\end{aligned} \tag{15}$$

where $\dot{S}(t_i) = \mu_x^{1/2}A\exp^{t_iA}\mu_x^{1/2}$.

6.2 Numerical differentiation for the second derivative of *g*

Riemannian CCA with exact projection can be optimized by Algorithm 1 in our main paper. Observe that the objective function of augmented Lagrangian method, $\mathcal{L}_A$ has ∇g in (11) in the main paper. Tge gradient of $\mathcal{L}_A$ involves the second gradient of g. Precisely, $\frac{d^2}{dwdt}g$ and $\frac{d^2}{dt^2}g$. These can be estimated by a finite difference method,

$$f^{'}(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \tag{16}$$

Obviously, $\frac{d^2}{dt^2}g$ is obtained by numerical recipe above using the analytic first derivative $\frac{d}{dt}g$. For $\frac{d^2}{dwdt}g$, we used orthonormal basis in $T_{\mu_x}\mathcal{M}$ to approximate derivative. By definition of directional derivative, we have

$$\lim_{h\to 0}\sum_i^d\left(\frac{f(x+hu_i)-f(x)}{h}\right)u_i = \sum_i^d\langle\nabla_x f(x), u_i\rangle u_i = \nabla_x f(x) \tag{17}$$

where $x \in \mathcal{X}$, d is dimension of $\mathcal{X}$, and $\{u_i\}$ is orthonormal basis of $\mathcal{X}$. Hence, perturbation along orthonormal basis enables to approximate gradient. For example, on P_n manifolds, the orthonormal basis in arbitrary tangent space $T_p\mathcal{M}$ can be obtained by three steps: 1. Pick orthonormal basis $\{e_i\}$ of $R^{n(n+1)/2}$, 2. convert $\{e_i\}$ into n-by-n symmetric matrices $\{u_i\}$ in $T_I\mathcal{M}$, in short,$\{u_i\}$ =mat($\{e_i\}$), 3. Transform basis $\{u_i\}$ from $T_I\mathcal{M}$ to $T_p\mathcal{M}$.

7 Multimodal risk analysis

MR image acquisition and processing: All the MRI data was acquired on a GE 3.0 Tesla scanner Discovery MR750 MRI system with an 8-channel head coil and parallel imaging (ASSET). The DWI data was acquired using a diffusion-weighted, spin-echo, single-shot, echo planar imaging radio-frequency (RF) pulse sequence. The images were acquired with diffusion weighting in 40 non-collinear directions at $b = 1300s/mm^2$ in addition to 8 $b = 0$ (non diffusion weighted or T2-weighted) images. The cerebrum was covered using contiguous 2.5 mm thick axial slices, FOV = 24 cm, TR = 8000 ms, TE = 67.8 ms, matrix = 96 x 96, resulting in isotropic 2.5 mm^3 voxels. High order shimming was performed prior to the DTI acquisition to optimize the homogeneity of the magnetic field across the brain and to minimize EPI distortions. The brain region was extracted using the first $b = 0$ image as input to the brain extraction tool (BET), also part of the FSL.

Eddy current related distortion and head motion of each data set were corrected using FSL software package [5]. The b-vectors were rotated using the rotation component of the transformation matrices obtained from the correction process. Geometric distortion from the inhomogeneous magnetic field applied was corrected with the b=0 field map and PRELUDE (phase region expanding labeler for unwrapping discrete estimates) and FUGUE (FMRIBs utility for geometrically unwarping EPIs) from FSL. Twenty-nine subjects did not have field maps acquired during their imaging session. Because these participants did not differ on APOE4 genotype, sex, or age compared to the participants that had field map correction, they were included in order to enhance the final sample size. The diffusion tensors were then estimated from the corrected DWI data using non-linear least squares estimation using the Camino library [6].

Individual maps were registered to a population specific template constructed using diffusion tensor imaging toolKit (DTI-TK[1]), which is an optimized DTI spatial normalization and atlas construction tool that has been shown to perform superior registration compared to scalar based registration methods [7]. The template is constructed in an unbiased way that captures both the average diffusion features (e.g. diffusivities and anisotropy) and anatomical shape features (tract size) in the population. A subset of 80 diffusion tensor maps was used to create a common space template. All diffusion tensor maps were normalized to the template with first rigid followed by affine and then symmetric diffeomorphic transformations. The diffeomorphic coordinate deformations themselves are smooth and invertible, that is neuroanatomical neighbors remain neighbors under the mapping. At the same time, the algorithms used to create these deformations are *symmetric* in that they are not biased towards the reference space chosen to compute the mappings. Moreover, these topology-preserving maps capture the large deformation necessary to aggregate populations of images in a common space. The spatially normalized data was interpolated to 2 mm × 2 mm × 2 mm voxels for the final CCA analysis.

Along with the DWI data T1-weighted images were acquired using BRAVO pulse sequence which uses 3D inversion recovery (IR) prepared fast spoiled gradient recalled echo (FSPGR) acquisition to produce isotropic images at 1 mm × 1 mm × 1mm resolution. We extract the brain regions again using BET. We compute an optimal template space i.e. a population-specific, unbiased average shape and appearance image derived from our population [8]. We use the openly available advanced normalization tools (ANTS[2]) to develop our template space and also perform the registration of the individual subjects to that space [9]. ANTS encodes current best practice in image registration, optimal template

[1] http://dti-tk.sourceforge.net/pmwiki/pmwiki.php

[2] http://www.picsl.upenn.edu/ANTS/

construction and segmentation [10]. The coordinate deformations in this case are also symmetric and diffeomorphic Once we perform the registrations we extract the Jacobian matrices per voxel per subject from these deformation fields for performing CCA analysis.

Additional experiments: We examine the following multimodal linear relations (as we examined in the main paper) but by performing CCA on the entire gray and white matter voxels. The gray matter region was defined as follows. First we perform a three tissue segmentation of each of the spatially normalized T1-weighted images into gray, white and cerebral spinal fluid using FAST segmentation algorithm ([11]) implemented in FSL. Then we take the average of the gray matter probabilities using all the subjects and threshold it to obtain the final binary mask. The white matter region is simply defined as the region with fractional anisotropy (FA) obtained from the diffusion tensors > 0.2.

$$\begin{aligned} Y_{\mathrm{DTI}} &= \beta_0 + \beta_1 \mathrm{Gender} + \beta_2 X_{\mathrm{T1W}} + \beta_3 X_{\mathrm{T1W}} \cdot \mathrm{Gender} + \varepsilon, \\ Y_{\mathrm{DTI}} &= \beta_0' + \beta_1' \mathrm{AgeGroup} + \beta_2' X_{\mathrm{T1W}} + \beta_3' X_{\mathrm{T1W}} \cdot \mathrm{AgeGroup} + \varepsilon, \end{aligned}$$

The statistical tests would be to see if we can reject the null-hypotheses $\beta_3 = 0$ and $\beta_3' = 0$ using our data at $\alpha = 0.05$. In addition to the above linear relationships CCA can also facilitate testing the following relationships by using the weight-vectors as regions of interest.

$$\begin{aligned} \overline{MD} &= \beta_0 + \beta_1 \delta_{\mathrm{Female}} + \beta_2 \delta_{\mathrm{Male}} + \varepsilon, \\ \log |J| &= \beta_0' + \beta_1' \delta_{\mathrm{Female}} + \beta_2' \delta_{\mathrm{Male}} + \varepsilon, \end{aligned}$$

where the null-hypotheses to be tested are $\beta_1 = \beta_2$ and $\beta_1' = \beta_2'$. These models can also be tested to find statistically significant AgeGroup differences similarly.

Below we show the gender and age distributions of the sample in Fig. 1. In presenting the CCA results, we first show the montages of all the slices of the brain overlaid by the weight vectors obtained by performing our CCA from

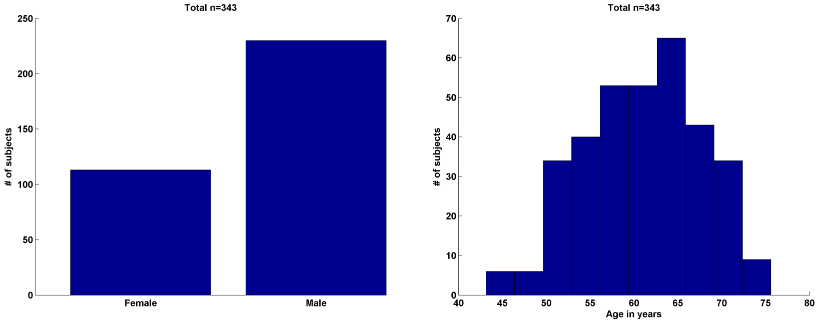

Fig. 1: The sample characteristics in terms of gender and age distributions.

Figs. 2 to 7. The underlays for T1W results are the slices from T1W population specific template created using ANTS. Similarly the underlays for the DTI results are the FA maps of the population specific template created using DTI-TK (please see the image processing section above for additional details). We can observe that the sparsity constraint actually is forcing many voxels to have zero weights but an interesting observation is that the voxels with non-zeros weights (highlighted in red-boxes) are spatially complimentary in DTI and T1W. Even more interestingly our CCA finds the cingulum regions in the white matter for the DTI modality. In our experiments we observe the same regions for various settings of the sparsity cost parameter. Similarly, Figs. 11 to 13 show the results using the Euclidean CCA performed using the full tensor information [12]. We can observe that the results are similar to those in Figs. 2 to 7 but the regions are thinner in the Euclidean version. Interpreting subtle differences between the Euclidean CCA with vectorized tensors and the Riemannian CCA with the full tensors is a complicated task because these differences can be amplified or subdued depending on the sparsity cost parameters and pre-processing of the images [13]. However when deciding from the *first principles* it is always recommended to be as faithful to the true mathematical nature of the data as possible when performing population analysis of the MRI data.

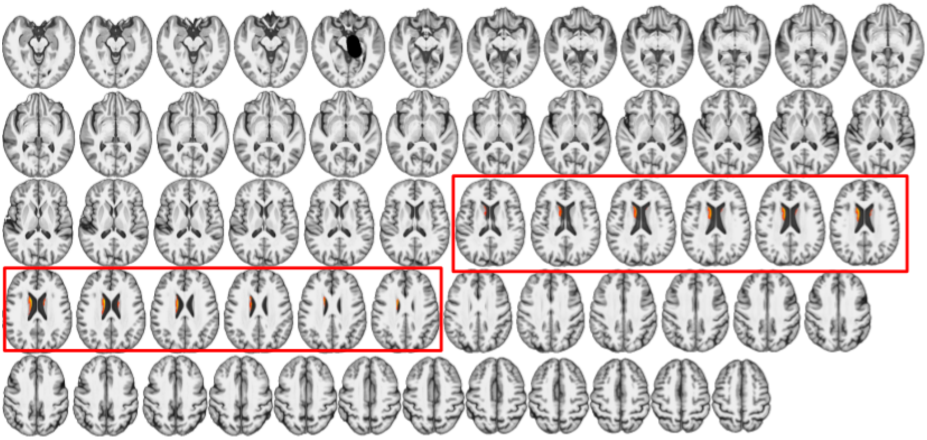

Fig. 2: Weight vector (obtained by Riemannian (tensor) CCA) visualization of T1W **axial** slices.

The scatter and bar plots for the testing linear relationships using both Riemannian and Euclidean CCA are shown in Figs. 14 and 15.

We have presented in this supplement that CCA when performed using the intrinsic properties of the MRI data can reveal biologically meaningful patterns without any *a priori* biological input to the model. We showed that one can perform various types of multi-modal hypothesis testing of linear relationships using the projection vectors from the CCA. We can even envision discriminant

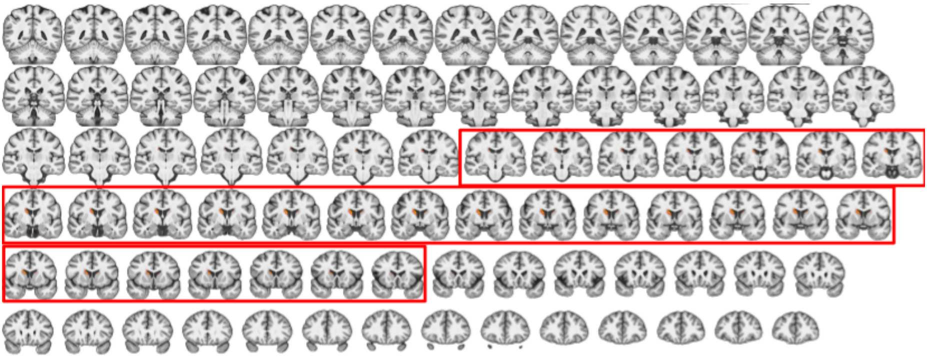

Fig. 3: Weight vector (obtained by Riemannian (tensor) CCA) visualization of T1W **coronal** slices.

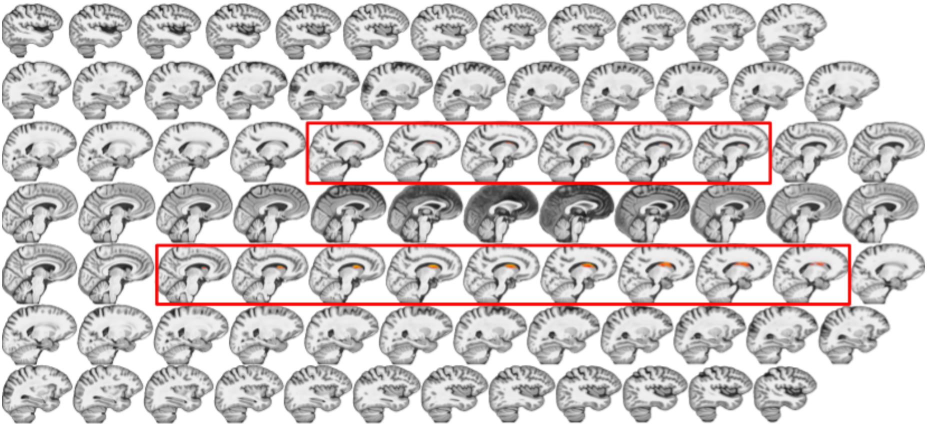

Fig. 4: Weight vector (obtained by Riemannian (tensor) CCA) visualization of T1W **sagittal** slices.

analysis (predicting gender and age group using the multi-modal brain data) using the CCA projection vectors. CCA can be extended beyond multi-modal imaging data, where one can try to directly maximize the correlation between imaging and non-imaging data using a cross-validation technique [14]. Our Riemannian CCA can provide a fitting extension to such investigations. Finally we would like to note that the distinctions between the Euclidean CCA with vectorization [12] and the proposed Riemannian CCA can become truly significant when the entire processing pipeline for the MRI data is faithful to the intrinsic properties of the MRI data and noise distributions. For example, currently the non-linear least squares tensor estimation in Camino might perform algebraic corrections (i.e. thresholding the eigen values to be positive) to make the tensors positive definite rather than performing a geometrically constrained estimation [13]. And our experimental evidence presented both in the main paper and this

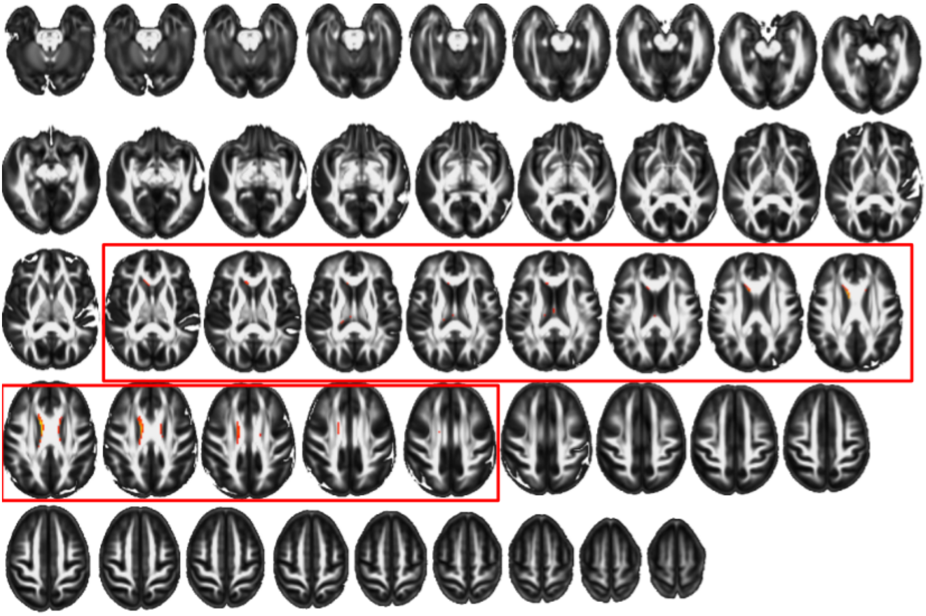

Fig. 5: Weight vector (obtained by Riemannian (tensor) CCA) visualization of DTI **axial** slices.

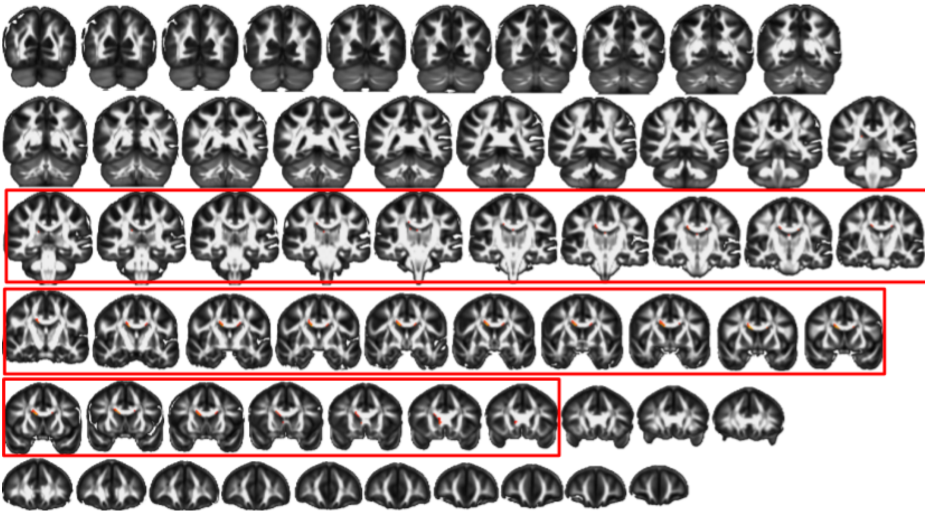

Fig. 6: Weight vector (obtained by Riemannian (tensor) CCA) visualization of DTI **coronal** slices.

supplement making a convincing case that manifold based analysis produces biologically meaningful results.

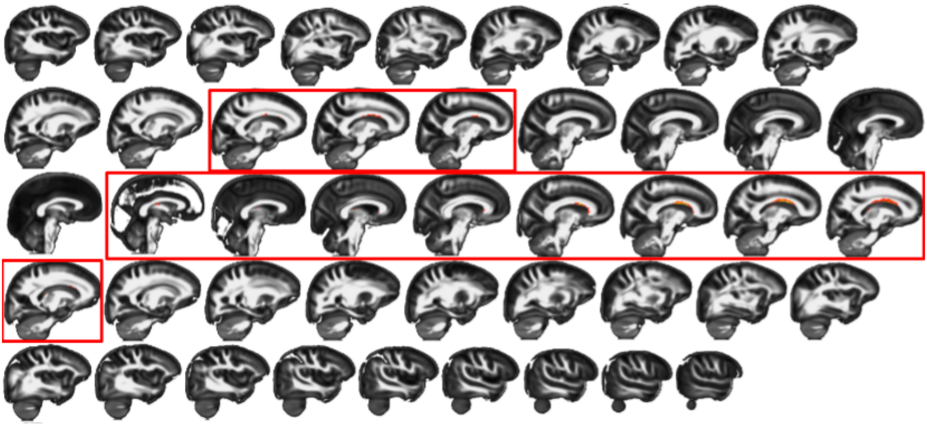

Fig. 7: Weight vector (obtained by Riemannian (tensor) CCA) visualization of DTI **sagittal** slices.

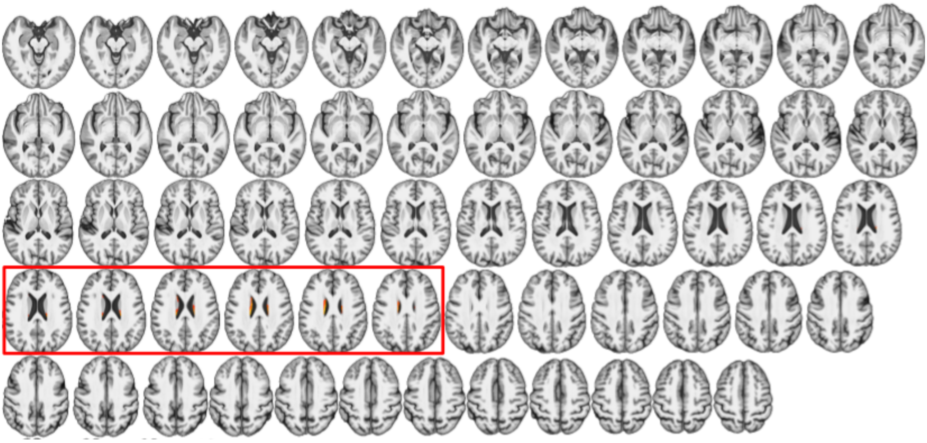

Fig. 8: Weight vector (obtained by Euclidean CCA (tensors)) visualization of T1W **axial** slices.

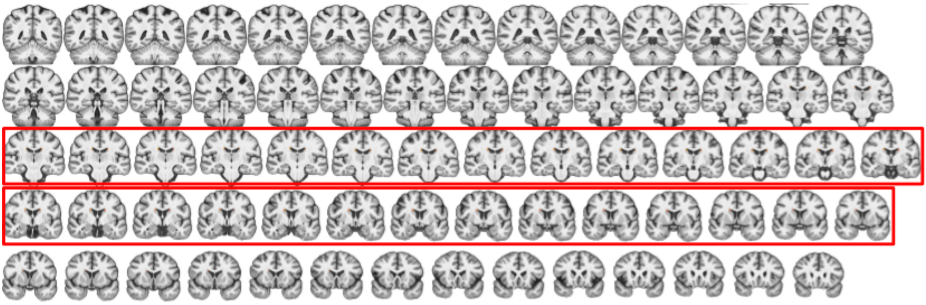

Fig. 9: Weight vector (obtained by Euclidean CCA (tensors)) visualization of T1W **coronal** slices.

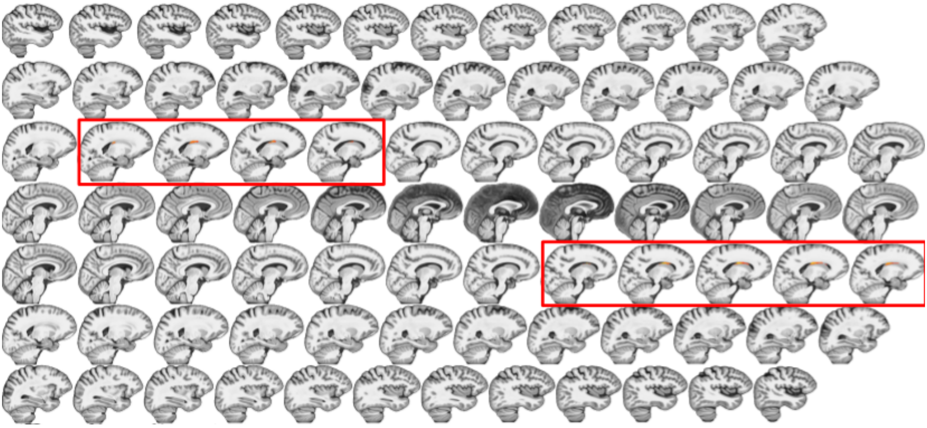

Fig. 10: Weight vector (obtained by Euclidean CCA (tensors)) visualization of T1W **sagittal** slices.

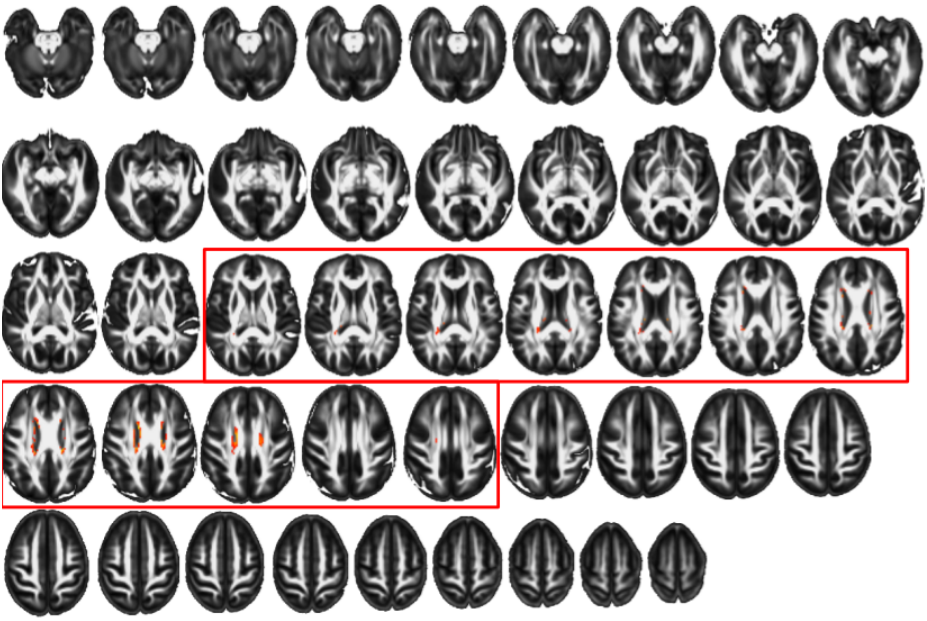

Fig. 11: Weight vector (obtained by Euclidean CCA (tensors)) visualization of DTI **axial** slices.

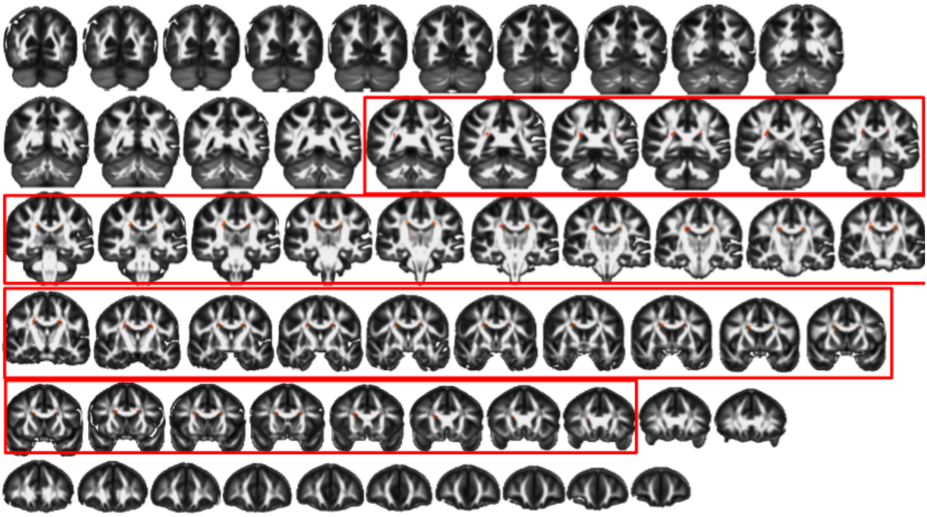

Fig. 12: Weight vector (obtained by Euclidean CCA (tensors)) visualization of DTI **coronal** slices.

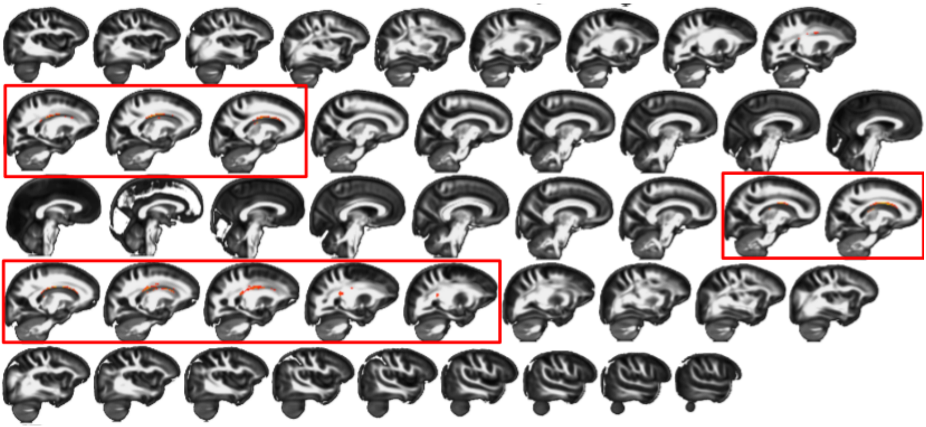

Fig. 13: Weight vector (obtained by Euclidean CCA (tensors)) visualization of DTI **sagittal** slices.

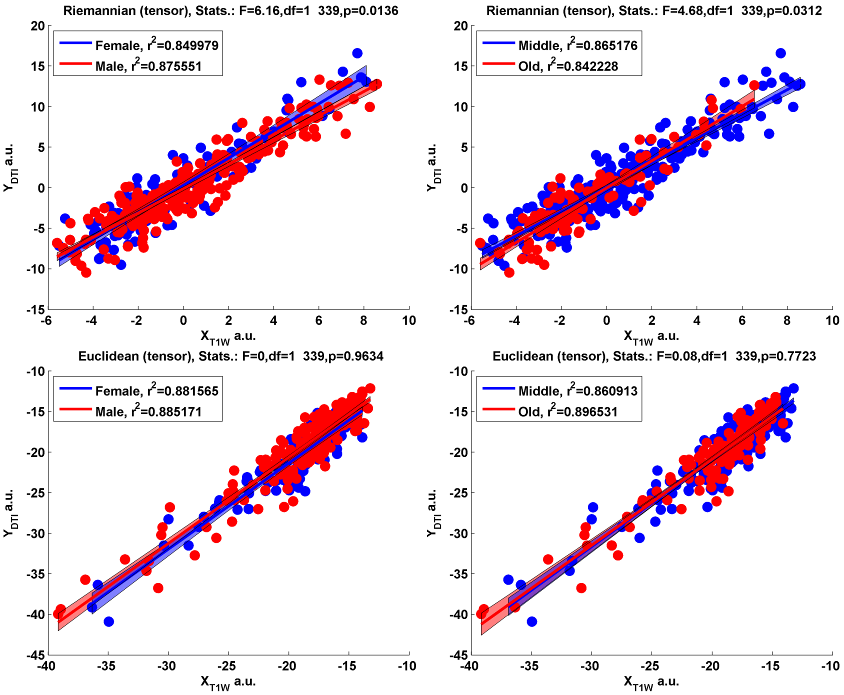

Fig. 14: CCA projections revealing statistically significant volume and diffusivity interactions with gender (left) and age-group (right). Top row shows the results using Riemannian CCA and the bottom row using the Euclidean CCA with vectorization. We can clearly see the improvement the statistical confidence (smaller p-values) for rejecting the null-hypotheses when using Riemannian CCA.

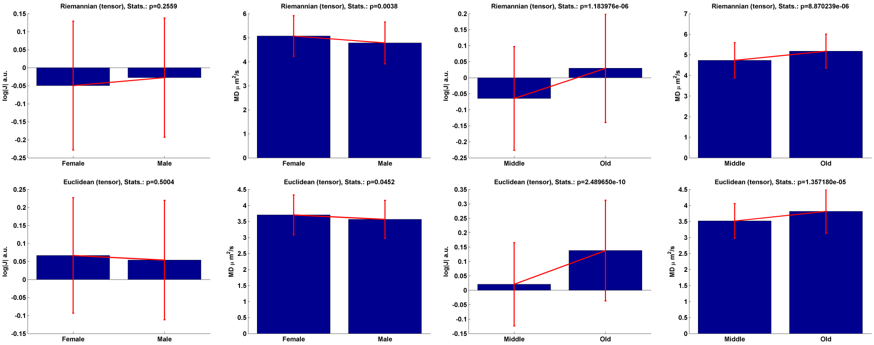

Fig. 15: Effect of gender and age-group on the mean diffusivity (MD) and the log Jacobian determinant. Top row shows the results using Riemannian CCA and the bottom row using Euclidean CCA with vectorization. Again our approach produces smaller or similar p-values in rejecting the null-hypotheses.

References

1. Ferreira, R., Xavier, J., Costeira, J.P., Barroso, V.: Newton method for riemannian centroid computation in naturally reductive homogeneous spaces. (2006)
2. Xie, Y., Vemuri, B.C., Ho, J.: Statistical analysis of tensor fields. MICCAI (2010) 682–689
3. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. Medical Imaging **23**(8) (2004) 995–1005
4. Moakher, M.: A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM Journal on Matrix Analysis and Applications **26**(3) (2005) 735–747
5. Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., Behrens, T., Johansen-Berg, H., Bannister, P., De Luca, M., Drobnjak, I., Flitney, D., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J., Matthews, P.: Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage **23** (2004) 208–219
6. Cook, P., Bai, Y., Nedjati-Gilani, S., Seunarine, K.K., Hall, M.G., Parker, G.J., Alexander, D.C.: Camino: Open-source diffusion-MRI reconstruction and processing. ISMRM (2006) 2759
7. Zhang, H., Yushkevich, P., Alexander, D., Gee, J.: Deformable registration of diffusion tensor MR images with explicit orientation optimization. Med. Img. Analysis **10** (2006) 764–785
8. Avants, B., Gee, J.: Geodesic estimation for large deformation anatomical shape and intensity averaging. NeuroImage (2004) S139–150
9. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis **12** (2008) 26–41
10. Klein, A., Andersson, J., Ardekani, B., Ashburner, J., Avants, B., Chiang, M., Christensen, G., Collins, L., Hellier, P., Song, J., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R., Mann, J., Parsey, R.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage (2009)
11. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Img. **20**(1) (2001) 45–57
12. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3) (2009) 515–534
13. Koay, C., Chang, L., Carew, J., Pierpaoli, C., Basser, P.: A unifying theoretical and algorithmic framework for least squares methods of estimation in diffusion tensor imaging. Journal of Magnetic Resonance **182** (2006) 115–125
14. Avants, B., Libonc, D., Rascovsky, K., Boller, A., McMillan, C., Massimo, L., Coslett, H., Chatterjee, A., Gross, R., Grossman, M.: Sparse canonical correlation analysis relates network-level atrophy to multivariate cognitive measures in a neurodegenerative population. NeuroImage **84**(1) (2014) 698–711