GEO-NN: AN END-TO-END FRAMEWORK FOR GEODESIC MEAN ESTIMATION ON THE MANIFOLD OF SYMMETRIC POSITIVE DEFINITE MATRICES

Anonymous authors
Paper under double-blind review

ABSTRACT

The manifold of symmetric positive definite (SPD) matrices plays a key role in many domains, from network science to differential geometry to signal and image processing. However, leveraging the SPD manifold geometry during inference is challenging, as simple operations, such as mean estimation, do not have a closed-form or easily computable solution. In this paper, we propose an end-to-end deep learning framework, which we call a Geometric Neural Network (Geo-NN), to efficiently compute the geodesic mean of a collection of matrices lying on the SPD manifold. Geo-NN utilizes a Matrix-Autoencoder (MAE) architecture with intersecting fully connected layers as its backbone. We illustrate that the matrix-normal equation arising from Fréchet mean estimation can be converted into a loss function for optimizing the Geo-NN, which in turn approximates the geodesic mean of a collection of SPD matrices. We demonstrate the efficacy of our framework in both synthetic and real-world scenarios, as compared to commonly used alternative methods. Our simulated experiments demonstrate that Geo-NN is robust to various noise conditions and is scalable to increasing dataset size and dimensionality. Our real-world application of Geo-NN to functional connectomics data allows us to extract network patterns associated with patient/control differences.

1 INTRODUCTION

Symmetric Positive Definite (SPD) matrices are ubiquitous across many areas of data science. For example, they arise as covariance matrices in statistical signal processing Fuhrmann & Miller (1988), elemental computational objects in convex and semidefinite programming Fletcher (1985), adjacency matrices in network science, and kernels in machine learning and graph theory Dodero et al. (2015). Thus, the study of SPD manifolds has immense real-world utility, particularly in medical informatics. Applications include statistical shape analysis Pérez & González-Farias (2013), designing Brain Computer Interfaces (BCI) Barachant et al. (2010), or the analysis of diffusion tensors Wang & Vemuri (2005) and functional connectomics D'Souza et al. (2021) data. In all these cases, one of the key goals is to predict an outcome of interest based on input data lying on an SPD manifold. Methods in this inferential setting include geometry-aware principal component analysis Horev et al. (2016), Reimannian machine learning Yger et al. (2012); Banerjee et al. (2015), and deep learning architectures Nguyen et al. (2019); Nguyen (2021) and optimization techniques Brooks et al. (2019) geared towards the SPD manifold. Such methods rely on accurate and robust geometric mean estimation on the SPD manifold as an integral intermediate step within their algorithmic implementations.

While the geometric properties underlying the SPD manifold have been studied through multiple lenses, from matrix analyis to operator theory to differential geometry, efficient mean estimation on the SPD manifold remains far more challenging than statistical estimation in Euclidean data spaces. This is because extensions of elementary operations such as addition, subtraction, and distances on the SPD manifold entail significant computational overhead Moakher (2005); Moakher & Batchelor (2006). Among different mean definitions, the Fisher Information distance metric and corresponding geodesic mean Georgiou (2007) are the most sought after due its desirable properties, such as congruence invariance, determinant identity, and self duality Bhatia (2009). Unlike Euclidean spaces, the geodesic mean on the SPD manifold does not have a closed-form solution and is often computed via iterative optimization algorithms Jeuris et al. (2012); Poloni (2009).

1.1 RELATED WORK

The most common approach for estimating the geodesic mean on the SPD manifold is via gradient descent Pennec et al. (2006). While this method relies on first-order information, and thus, requires moderate computational overhead, it does not provide guaranteed convergence from arbitrary stationary points and is highly sensitive to step size selection. To mitigate this issue, Armijo step-size adaptions or second order optimization techniques such as conjugate gradient descent, trust-region optimization, and BFGS methods may be employed Afsari et al. (2013). Other alternatives include the majorization-maximization algorithm Zhang (2013), Riemannian optimization methods Jeuris (2015), and fixed-point iterations Congedo et al. (2017). While these extensions have desirable convergence properties, they significantly increase the computational complexity per iteration and do not scale well to higher input dimensionality and larger numbers of samples Congedo et al. (2015).

In contrast to gradient methods, the work of Congedo et al. (2015) leverages the approximate joint diagonalization Pham (2001) of matrices on the SPD manifold. This representation arises from the Common Principal Components (CPC) formulation Jolliffe & Cadima (2016). Leveraging the invariance properties of the geodesic mean and its interplay with the CPC objective, the authors propose an iterative estimation procedure using the CPC solution as an initialization. While the approximate joint diagonalization method provides guaranteed convergence to a fixed point, the accuracy and stability of the optimization is sensitive to the deviation of the generating process from the CPC model in practice. In the presence of severe deviations, the algorithm often diverges.

1.2 OUR CONTRIBUTIONS

We propose a novel end-to-end framework to estimate the geodesic mean of data on the SPD manifold. Our method, the Geometric Neural Network (Geo-NN), leverages a matrix autoencoder formulation D'Souza et al. (2021) that performs a series of bi-linear transformations on the input SPD matrices. This strategy ensure that the estimated mean remains on the manifold at each iteration. Our loss function for training the network is designed to approximate a first order matrix-normal condition arising from Fréchet mean estimation Moakher (2005). Using conventional backpropagation via stochastic iterative optimization, the Geo-NN automatically learns to estimate the geodesic mean of the input data. We demonstrate the robustness of our framework using simulation studies and show that Geo-NN can handle both input noise and variations in the data generating process better than current iterative methods. In addition, Geo-NN can be applied to high-dimensional data, a notable bottleneck for the second-order methods described above. Finally, we examine the applicability of the Geo-NN to a real-world functional connectomics study and discover consistent group differences between patients diagnosed with ADHD-Autism comorbidities and healthy controls. Importantly, Geo-NN minimal assumptions about the data and can be easily adapted other domains, both as a standalone estimation module or as a part of a larger deep learning architecture.

2 GEO-NN: GEOMETRIC MEAN ESTIMATION ON THE SPD MANIFOLD

2.1 PRELIMINARIES

Let matrices $\{\mathbf{\Gamma}_n\}_{n=1}^N \in \mathcal{M}$ be a collection of N datapoints belong to the manifold $\mathcal{M}$ of Symmetric Positive Definite (SPD) matrices of dimensionality $P \times P.$, i.e. $\mathcal{M} \in \mathcal{P}_P^+$. By definition, each $\mathbf{\Gamma}_n$ has P strictly positive eigenvalues and is a point on a convex cone in $P(P+1)/2$ dimensions. Additionally, $\mathcal{M}$ is a real and smooth Reimannian manifold that is locally similar in geometry to Euclidean space. More precisely, $\mathcal{M}$ is equipped with an inner product that varies smoothly at each vector $\mathcal{T}_{\mathbf{\Gamma}}(\mathcal{M})$ in the tangent space defined at any point $\mathbf{\Gamma} \in \mathcal{M}$. Finally, a *geodesic* denotes a path joining any two points on the manifold by following the manifold surface.

2.1.1 ELEMENTARY OPERATIONS ON THE SPD MANIFOLD

Geodesic Mappings: The matrix exponential and the matrix logarithm maps allow us to translate geodesics on the manifold back and forth to the local tangent space at a reference point.

The matrix exponential mapping translates a vector $\mathbf{V} \in \mathcal{T}_{\mathbf{\Phi}}(\mathcal{M})$ in the tangent space at $\mathbf{\Phi} \in \mathcal{M}$ to a point on the manifold $\mathbf{\Gamma} \in \mathcal{M}$ via the geodesic emanating from $\mathbf{\Phi}$. Mathematically, the exponential

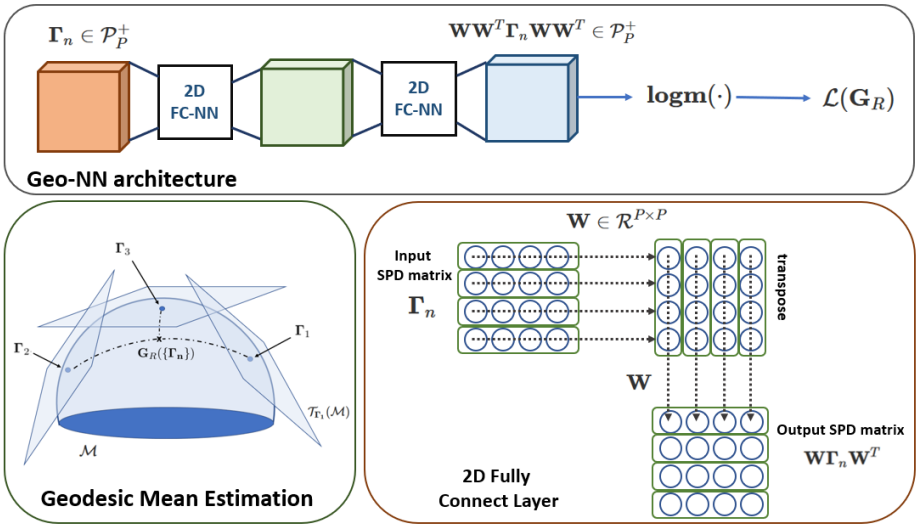

Figure 1: **The Geo-NN architecture:** The input is transformed by a cascade of 2D fully connected layers Dong et al. (2017); Huang & Van Gool (2017). The matrix logarithm function to obtain the matrix normal form in Eq. (2.1), which serves as the loss function for Geo-NN during training.

map is parameterized as follows:

$$\mathbf{\Gamma} = \mathbf{Expm}_{\mathbf{\Phi}}(\mathbf{V}) = \mathbf{\Phi}^{1/2}\mathbf{expm}(\mathbf{\Phi}^{-1/2}\mathbf{V}\mathbf{\Phi}^{-1/2})\mathbf{\Phi}^{1/2} \tag{1}$$

Conversely, the matrix logarithm map translates the geodesic between $\mathbf{\Phi} \in \mathcal{M}$ to $\mathbf{\Gamma} \in \mathcal{M}$ back to the tangent vector $\mathbf{V} \in \mathcal{T}_{\mathbf{\Phi}}(\mathcal{M})$ and is expressed in closed form as:

$$\mathbf{V} = \mathbf{Logm}_{\mathbf{\Phi}}(\mathbf{\Gamma}) = \mathbf{\Phi}^{1/2}\mathbf{logm}(\mathbf{\Phi}^{-1/2}\mathbf{\Gamma}\mathbf{\Phi}^{-1/2})\mathbf{\Phi}^{1/2} \tag{2}$$

Here, $\mathbf{expm}(\cdot)$ and $\mathbf{logm}(\cdot)$ refer to the matrix exponential and logarithm respectively. Computationally, each of these operations require an eigenvalue decomposition of the argument matrix, a transformation applied point-wise to the eigenvalues, and a matrix reconstruction.

Distance Metric: Given two points on the manifold $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \in \mathcal{M}$, the Fisher Information distance between them is the length of the geodesic connecting the two points, given by:

$$\delta_R(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2) = ||\mathbf{logm}(\mathbf{\Gamma}_1^{-1}\mathbf{\Gamma}_2)||_F = ||\mathbf{logm}(\mathbf{\Gamma}_2^{-1}\mathbf{\Gamma}_1)||_F, \tag{3}$$

where $||\cdot||_F$ denotes the Frobenius norm. Similarly, the Reimannian norm of a matrix $\mathbf{\Gamma}$ is defined as the geodesic distance from the identity matrix $\mathcal{I}$ i.e. $||\mathbf{\Gamma}||_R = ||\mathbf{logm}(\mathbf{\Gamma})||_F$

2.2 Geodesic Mean Estimation via the Geo-NN

The geodesic mean of a collection of SPD matrices $\{\mathbf{\Gamma}_n\} \in \mathcal{M}$ is defined as the matrix $\mathbf{G}_R \in \mathcal{M}$ whose sum of squared geodesic distances (Eq. (3)) to each element of the collection is minimal Ando et al. (2004). A pictorial illustration is provided in the green box in Fig 1. Mathematically, this condition can be expressed as follows:

$$\mathcal{G}_R(\{\mathbf{\Gamma}_n\}) = \arg\min_{\mathbf{G}_R} \mathbf{L}(\mathbf{G}_R) = \arg\min_{\mathbf{G}_R} \sum_n \delta_R^2(\mathbf{G}_R, \mathbf{\Gamma}_n) = \arg\min_{\mathbf{G}_R} \sum_n ||\mathbf{logm}(\mathbf{G}_R^{-1}\mathbf{\Gamma}_n)||_F^2 \tag{4}$$

Eq. (4) is known to have a closed form solution for $N = 2$, but not for larger collections. Eq. (4) is convex and smooth with respect to the unknown quantity $\mathbf{G}_R(\cdot)$ Moakher (2005). The point at which the gradient of Eq. (4) vanishes is thus the unique minima for mean estimation. Making use of this observation, the following theorem from Moakher (2005) can be stated.

Theorem 2.1. *The geometric mean $\mathbf{G}_R$ of a collection of N SPD matrices $\{\mathbf{\Gamma}_n\}$ is the unique symmetric positive-definite solution to the nonlinear matrix equation*

$$\sum_n \mathbf{logm}(\mathbf{G}_R^{-1/2}\mathbf{\Gamma}_n\mathbf{G}_R^{-1/2}) = \mathbf{0} \tag{5}$$

where the symbol $\mathbf{0}$ denotes a $P \times P$ matrix of all zeros.

Proof: The proof follows by computing the first order necessary (and here, sufficient) condition for optimality for Eq. (4). First, we express the derivative of a real-valued function of the form $\mathbf{H}(\mathbf{S}(t)) = \frac{1}{2}||\mathbf{logm}(\mathbf{C}^{-1}\mathbf{S}(t))||_F^2$ with respect to t. In this expression, the argument $\mathbf{S}(t) = {\mathbf{G_R}}^{1/2}\mathbf{expm}(t\mathbf{A}){\mathbf{G_R}}^{1/2}$ is the geodesic arising from $\mathbf{G}_R$ in the direction of $\mathbf{\Delta} = \dot{\mathbf{S}}(\mathbf{0}) = {\mathbf{G_R}}^{1/2}\mathbf{A}{\mathbf{G_R}}^{1/2}$, and the matrix $\mathbf{C} \in \mathcal{P}_P^+$ is a constant SPD matrix of dimension P.

By using the cyclic properties of the trace function and the distributive equivalence of $\mathbf{logm}(\mathbf{A}^{-1}[\mathbf{B}]\mathbf{A}) = \mathbf{A}^{-1}[\mathbf{logm}(\mathbf{B})]\mathbf{A}$, we obtain the following condition:

$$\mathbf{H}(\mathbf{S}(t)) = \frac{1}{2}||\mathbf{logm}(\mathbf{C}^{-1/2}\mathbf{S}(t)\mathbf{C}^{-1/2})||_F^2$$

By the symmetry of the term $\mathbf{logm}(\mathbf{C}^{-1/2}\mathbf{S}(t)\mathbf{C}^{-1/2})$ we have that:

$$\therefore \frac{d}{dt}\mathbf{H}(\mathbf{S}(t))\Big|_{t=0} = \frac{1}{2}\frac{d}{dt}\mathrm{Tr}\left([\mathbf{logm}(\mathbf{C}^{-1/2}\mathbf{S}(t)\mathbf{C}^{-1/2})]^2\right)\Big|_{t=0}$$

$$\therefore \frac{d}{dt}\mathbf{H}(\mathbf{S}(t))\Big|_{t=0} = \mathrm{Tr}\left([\mathbf{logm}(\mathbf{C}^{-1}\mathbf{G}_R)\mathbf{G}_R^{-1}\mathbf{\Delta}]\right) = \mathrm{Tr}[\mathbf{\Delta}\mathbf{logm}(\mathbf{C}^{-1}\mathbf{G}_R)\mathbf{G}_R^{-1}]$$

$$\therefore \nabla\mathbf{H} = \mathbf{logm}(\mathbf{C}^{-1}\mathbf{G}_R)\mathbf{G}_R^{-1} = \mathbf{G}_R^{-1}\mathbf{logm}(\mathbf{G}_R\mathbf{C}^{-1})$$

Notice that since $\nabla\mathbf{H}$ is symmetric, it indeed belongs to the tangent space $\mathcal{S}_P$ of $\mathcal{P}_P^+$. For $\mathbf{L}(\mathbf{G}_R)$ defined in Eq. (4), we can correspondingly express the gradient as follows:

$$\mathbf{L}(\mathbf{G}_R) = \sum_n ||\mathbf{logm}(\mathbf{G}_R^{-1}\mathbf{\Gamma}_n)||_F^2 \quad \implies \nabla\mathbf{L}(\mathbf{G}_R) = \mathbf{G}_R^{-1}\sum_n \mathbf{logm}(\mathbf{G}_R\mathbf{\Gamma}_n^{-1})$$

Since $\mathbf{L}(\mathbf{G}_R)$ is a sum of convex functions, the first order stationary point is the necessary and sufficient condition for $\mathbf{G}_R$ being the unique minima.

$$\therefore \arg\min_{\mathbf{G}_R} \mathbf{L}(\mathbf{G}_R) \implies \sum_n \mathbf{logm}(\mathbf{G}_R\mathbf{\Gamma}_n^{-1}) = \sum_n \mathbf{logm}(\mathbf{G}_R^{-1/2}\mathbf{\Gamma}_n\mathbf{G}_R^{-1/2}) = \mathbf{0}$$

Denoting $\mathbf{G}_R^{-1/2} = \mathbf{V} \in \mathcal{P}_P^+$, the matrix multiplications in the argument of the $\mathbf{logm}(\cdot)$ term can be efficiently expressed within the feed-forward operations of a neural network with unknown parameters $\mathbf{V}$. Correspondingly, the Geo-NN architecture uses the form of the matrix normal equation in Theorem 2.1 to perform the estimation of the geometric mean.

2.3 GEO-NN ARCHITECTURE

The Geo-NN is a matrix autoencoder with tied weights, as illustrated in Fig. 1. The encoder of the Geo-NN is a 2D fully-connected neural network (FC-NN) Dong et al. (2017); Huang & Van Gool (2017) layer $\mathbf{\Psi}_{\text{enc}}(\cdot) : \mathcal{P}_P^+ \to \mathcal{P}_P^+$ that projects the input matrices $\mathbf{\Gamma}_n$ into a latent representation. This mapping is parameterized by weights $\mathbf{W} \in \mathcal{R}^{P\times P}$ and is computed as a cascade of two linear layers with tied weights, i.e., $\mathbf{\Psi}_{\text{enc}}(\mathbf{\Gamma}_n) = \mathbf{W}\mathbf{\Gamma}_n\mathbf{W}^T$ The decoder $\mathbf{\Psi}_{dec}(\cdot)$ has the same architecture as the encoder, but with transposed weights $\mathbf{W}^T$. The overall transformation can be written as:

$$\text{Geo-NN}(\mathbf{\Gamma}_n) = \mathbf{\Psi}_{\text{dec}}(\mathbf{\Psi}_{\text{enc}}(\mathbf{\Gamma}_n)) = \mathbf{W}\mathbf{W}^T(\mathbf{\Gamma}_n)\mathbf{W}\mathbf{W}^T = \mathbf{V}(\mathbf{\Gamma}_n)\mathbf{V} \tag{6}$$

where $\mathbf{V} \in \mathcal{R}^{P\times P}$ and is symmetric and positive definite by construction.

We would like our loss function to minimize Eq. (4) in order to estimate the first order stationary point as $\mathbf{V} = \mathbf{G}_R^{-1/2}$. [1] Therefore, we choose to utilize the following loss function:

$$\mathcal{L}(\cdot) = \frac{1}{P^2}\Big|\Big|\frac{1}{N}\sum_n \mathbf{logm}\Big[\mathbf{W}\mathbf{W}^T(\mathbf{\Gamma}_n)\mathbf{W}\mathbf{W}^T\Big]\Big|\Big|_F^2 \tag{7}$$

[1] Note that to ensure invertibility of $\mathbf{V}$, $\mathbf{W}$ should be full rank. In practice, we add a small bias to the weights for regularization, i.e., $\tilde{\mathbf{W}} = \mathbf{W} + \lambda\mathcal{I}_P$ and stability. Empirically, we verified the rank of $\tilde{\mathbf{W}}$ and of the $\mathbf{logm}(\cdot)$ arguments at each iteration during training and did not observe any degenerate behavior.

Formally, an error of $\mathcal{L}(\cdot) = 0$ implies that the norm argument goes to zero and therefore satisfies the matrix normal equation exactly under the parameterization $\mathbf{V} = \mathbf{W}\mathbf{W}^T = \mathbf{G}_R^{-1/2}$. Therefore, Eq. (7) is a suitable candidate for estimation of the geodesic mean on the SPD manifold.

Since the operations in Geo-NN can be implemented as differentiable modules, we utilize standard backpropagation to optimize Eq. (7). From an efficiency standpoint, the Geo-NN architecture maps onto a relatively shallow neural network. Therefore, this module can be easily integrated into other deep learning inference frameworks for example, for batch normalization on the SPD manifold. This flexibility is the key advantage over classical methods (see Section 1.1), in which integrating the geometric mean estimation within a larger framework is not straightforward. Finally, the extension of Eq. (7) to the estimation of a weighted mean (with positive weights $\{w_n\}$) also follows naturally.

Implementation Details: We train Geo-NN for a maximum of 200 epochs with an initial learning rate of 0.001 decayed by 0.8 every 50 epochs. The tolerance criteria for the training loss is set at $1e^{-4}$. Geo-NN implemented in PyTorch (v1.5.1), Python 3.5 and experiments were run on an 4.9 GB Nvidia K80 GPU. We utilize the stochastic ADAM Kingma & Ba (2014) optimizer during training and a default PyTorch initialization for the model weights LeCun et al. (2012). The computational complexity per iteration of Geo-NN scales as $\mathcal{O}(N \times P^3)$ with an extra matrix multiplication and square root operation to estimate the final geodesic mean $\mathbf{G}_R$.

3 Evaluation and Results

3.1 Experiments on Synthetic Data

We evaluate the scalability, robustness, and adaptability of Geo-NN using simulated data. Here, we choose the data generating process to be the Common Principal Components (CPC) framework Jolliffe & Cadima (2016). We compare the Geo-NN against two popular mean estimation algorithms introduced in Section 1.1. The first algorithm is Riemannian gradient descent Pennec et al. (2006) on the objective in Eq. (4). The second algorithm is the recently proposed **A**pproximate Joint Diagonalization **L**og **E**uclidean (ALE) mean estimation Congedo et al. (2015), which leverages properties of the approximate joint diagonalization objective Pham (2001).

Using the CPC formulation, each input SPD matrix $\mathbf{\Gamma}_n \in \mathcal{R}^{P \times P}$ is derived from a set of components $\mathbf{B} \in \mathcal{R}^{P \times P}$ common to the collection and a set of example specific (and strictly positive) weights across the components $\mathbf{c}_n \in \mathcal{R}^{(+)P \times 1}$. Let the diagonal matrix $\mathbf{C}_n$ be defined as $\mathbf{C}_n = \mathbf{diag}(\mathbf{c}_n) \in \mathcal{R}^{(+)P \times P}$. Each $\mathbf{\Gamma}_n$ is expressed as an outer-product $\mathbf{\Gamma}_n = \mathbf{B}\mathbf{C}_n\mathbf{B}^T$.

In the absence of corrupting noise, the theoretically optimal geodesic mean of the examples $\{\mathbf{\Gamma}_n\}_{n=1}^N$ can be computed in closed form as follows Congedo et al. (2015) (Proof in Appendix A.1):

$$\mathbf{G}_R^* = \mathbf{B}\ \mathbf{expm}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{logm}(\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T})\right]\mathbf{B}^T \tag{8}$$

3.1.1 Scalability of Geo-NN

We first use the noiseless CPC setup to evaluate the scalability of the Geo-NN when varying the dataset dimensionality P and the number of examples N. In this case, we compare the solution of each algorithm to the theoretically optimal geodesic mean in Eq. (8).

We randomly sample columns of the component matrix $\mathbf{B}$ from a standard normal, i.e., $\mathbf{B}[:, j] \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_P)\ \forall\ j \in \{1, \ldots, P\}$, where $\mathcal{I}_P$ is an identity matrix of dimension P. In parallel, we sample the component weights $\mathbf{c}_{nk}$ according to $\mathbf{c}_{nk}^{1/2} \sim \mathcal{N}(0, 1)\ \forall\ k \in \{1, \ldots, P\}$. To avoid degenerate behavior when the inputs are not full-rank, we clip $\mathbf{c}_{nk}$ to a minimum value of 0.001.

We consider two experimental scenarios. In **Experiment 1**, we fix the data dimensionality at $P = 30$ and sweep the dataset size as $N \in \{5, 10, 20, 50, 100, 200\}$. In **Experiment 2**, we fix the dataset size at $N = 20$ and sweep the dimensionality as $P \in \{5, 10, 20, 50, 100, 200\}$. For each setting, we run all estimation algorithms ten times using different random initializations.

Metrics: We score the performance of each algorithm based on the correctness of the solution and the execution time in seconds. The first metric of correctness is he final condition fit $\mathcal{L}(\mathbf{G}_R^{\text{est}})$ from

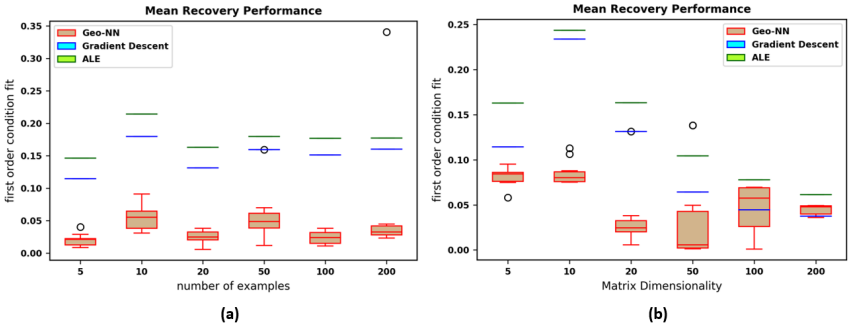

Figure 2: First-order condition fit (Eq. 7) for estimates for Geo-NN, gradient descent and ALE for varying **(a)** Dataset size N in Experiment 1 and **(b)** Data Dimension P in Experiment 2.

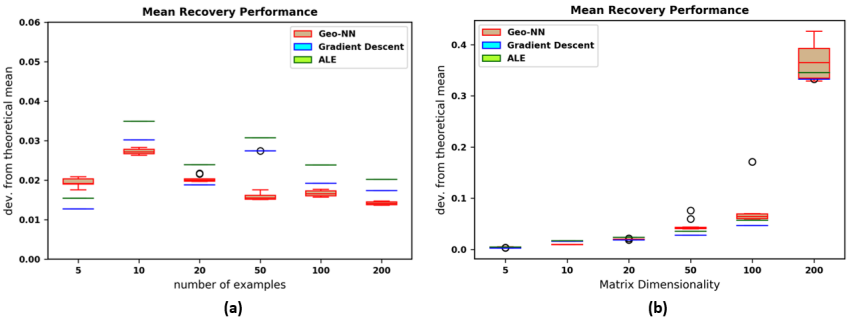

Figure 3: Deviation from the theoretical solution Eq. 8 for Geo-NN, gradient descent and ALE for varying **(a)** Dataset size N in Experiment 1 and **(b)** Data Dimension P in Experiment 2.

Eq. (7). It denotes the deviation of each estimate from the theoretical first order stationary point. The second metric of correctness is the scaled squared Riemannian distance from the theoretically optimal mean in Eq. (8). Mathematically, this distance is computed as $d_{\text{mean}} = d_R^2(\mathbf{G}_R^{\text{est}}, \mathbf{G}_R^*)/||\mathbf{G}_R^*||_R^2$. Lower values of the condition fit $\mathcal{L}(\mathbf{G}_R)$ and deviation d_{mean} imply a better quality solution.

Results: Fig. 2 illustrates the performance of Geo-NN, gradient descent and ALE mean estimation with respect to the first-order condition fit $\mathcal{L}(\mathbf{G}_R^{\text{est}})$. Fig. 2(a) plots the results when varying the dataset size N for a fixed matrix dimensionality P (Experiment 1), while Fig. 2(b) considers the opposite scenario (Experiment 2). Likewise, Fig. 3(a)-(b) plot d_{mean} for Experiment 1 and Experiment 2, respectively. We observe that the first order condition fit for the Geo-NN is better than the ALE for all settings, and better than the gradient descent for a majority of the settings. From Fig. 3, we note that the recovery performance of Geo-NN is better than the baselines in most cases while being a close approximation in the remaining ones.

Finally, Fig. 4(a)-(b) illustrate the time to convergence for each algorithm in Experiment 1 and Experiment 2, respectively. As seen, the performance of Geo-NN scales with dataset size but not matrix dimensionality. In all cases, it either beats or is competitive with ALE. We additionally compare the concordance of the recovered Geo-NN solutions in the Appendix (Fig. 7), and observe that the framework converges reliably.

Going one step further, we evaluate the efficacy of the Geo-NN framework when there is deviation from the ideal CPC generating process. We observe that the Geo-NN is robust to increasing levels of additive structured noise when compared with the baselines (Refer to the Appendix A.2.1).

3.1.2 Adaptability Beyond the CPC Generating Process

Finally, we consider a low-rank data generating process. Namely, SPD matrices in many real-world settings are assumed to be generated from a mixed effect setup, where K components ($K < P$) are common to the dataset (e.g., group mean), with the remaining $(P - K)$ components being unique

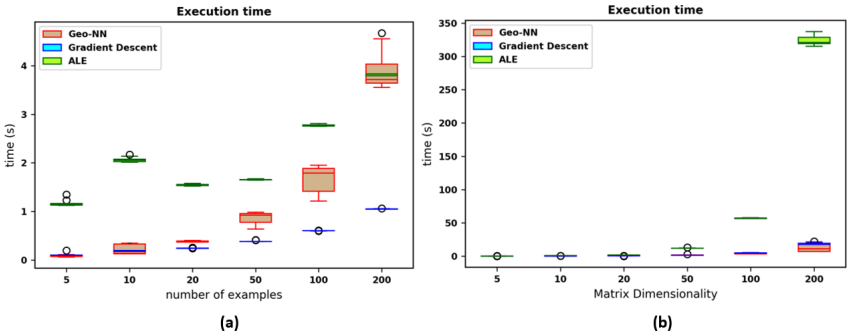

Figure 4: Execution time in seconds for Geo-NN, gradient descent and ALE for varying **(a)** Dataset size N in Experiment 1 and **(b)** Data Dimension P in Experiment 2.

to each data sample, or alternatively, treated as random noise. Thus, the input data $\{\Gamma_n\}$ is endowed with intrinsic low-rank structure despite individual matrices being full rank. This assumption is encountered in many application domains, such as spatio-temporal data modeling Schiratti et al. (2015), functional connectomics D'Souza et al. (2020), neural population analysis Keeley et al. (2020), and longitudinal data analysis McNeish & Bauer (2022).

Formally, we generate the SPD input matrices $\mathbf{\Gamma}_n$ via the following mixture model:

$$\mathbf{\Gamma}_n = \mathbf{B}\mathbf{C}_n\mathbf{B}^T + \mathbf{Q}_n\mathbf{diag}(\mathbf{d}_n)\mathbf{Q}_n^T/P$$
$$\mathbf{B}[:,j] \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_P) \;\; ; \;\; \mathbf{c}_{nj}^{1/2} \sim \mathcal{N}(0,1) \;\; \forall j \in \{1,\ldots,K\} K < P$$
$$\mathbf{Q}_n \in \mathbf{Null}(\mathbf{B}) \;\; ; \mathbf{d}_{nk}^{1/2} \sim \mathcal{N}(0,\sigma^2) \;\; \forall k \in \{1,\ldots,P-K\}$$

The columns of $\mathbf{Q}_n$ are scaled to unit norm and to lie in the null space of $\mathbf{B}$. We consider two scenarios of interest. In **Experiment 1:**, we fix the dimensionality K of the common generating process and increase the noise level parameterized by σ^2. Specifically, the low-rank dimension is

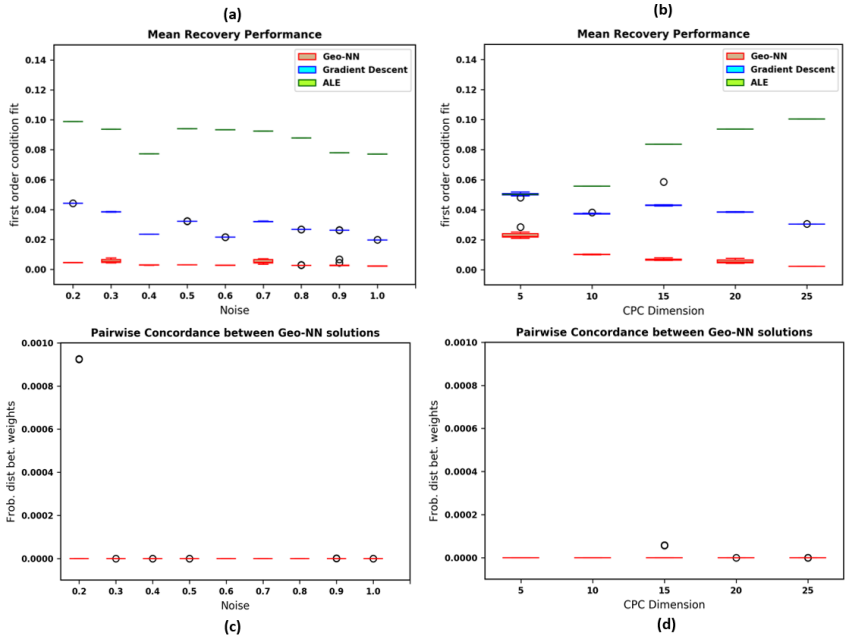

Figure 5: Performance of Geo-NN, gradient descent, and ALE under a low rank generating process. First order condition fit (Eq. 7) for **(a)** varying noise and **(b)** varying CPC dimensionality. Pairwise distance between recovered solutions for **(c)** varying noise and **(d)** varying CPC dimensionality.

held constant at $K = 20$, and the noise is varied in increments of 0.1 in the range $\sigma \in [0.2, 1]$. In **Experiment 2**, we vary the dimensionality K while keeping the noise level fixed. Here, we hold the noise constant at $\sigma = 0.3$, and vary the rank as $K \in \{5, 10, 15, 20, 25\}$. The dataset size and matrix dimension are fixed at $\{P = 30, N = 50\}$ in both experimental settings. Since there is no closed-form solution for the theoretical mean in the mixed effects case, we use the first order condition fit $\mathcal{L}(\mathbf{G}_R^{\text{est}})$ to compare performance across algorithms. We also evaluate the consistency of the Geo-NN solution across random initializations to quantify reliability.

Results: Fig. 5(a-b) report the first-order condition fit achieved by Geo-NN, gradient descent and ALE for Experiment 1 and Experiment 2, respectively. Likewise, Fig. 5(c-d) illustrates the consistency of the Geo-NN estimate for each experimental scenario. Once again, we observe that the solution recovered by the Geo-NN satisfies the first order condition more closely than both baselines algorithms. Empirically, we also encounter convergence issues using the ALE mean estimation in the low-rank setting. This can be seen for $K = 5$ in Fig. 5(b), where we have not plotted any values for the ALE algorithm because it does not converge to any finite estimate.

Taken together, the experimental results in Sections 3.1.1,A.2.1, and 3.1.2 demonstrate that Geo-NN is a robust and generalizable mean estimation algorithm across a variety of data generation scenarios. Encouraged by these results, we adopt Geo-NN to study a clinical neuroscience application below.

3.2 Experiments on Real-World Connectomics Data

Dataset: We adopt the Geo-NN for a groupwise discrimination task on the publicly available CNI 2019 Challenge dataset Schirmer et al. (2021). Mean regional time series are provided for 158 subjects diagnosed with Attention Deficit Hyperactivity Disorder (ADHD), 92 subjects with Autism Spectrum Disorder (ASD) with an ADHD comorbidity Leitner (2014), and 257 healthy controls. Functional connectomes (FC) are estimated via the Pearson's correlation matrix, regularized to be full-rank. We experiment on two different parcellations, the Automated Anatomical Atlas (AAL) Tzourio-Mazoyer et al. (2002) ($P = 116$) and the Craddocks 200 atlas Craddock et al. (2012) ($P = 200$). Further details about the data and pre-processing steps are provided in Appendix A.3.

Groupwise Discrimination: Given the comorbidity Leitner (2014), we expect that FC differences between the ASD and ADHD cohorts are harder to tease apart than differences between ADHD and controls Schirmer et al. (2021). We test this hypothesis by comparing the geodesic means estimated via Geo-NN for the three cohorts. For robustness, we perform bootstrapped trials for mean estimation by sampling 25 random subjects from a given group (ADHD/ASD/Controls). We then compute the Riemannian distance $d(\mathbf{G}_R(\{\mathbf{\Gamma}_{g1}\}), \mathbf{G}_R(\{\mathbf{\Gamma}_{g2}\}))$ between the Geo-NN means associated with groups $g1$ and $g2$. We run a Wilcoxon signed rank test to qualify differences in the distribution of $d(\cdot)$. A higher value of $d(\cdot)$ implies a better separation between the groups.

Fig. 6 illustrates the pairwise distances between cohorts $g1 - g2$ across bootstrapped trials. As a sanity check, we note that the mean estimates across samples within the same cohort (ADHD-ADHD) are closer than those across cohorts (ADHD-controls, ASD-controls, ADHD-ASD). More interest-

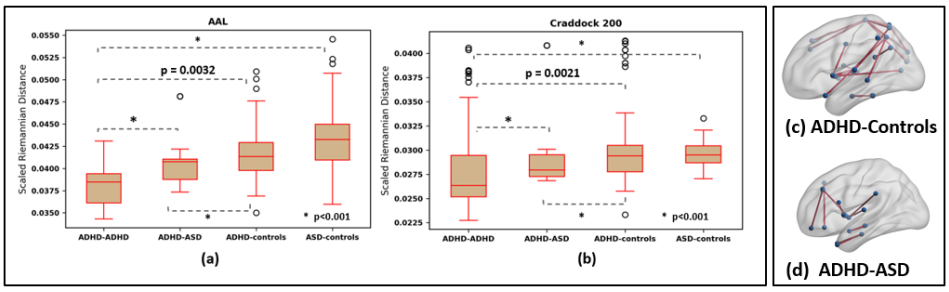

Figure 6: Groupwise discrimination between the FC matrices estimated via the **(a)** AAL **(b)** Craddock's 200 atlas, for the ADHD/ASD/Controls cohorts from the CNI 2019 Challenge Dataset. Pairwise distances are calculated between the Geo-NN mean estimates. Results of pairwise connectivity comparisons between Geo-NN group means for **(c)** ADHD-Controls **(d)** ADHD-ASD groups for the AAL parcellation. The red connections differ significantly ($p < 0.001$) across groups.

ingly, we observe that ADHD-controls separation is consistently larger than that of the ADHD-ASD groups for both parcellations. This result confirms the hypothesis that the overlapping diagnosis for the two classes translates to a reduced separability in the space of FC matrices and indicates that Geo-NN is able to robustly uncover population level differences in FC.

Classification: Building on the observation that Geo-NN provides reliable group-separability, we adopt this framework for classification. Using the AAL parcellation, we randomly sample 25 subjects from each class for training, and set aside the rest for evaluation with a $10\%/90\%$ validation/test split. We estimate the geodesic mean for each group across the training samples via 10 bootstrapped trials, in which we sub-sample 80% of the training subjects from the respective group. Permutation testing is performed on the mean estimates Zalesky et al. (2010), and functional connections (i.e., entries of $\mathbf{\Gamma}_n$) that differ with an FDR-corrected threshold of $p < 0.001$ are retained for classification. Finally, a Random Forest classifier is trained on the selected features to classify ADHD vs Controls. The train-validation-test splits are repeated 10 times to compute confidence intervals.

We use classification accuracy and area under the receiver operating curve (AU-ROC) as metrics for evaluation. The Geo-NN feature selection plus Random Forest approach provides an accuracy of 0.62 ± 0.031 and an AU-ROC of 0.60 ± 0.04 for ADHD-Control classification on the test samples. We note that this approach outperforms all but one method on the CNI challenge leaderboard Schirmer et al. (2021). Moreover, one focus of the challenge is to observe how models trained on the ADHD vs Control discrimination task translate to ASD (with ADHD comorbidity) vs Control discrimination in a transfer learning setup. Accordingly, we apply the learned classifiers in each split to ASD vs Control classification and obtain an accuracy of 0.54 ± 0.044 and an AU-ROC of 0.53 ± 0.03. This result is on par with the best performing algorithm in the CNI-TL challenge. The drop in accuracy and AU-ROC is consistent with the performance profile of all the challenge submissions. These results suggest that despite the comorbidity, connectivity differences between the cohorts are subtle and hard to reliably capture. Nonetheless, the Geo-NN+RF framework is a first step to underscoring stable, yet interpretable (see below) connectivity patterns that can discriminate between diseased and healthy populations.

Qualitative Analysis: To better understand the group-level connectivity differences, we plot the most consistently selected features (top 10 percent) from the previous experiment (ADHD-control feature selection) in Fig. 6(c). We utilize the BrainNetViewer Software for visualization. The blue circles are the AAL nodes, while the solid lines denote edges between nodes. We observe that the highlighted connections appear to cluster in the sensorimotor and visual areas of the brain, along with a few temporal lobe contributions. Altered sensorimotor and visual functioning has been previously reported among children and young adults diagnosed with ADHD Duerden et al. (2012); Ahrendts et al. (2011). Adopting a similar procedure, we additionally highlight differences among the ASD and ADHD cohorts in Fig. 6(d). The selected connections concentrate around the prefrontal areas of the brain, which is believed to be associated with altered social-emotional regulation in Autism Pouw et al. (2013). We additionally provide an extended version of the group connectivity difference results across trials in Fig. 9 (ADHD vs Controls) and Fig. 10 (ADHD vs ASD) in the Appendix. Across train-test-val splits, we observe that the connectivity differences are fairly consistent. Overall, the patterns highlighted via statistical comparisons on the Geo-NN estimates are both robust as well as in line with the physiopathology of ADHD and ASD reported in the literature.

4 Conclusion

We have proposed a novel geometric neural network framework, i.e. the Geo-NN, designed to reliably estimate the geodesic mean of SPD matrices. We devise a loss function that can optimize the first-order matrix normal condition for mean estimation via conventional stochastic optimization. Through extensive simulation studies, we demonstrate that the Geo-NN scales well to high-dimensional data, can handle input noise, and is more robust to variations in the data generating process when compared with current iterative methods. We also demonstrate the applicability of the Geo-NN to a real-world functional connectomics study for discovering consistent group differences between patients diagnosed with ADHD-Autism comorbidities and healthy controls. We also demonstrate the applicability of the framework for feature selection and classification in the functional connectomics setting. Given that the Geo-NN makes few assumptions, we envision it to be a valuable tool for research in geometric deep learning and beyond.

REFERENCES

Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

Johannes Ahrendts, Nicolas Rüsch, Marko Wilke, Alexandra Philipsen, Simon B Eickhoff, Volkmar Glauche, Evgeniy Perlov, Dieter Ebert, Jürgen Hennig, and Ludger Tebartz van Elst. Visual cortex abnormalities in adults with adhd: a structural mri study. *The World Journal of Biological Psychiatry*, 12(4):260–270, 2011.

Tsuyoshi Ando, Chi-Kwong Li, and Roy Mathias. Geometric means. *Linear algebra and its applications*, 385:305–334, 2004.

Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, David Vaillancourt, and Baba C Vemuri. Nonlinear regression on riemannian manifolds and its applications to neuro-image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 719–727. Springer, 2015.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to bci classification. In *International conference on latent variable analysis and signal separation*, pp. 629–636. Springer, 2010.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (compcor) for bold and perfusion based fmri. *Neuroimage*, 37(1):90–101, 2007.

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Marco Congedo, Bijan Afsari, Alexandre Barachant, and Maher Moakher. Approximate joint diagonalization and geometric mean of symmetric positive definite matrices. *PloS one*, 10(4):e0121423, 2015.

Marco Congedo, Alexandre Barachant, and Ehsan Kharati Koopaei. Fixed point algorithms for estimating power means of positive definite matrices. *IEEE Transactions on Signal Processing*, 65(9):2211–2220, 2017.

Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

R Cameron Craddock, G Andrew James, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–1928, 2012.

Luca Dodero, Ha Quang Minh, Marco San Biagio, Vittorio Murino, and Diego Sona. Kernel-based classification for brain connectivity graphs on the riemannian manifold of positive definite matrices. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 42–45. IEEE, 2015.

Zhen Dong, Su Jia, Chi Zhang, Mingtao Pei, and Yuwei Wu. Deep manifold learning of symmetric positive definite matrices with application to face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Emma G Duerden, Rosemary Tannock, and Colleen Dockstader. Altered cortical morphology in sensorimotor processing regions in adolescents and adults with attention-deficit/hyperactivity disorder. *Brain research*, 1445:82–91, 2012.

Niharika Shimona D'Souza, Mary Beth Nebel, Nicholas Wymbs, Stewart H Mostofsky, and Archana Venkataraman. A joint network optimization framework to predict clinical severity from resting state functional mri data. *NeuroImage*, 206:116314, 2020.

Niharika Shimona D'Souza, Mary Beth Nebel, Deana Crocetti, Joshua Robinson, Stewart Mostofsky, and Archana Venkataraman. A matrix autoencoder framework to align the functional and structural connectivity manifolds as guided by behavioral phenotypes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 625–636. Springer, 2021.

Roger Fletcher. Semi-definite matrix constraints in optimization. *SIAM Journal on Control and Optimization*, 23(4):493–513, 1985.

Daniel R Fuhrmann and Michael I Miller. On the existence of positive-definite maximum-likelihood estimates of structured covariance matrices. *IEEE transactions on information theory*, 34(4):722–729, 1988.

Tryphon T Georgiou. Distances and riemannian metrics for spectral density functions. *IEEE Transactions on Signal Processing*, 55(8):3995–4003, 2007.

Inbal Horev, Florian Yger, and Masashi Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *Asian Conference on Machine Learning*, pp. 1–16. PMLR, 2016.

Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

Ben Jeuris. Riemannian optimization for averaging positive definite matrices. 2015.

Ben Jeuris, Raf Vandebril, and Bart Vandereycken. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electronic Transactions on Numerical Analysis*, 39(ARTICLE):379–402, 2012.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Stephen Keeley, Mikio Aoi, Yiyi Yu, Spencer Smith, and Jonathan W Pillow. Identifying signal and noise structure in neural population activity with gaussian process factor models. *Advances in Neural Information Processing Systems*, 33:13795–13805, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

Yael Leitner. The co-occurrence of autism and attention deficit hyperactivity disorder in children–what do we know? *Frontiers in human neuroscience*, 8:268, 2014.

Daniel McNeish and Daniel J Bauer. Reducing incidence of nonpositive definite covariance matrices in mixed effect models. *Multivariate Behavioral Research*, 57(2-3):318–340, 2022.

Maher Moakher. 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.

Maher Moakher and Philipp G Batchelor. Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and processing of tensor fields*, pp. 285–298. Springer, 2006.

Xuan Son Nguyen. Geomnet: A neural network based on riemannian geometries of spd matrix space and cholesky space for 3d skeleton-based interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13379–13389, 2021.

Xuan Son Nguyen, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. A neural network based on spd manifold learning for skeleton-based hand gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12036–12045, 2019.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

Raúl Alberto Pérez and Graciela González-Farias. Partial least squares regression on symmetric positive-definite matrices. *Revista Colombiana de Estadística*, 36(1):177–192, 2013.

Dinh Tuan Pham. Joint approximate diagonalization of positive definite hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1136–1152, 2001.

Federico Poloni. Constructing matrix geometric means. *arXiv preprint arXiv:0906.3132*, 2009.

Lucinda BC Pouw, Carolien Rieffe, Lex Stockmann, and Kenneth D Gadow. The link between emotion regulation, social functioning, and depression in boys with asd. *Research in Autism Spectrum Disorders*, 7(4):549–556, 2013.

Jean-Baptiste Schiratti, Stéphanie Allassonniere, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. *Advances in neural information processing systems*, 28, 2015.

Markus D Schirmer, Archana Venkataraman, Islem Rekik, Minjeong Kim, Stewart H Mostofsky, Mary Beth Nebel, Keri Rosch, Karen Seymour, Deana Crocetti, Hassna Irzan, et al. Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge. *Medical image analysis*, 70:101972, 2021.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

Zhizhou Wang and Baba C Vemuri. Dti segmentation using an information theoretic tensor dissimilarity measure. *IEEE transactions on medical imaging*, 24(10):1267–1277, 2005.

Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

Florian Yger, Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy. Adaptive canonical correlation analysis based on matrix manifolds. *arXiv preprint arXiv:1206.6453*, 2012.

Andrew Zalesky, Alex Fornito, and Edward T Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010.

Teng Zhang. A majorization-minimization algorithm for the karcher mean of positive definite matrices. *arXiv preprint arXiv:1312.4654*, 2013.

A Appendix

A.1 Proof of the result in Eq. 8

Let $\mathbf{B}$ be an invertible matrix. Writing out the geodesic mean of the collection $\{\mathbf{\Gamma}_n\} \in \mathcal{P}_P^+$, we can use the congruence invariance property to establish the following:

$$\mathbf{G}_R(\{\mathbf{\Gamma}_n\}) = \mathbf{B}[\mathbf{G}_R(\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T})]\mathbf{B}^T \tag{9}$$

If the matrices $\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T}$ are exactly diagonal (i.e. solutions to the CPC objective Jolliffe & Cadima (2016)), then they commute in multiplication. The geodesic mean has a closed form, that can be computed by first averaging their matrix logarithms, and then applying a matrix exponential on the average. We can use this property to compute the geodesic mean of $\{\mathbf{\Gamma}_n\}$

$$\mathbf{G}_R^*(\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T}) = \mathbf{expm}\Big[\frac{1}{N}\sum_n[\mathbf{logm}(\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T})]\Big]$$

$$\therefore \quad \mathbf{G}_R^*(\{\mathbf{\Gamma}_n\}) = \mathbf{B}\mathbf{expm}\Big[\frac{1}{N}\sum_{n=1}^{N}[\mathbf{logm}(\mathbf{B}^{-1}\mathbf{\Gamma}_n\mathbf{B}^{-T})]\Big]\mathbf{B}$$

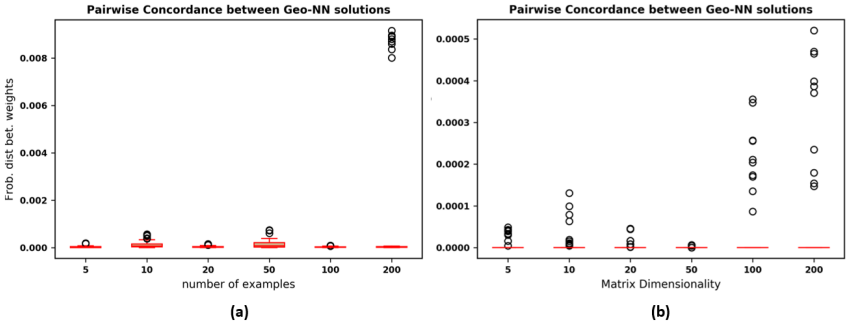

Figure 7: Pairwise distances between the recovered Geo-NN weights across initialization for varying **(a)** Dataset size **(b)** Data Dimensionality. Lower values indicate better stability

A.2 Synthetic Experiments: Quantifying Reliability of Geo-NN Optimization

For the experiment in Section 3.1.1, we additionally quantify the stability of the Geo-NN solution. To this end, we calculate the pairwise concordance of the final Geo-NN weights $\mathbf{W}_{\text{est}}$ across different initializations via the element-wise average of the Frobenius distance between solutions. Since the geodesic mean is computed as $\mathbf{W}\mathbf{W}^T = \mathbf{G}_R^{-1/2}$, lower values of d_{weights} indicate better agreement between recovered solutions. Fig. 7 plots d_{weights} against (a) varying number of examples, (b) for varying matrix dimensionality. From the scale of the error on the y-axis Fig. 7, we observe that the Geo-NN final solutions are in close agreement. In turn, this indicates that the optimization is robust to initialization and that the Geo-NN reliably converges to the same final solution. Combined with the results in Section 3.1.1, we conclude that the recovered estimates deviate only slightly from the optimal solution and consistently converge to the theoretical mean in almost all simulation settings.

A.2.1 Robustness to Noise

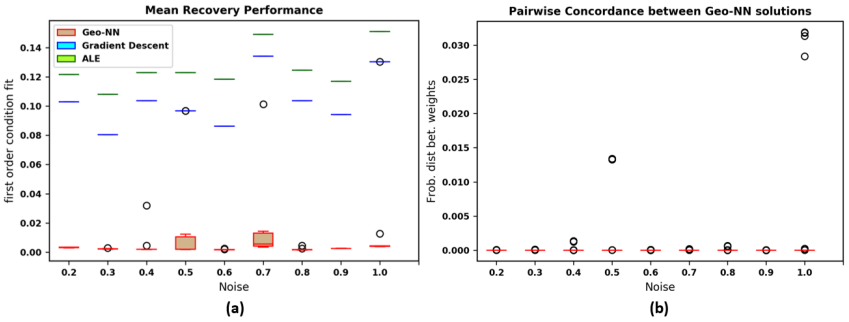

Figure 8: Performance of the Geo-NN, gradient descent and ALE estimation under increasing additive noise: **(a)** First order condition fit (Eq. 7) **(b)** Pairwise distance between the recovered Geo-NN solutions across random initializations. Lower values indicate better performance in each case.

In this experiment, we evaluate the performance when our simulated data deviates from the ideal CPC process. In this case, we add rank-one structured noise to obtain the input data: $\mathbf{\Gamma}_n = \mathbf{B}\mathbf{C}_n\mathbf{B}^T + \frac{1}{P}\mathbf{x}_n\mathbf{x}_n^T$. As before, the bases and coefficients are randomly sampled as $\mathbf{B}[:, j] \sim \mathcal{N}(\mathbf{0}, \mathcal{I}_P)$ and $\mathbf{c}_{nj}^{1/2} \sim \mathcal{N}(0, 1) \;\; \forall \;\; j \in \{1, \ldots, P\}$. In a similar vein, the structured noise is generated as $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathcal{I}_P) \in \mathcal{R}^{P\times 1}$, with σ^2 controlling the extent of the deviation. For this experiment, we set $P = 30, N = 20$ and vary the noise over the range $[0.2-1]$ in increments of 0.1.

One caveat in this setup is that Eq. (8) is no longer the theoretically optimal mean and cannot be used to evaluate performance. Hence, we report only the first-order condition fit $\{\mathcal{L}(\mathbf{G}_R)$ We also calculate the pairwise concordance d_{weights} of the final Geo-NN weights for different initializations.

Results: Fig. 8(a) illustrates the first-order condition fit $\mathcal{L}(\mathbf{G}_R^{\text{est}})$ across all three methods for increasing noise σ. As seen, $\mathcal{L}(\mathbf{G}_R^{est})$ for the Geo-NN is consistently lower than the corresponding value for the gradient descent and ALE algorithm, suggesting improved performance despite increasing corruption to the CPC process. We note that the ALE algorithm also optimizes the first-order condition fit, and its poor performance suggests that it is particularly susceptible to noise. Fig. 8(b) plots the pairwise distances between the geodesic means estimated by Geo-NN across the 10 random initializations. As seen, Geo-NN produces a consistent solution, thus underscoring its robustness.

A.3 RS-FMRI DATA PRE-PROCESSING

The CNI 2019 challenge data consists of preprocessed time resting-state fMRI (rs-fMRI) time series and demographic information Schirmer et al. (2021). The rs-fMRI data was acquired on a Phillips 3T Achieva scanner using a single shot, partially parallel, gradient-recalled EPI sequence with TR/TE = $2500/30$ms, flip angle 70, voxel resolution = $3.05 \times 3.15 \times 3$mm, with a scan duration of either 128 or 156 time samples (TR). Subjects were instructed to focus on a central cross-hair while remaining still and relax with their eyes open for the duration of the scan.

Rs-fMRI pre-processing consisted of slice time correction, rigid body realignment, and normalization to the EPI version of the MNI template. Temporal detrending was performed on the time courses, and spatially coherent noise was removed from the white matter and ventricles, along with the linearly detrended versions of the six rigid body realignment parameters and their first derivatives by using CompCorr Behzadi et al. (2007). The data were spatially smoothed with a 6mm FWHM Gaussian kernel and bandpass filtered between $0.01 - 0.1$Hz. Finally, the AFNI package Cox (1996) was used to perform spike correction in lieu of motion scrubbing.

A.4 UNCOVERING GROUP-LEVEL CONNECTIVITY DIFFERENCES VIA THE GEO-NN

We plot the selected features from the experiments in Section 3.2 for ADHD-control feature selection) in Fig. 9 for all train-test-val splits. Similarly, we plot the selected features for ADHD vs ASD differences in Fig. 10. We employ the BrainNetViewer Software Xia et al. (2013) for visualization. The blue circles are the AAL nodes, while the solid lines denote edges between nodes. We observe that across trials, i.e. random sub-samples across the cohort, several connectivity patterns show up fairly consistently. Bolstered by this observation, we utilize the group comparison approach as a viable feature selection framework to inform a downstream ADHD vs controls classification and for transfer learning (i.e. ASD vs controls classification).

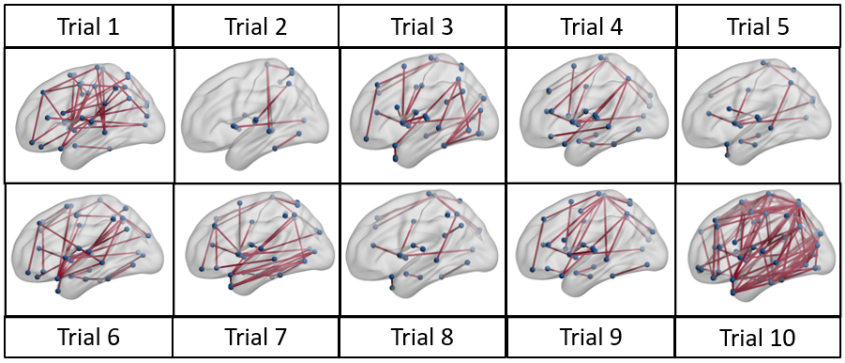

Figure 9: Results of pairwise connectivity comparisons between Geo-NN group means for ADHD-Controls groups (for training subjects) across train-test-validation aplits. The red connections differ significantly ($p < 0.001$) across groups

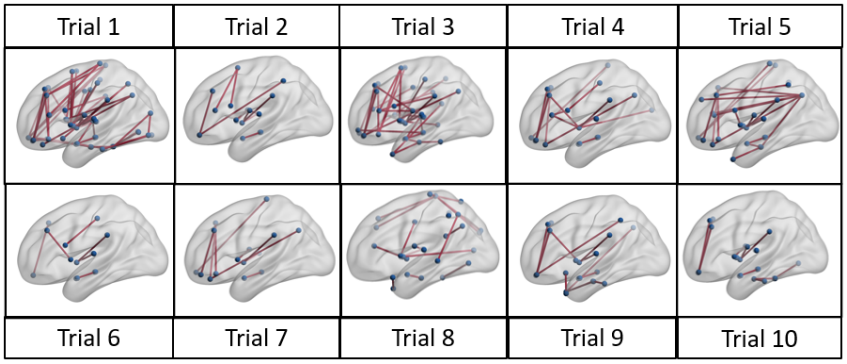

Figure 10: Results of pairwise connectivity comparisons between Geo-NN group means for ADHD vs ASD groups (for training subjects) across train-test-validation aplits. The ASD subjects have ADHD comorbidities. The red connections differ significantly ($p < 0.001$) across groups