scientific reports

Check for updates

OPEN

Scalable geometric learning with correlation-based functional brain networks

Kisung You[1], Yelim Lee[2] & Hae-Jeong Park[2,3,4]✉

Correlation matrices serve as fundamental representations of functional brain networks in neuroimaging. Conventional analyses often treat pairwise interactions independently within Euclidean space, neglecting the underlying geometry of correlation structures. Although recent efforts have leveraged the quotient geometry of the correlation manifold, they suffer from computational inefficiency and numerical instability, especially in high-dimensional settings. We propose a novel geometric framework that uses diffeomorphic transformations to embed correlation matrices into a Euclidean space while preserving critical manifold characteristics. This approach enables scalable, geometry-aware analyses and integrates seamlessly with standard machine learning techniques, including regression, dimensionality reduction, and clustering. Moreover, it facilitates population-level inference of brain networks. Simulation studies demonstrate significant improvements in both computational speed and predictive accuracy over existing manifold-based methods. Applications to real neuroimaging data further highlight the framework's versatility, improving behavioral score prediction, subject fingerprinting in resting-state fMRI, and hypothesis testing in EEG analyses. To support community adoption and reproducibility, we provide an open-source MATLAB toolbox implementing the proposed techniques. Our work opens new directions for efficient and interpretable geometric modeling in large-scale functional brain network research.

A widely accepted view of the human brain is that it operates as a network formed by interactions among distributed regions[1]. These interactions are often quantified using second-order statistics, including covariance, precision, and correlation matrices, which capture spontaneous fluctuations observed in resting-state functional magnetic resonance imaging (rs-fMRI)[2] or through electroencephalogram (EEG) and magnetoencephalogram (MEG) recordings[3,4].

In most studies employing correlation matrices, interactions along individual edges are analyzed either independently of other edges[5–7] or collectively, to identify sets of edges that interact synergistically[6,8]. However, these approaches often fail to account for the intrinsic dependence structure among edges in a correlation matrix. A correlation matrix contains richer information as a whole than its individual pairwise correlations suggest, underscoring the need to treat it as a manifold-valued object with well-defined geometric properties.

Mathematically, the correlation matrix belongs to the class of symmetric, positive-definite (SPD) matrices[9], the collection of which constitutes a Riemannian manifold. This SPD perspective has been actively employed across various tasks in brain functional network analysis, including dynamic modeling[10], multi-frequency network fusion[11], multi-site representation learning[12], and identification of brain network reconfiguration under substance-use disorder[13], among recent works. Note that an increasing number of studies have directly adopted the SPD-manifold framework for analyzing correlation-valued networks[14–17]. A significant challenge in treating correlation matrices as SPD-manifold objects arises because operations on these matrices often yield outputs that deviate from a valid correlation matrix, necessitating post hoc normalization to enforce unit diagonal elements. In a previous study[18], we addressed this issue by iteratively normalizing the matrices at each intermediate step. Although effective in most scenarios, this heuristic lacks mathematical rigor and does not guarantee exact solutions.

Unlike the SPD manifold, the space of correlation matrices, referred to as the elliptope[19], has received relatively limited attention. Only a few notable studies have explored this space, yet these efforts are hindered by either undesirable properties or a lack of efficient computational methods[20,21]. A promising alternative leverages

[1]Department of Mathematics, Baruch College, City University of New York, New York, USA. [2]Graduate School of Medical Science, Brain Korea 21 Project, Department of Nuclear Medicine, Psychiatry, Yonsei University College of Medicine, Seoul, Republic of Korea. [3]Center for Systems and Translational Brain Science, Institute of Human Complexity and Systems Science, Yonsei University, Seoul, Republic of Korea. [4]Department of Cognitive Science, Yonsei University, Seoul, Republic of Korea. ✉email: parkhj@yonsei.ac.kr

the quotient geometry of the SPD manifold, induced by the affine-invariant Riemannian metric, to represent the space of correlation matrices[22–24].

Building upon these advancements, our previous study[25] incorporated the quotient geometry of the correlation manifold into well-known algorithms in machine learning and statistical inference, specifically for FC analysis. Despite its mathematical soundness and high performance, this approach faces critical challenges. These include computational inefficiency and numerical instability when applied to high-dimensional data, raising concerns about the robustness of the results. Addressing these limitations requires new methods to enable routine learning tasks at a practical scale.

Recently, novel geometric structures for the correlation manifold based on specialized transformations were introduced[26]. These transformations preserve much of the geometric characteristics of the manifold while mapping correlation matrices to vectors by diffeomorphism, allowing the use of Euclidean geometry. This framework offers two key advantages: it facilitates the direct application of established algorithms from conventional learning paradigms and improves computational efficiency by confining expensive numerical operations to a one-time transformation.

The primary objective of the present study is to introduce these theoretical advancements to the neuroimaging community and demonstrate how this underutilized framework can enhance statistical learning with correlation-valued data in population-level FC analysis. The proposed approach achieves substantial computational speedups, thereby enabling correlation-based analyses for large-scale network studies, a significant limitation of previous methods, including our own.

This study is organized into three main sections. First, we revisit the foundational theory of Riemannian geometry and correlation manifolds. Next, we present the novel geometric structures and extend a suite of learning algorithms across multiple task categories. Finally, we evaluate the performance of these algorithms from computational and theoretical perspectives and apply the proposed pipeline to experimental data. To promote broader adoption, all algorithms have been implemented in a MATLAB toolbox (MathWorks, Inc., USA), which is freely available on a code-sharing platform for use by the neuroimaging community.

Background

Geometry of SPD and CORR manifolds

Brain functional connectivity (FC) is commonly represented using second-order statistics, such as covariance, correlation, or precision matrices. Mathematically, these matrices belong to the class of symmetric positive-definite (SPD) matrices, which are formally defined as follows:

Definition 1 $\mathscr{S}^n_{++}$ is the space of $(n \times n)$ symmetric positive-definite matrices:

$$\mathscr{S}^n_{++} = \{X \in \mathbb{R}^{n\times n} \mid X = X^\top,\ \lambda_{\min}(X) > 0\},$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of the matrix.

As a mathematical space, $\mathscr{S}^n_{++}$ has a dimension of $(n^2+n)/2$ and has attracted significant attention due to the frequent occurrence of such matrices in data analysis[9]. Among the various geometric structures available, the affine-invariant Riemannian metric (AIRM)[23] is one of the most prominent for $\mathscr{S}^n_{++}$, defining it as a Riemannian manifold.

Under AIRM, the geodesic distance $d_{\mathscr{S}^n_{++}}(P,Q)$ between two SPD matrices $P, Q \in \mathscr{S}^n_{++}$ is expressed as $d^2_{\mathscr{S}^n_{++}}(P,Q) = \|\log(P^{-1}Q)\|^2_F$, where $\|A\|_F = \sqrt{\mathrm{Tr}(A^\top A)}$ is the Frobenius norm, and $\log(\cdot)$ denotes the matrix logarithm[27]. For any symmetric positive-definite matrix, the matrix logarithm is computed through its eigendecomposition, making it a well-defined and computationally feasible operation.

Our primary focus is on representing functional connectivity (FC) using correlation matrices, which constitute a specialized subset of SPD matrices. The space of correlation matrices, denoted as $\mathscr{C}^n_{++}$, is formally defined as follows:

Definition 2 $\mathscr{C}^n_{++}$ is the space of $(n \times n)$ symmetric positive-definite matrices with unit diagonal elements:

$$\mathscr{C}^n_{++} = \{X \in \mathbb{R}^{n\times n} \mid X \in \mathscr{S}^n_{++},\ \mathrm{diag}(X) = 1_n\},$$

where $\mathrm{diag}(A)$ is a vector consisting of the diagonal elements of matrix A, and 1_n is the vector of length n with all elements equal to 1.

This definition establishes that $\mathscr{C}^n_{++}$ is a strict subset of $\mathscr{S}^n_{++}$. For illustration, consider the simple case $n = 2$, namely the collection of 2×2 SPD and correlation matrices. A convenient way to visualize $\mathscr{S}^2_{++}$ is as the interior of the open upper cone in $\mathbb{R}^3$[9]. Let C be a 2×2 correlation matrix,

$$C = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

Every matrix in $\mathscr{C}^2_{++}$ has exactly one free parameter, the off-diagonal element r. Therefore, $\mathscr{C}^2_{++}$ corresponds to a one-dimensional manifold in $\mathbb{R}^3$, appearing as an open line segment, as shown in Figure 1.

In more general settings, the space of correlation matrices can be endowed with a Riemannian manifold structure via the theory of quotient manifolds. In particular, the quotient-affine metric (QAM)[22] inherits the affine-invariant Riemannian metric (AIRM) from $\mathscr{S}^n_{++}$ and thereby provides a Riemannian framework for correlation matrices.

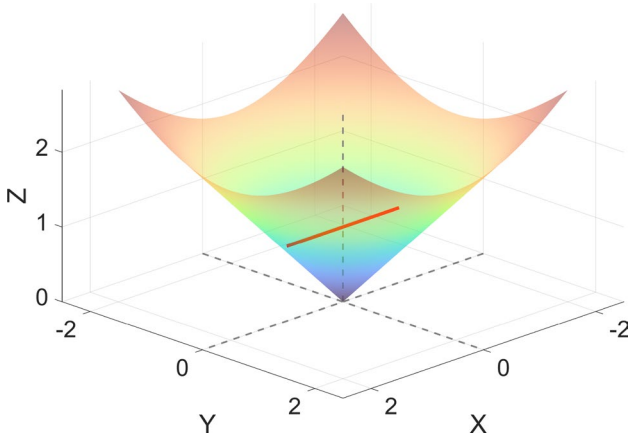

Fig. 1. Visualization of the 2×2 symmetric and positive-definite (SPD) manifold as the interior of the open upper cone in $\mathbb{R}^3$. The dashed gray lines indicate the coordinate axes, and the red line represents the correlation manifold, with its endpoints excluded, embedded within the SPD region.

Under the QAM, the geodesic distance between $P, Q \in \mathscr{C}^n_{++}$ is given by $d^2_{\mathscr{C}^n_{++}}(P,Q) = \min d^2_{\mathscr{S}^n_{++}}(P, DQD)$ for $D \in \mathscr{D}^n_{++}$, where $\mathscr{D}^n_{++}$ represents the set of $(n \times n)$ diagonal matrices with strictly positive entries, and $d_{\mathscr{S}^n_{++}}$ denotes the geodesic distance on the SPD manifold under AIRM.

Unlike the direct computation of geodesic distance in $\mathscr{S}^n_{++}$ using AIRM, calculating the geodesic distance on $\mathscr{C}^n_{++}$ under QAM involves solving a nonlinear optimization problem. Each iteration of this process requires eigendecomposition to compute matrix square roots and logarithms, making the procedure computationally intensive. For further details on the AIRM and QAM geometries in the context of FC analysis, we direct readers to our earlier works[18,25].

New geometries

While the development of QAM geometry offers a promising framework for geometric learning on $\mathscr{C}^n_{++}$, it becomes computationally prohibitive as the number of FC matrices matrices or the dimensionality of the regions of interest (ROIs) increases. In this section, we examine two alternative geometries for $\mathscr{C}^n_{++}$.

We start by establishing the notations used throughout this section. The Cholesky decomposition is defined as the mapping $Chol : \mathscr{S}^n_{++} \to \mathscr{L}^n_+$, where $\mathscr{L}^n_+$ denotes the set of $(n \times n)$ lower-triangular matrices with positive diagonal entries. For any $\Sigma \in \mathscr{S}^n_{++}$, the Cholesky decomposition $Chol(\Sigma) = L$ ensures that $\Sigma = LL^\top$. The symbols $\mathscr{L}^n_0$ and $\mathscr{L}^n_1$ represent the sets of lower-triangular matrices with zero diagonals and unit diagonals, respectively. Additionally, the operation $Diag(\cdot)$, when applied to a square matrix A, zeros out all off-diagonal elements, such that $Diag(A)_{i,j} = A_{i,j}$ if $i = j$ and 0 otherwise. These notations will be integral in describing and analyzing the alternative geometries for $\mathscr{C}^n_{++}$.

The Euclidean-Cholesky metric (ECM) represents the first of these geometries, which transforms a correlation matrix into a lower-triangular matrix with unit diagonals. This transformation is defined as $\Theta : \mathscr{C}^n_{++} \to \mathscr{L}^n_1$ such that for any $C \in \mathscr{C}^n_{++}$,

$$\Theta(C) = Diag(Chol(C))^{-1} \cdot Chol(C).$$

This mapping ensures that the resulting lower-triangular matrix belongs to $\mathscr{L}^n_1$, facilitating the application of Euclidean geometry in this transformed space.

The map Θ is smooth, allowing the use of the vector space structure of $\mathscr{L}^n_1$ to define a pullback metric through Θ, incorporating the logarithmic transformation of the diagonal elements in $\mathscr{L}^n_1$. Under the ECM geometry, the distance between two correlation matrices $C_1, C_2 \in \mathscr{C}^n_{++}$ is defined as

$$d_{\mathrm{ECM}}(C_1, C_2) = \|\Theta(C_1) - \Theta(C_2)\|_F, \tag{1}$$

where $\|\cdot\|_F$ denotes the standard Frobenius norm. The unique geodesic curve $\gamma : [0,1] \to \mathscr{C}^n_{++}$ connecting the two points C_1 and C_2 is expressed as

$$\gamma_{\mathrm{ECM}}(t) = \Theta^{-1}\left((1-t) \cdot \Theta(C_1) + t \cdot \Theta(C_2)\right),$$

with $\gamma_{\mathrm{ECM}}(0) = C_1$ and $\gamma_{\mathrm{ECM}}(1) = C_2$. The inverse mapping $\Theta^{-1} : \mathscr{L}^n_1 \to \mathscr{C}^n_{++}$ for any $L \in \mathscr{L}^n_1$ is explicitly available as the following:

$$\Theta^{-1}(L) = Diag(LL^\top)^{-1/2} \cdot LL^\top \cdot Diag(LL^\top)^{-1/2},$$

ensuring that $\Theta^{-1} \circ \Theta(C) = C$ for all $C \in \mathscr{C}^n_{++}$.

Building on ECM, the Log-Euclidean Cholesky metric (LEC) introduces a different vector space structure on $\mathscr{C}_{++}^n$ by applying the matrix logarithm to Θ. Recall that the logarithm of a square matrix Z is defined through the power series:

$$\log(Z) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} (Z - I_n)^k,$$

where I_n is the identity matrix of size $n \times n$. This series converges for matrices Z whose eigenvalues lie in the positive real half-plane, making it a suitable operation for matrices in $\mathscr{S}_{++}^n$.

Given $Z \in \mathscr{L}_1^n$, it follows that $Z - I_n \in \mathscr{L}_0^n$ because Z has unit diagonals. Consequently, $(Z - I_n)^k$ remains strictly lower-triangular for $k < n$ and becomes the zero matrix for $k \geq n$. This property allows the matrix logarithm to serve as a smooth mapping from $\mathscr{L}_1^n$ to $\mathscr{L}_0^n$, which involves only a finite number of matrix powers. The LEC defines the composite mapping $\log \circ \Theta : \mathscr{C}_{++}^n \to \mathscr{L}_0^n$, which acts as a diffeomorphism and equips $\mathscr{C}_{++}^n$ with the pullback metric of the standard Euclidean inner product. The distance between two points $C_1, C_2 \in \mathscr{C}_{++}^n$ under the LEC geometry is given by:

$$d_{\mathrm{LEC}}(C_1, C_2) = \| \log \circ \Theta(C_1) - \log \circ \Theta(C_2) \|_F, \tag{2}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Similar to the ECM framework, the geodesic curve $\gamma : [0, 1] \to \mathscr{C}_{++}^n$ under LEC geometry, connecting C_1 and C_2, is determined as:

$$\gamma_{\mathrm{LEC}}(t) = (\log \circ \Theta)^{-1} \big((1 - t) \cdot \log \circ \Theta(C_1) + t \cdot \log \circ \Theta(C_2) \big),$$

where $(\log \circ \Theta)^{-1}$ is the composite inverse of the matrix exponential and Θ, explicitly defined as $(\log \circ \Theta)^{-1} = \Theta^{-1} \circ \exp$. These transformations are illustrated in Figure 2.

We remark the advantages of adopting the geometries described above. First, these geometries exhibit the characteristics of standard Euclidean space via diffeomorphic transformations, leading to zero curvature as in Euclidean space. This property enables the straightforward application of interpolation, extrapolation, and the computation of unique centroids due to the homogeneous space property[28]. Furthermore, these geometries preserve critical manifold properties such as smooth manifold structure, existence and uniqueness of geodesics in their respective parametrizations, geodesic completeness, and uniqueness of Fréchet means, the components of which form theoretical soundness in many algorithms we introduce later.

Second, these geometries offer significant computational benefits. Both start with the Cholesky decomposition, which has a computational complexity of $O(n^3)$[29]. In the case of the LEC geometry, the $\log \circ \Theta$ mapping requires an additional matrix logarithm step, with complexity $O(n^\omega)$ for $\omega \in (2, 2.376)$[30]. Once the transformation is performed, subsequent computations follow standard multivariate analysis routines in the Euclidean space. In contrast, QAM geometry necessitates solving an optimization problem even for basic distance computations, making it significantly less efficient. For completeness, we describe memory and storage complexity as well as parallelization in the Supplementary Information.

This distinction is illustrated in the following example, where we compute the distance between two correlation matrices C_1 and $C_2 \in \mathscr{C}_{++}^n$. Here, C_1 is the identity matrix, and C_2 is derived from an AR(1) process with $C_2(i, j) = \rho^{|i-j|}$ for $\rho = 0.8$. Figure 3 summarizes the average runtime over 50 trials for computing distances between perturbed versions of C_1 and C_2 across varying dimensions $n = 10, 20, \ldots, 100$. The ECM and LEC geometries demonstrate remarkable computational efficiency, outperforming QAM geometry by several orders of magnitude.

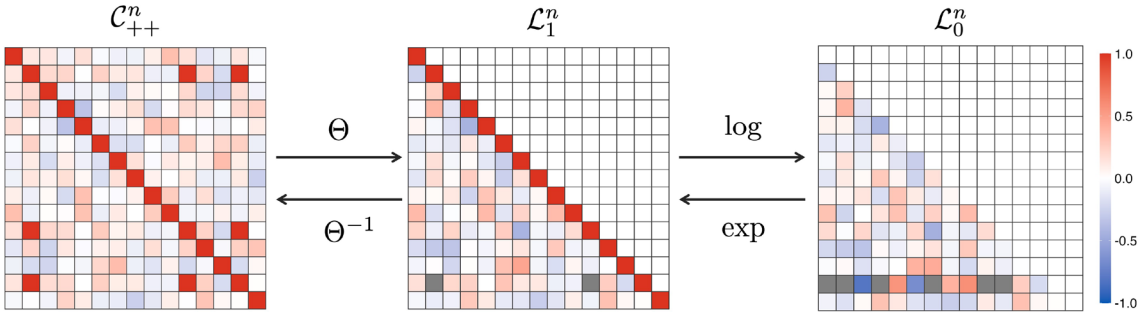

Fig. 2. Diagram of the transformation process for ECM and LEC geometries. Applying the mapping Θ to a full-rank correlation matrix (left) results in a lower-triangular matrix with unit diagonals (middle). The subsequent application of the matrix logarithm to $\mathscr{L}_1^n$ produces strictly lower-triangular matrices with zero diagonals (right).

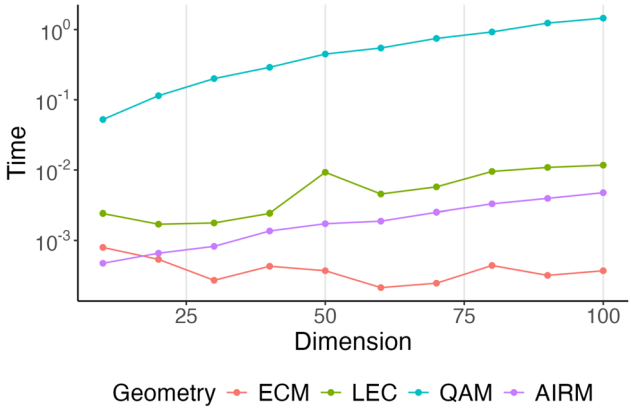

Fig. 3. Comparison of average wall-clock runtime over 50 trials for computing the distance between perturbed versions of two model correlation matrices, C_1 and C_2, at varying dimensions $n = 10, 20, \ldots, 100$. The y-axis represents the average runtime in seconds, displayed on a base-10 logarithmic scale.

Category	Function	Description
Exploratory Analysis	**corr_mean.m**	compute the Fréchet mean and variation
	corr_median.m	compute the Fréchet median and variation
Regression	**corr_gpreg.m**	Gaussian process regression
	corr_kernreg.m	kernel regression
	corr_svmreg.m	support vector regression
Dimensionality Reduction	**corr_ae.m**	shallow autoencoder
	corr_cmds.m	classical multidimensional scaling
	corr_mmds.m	metric multidimensional scaling
	corr_pga.m	principal geodesic analysis
	corr_tsne.m	t-stochastic neighbor embedding
Cluster Analysis	**corr_kmeans.m**	k-means clustering
	corr_kmedoids.m	k-medoids clustering
	corr_specc.m	spectral clustering
	corr_silhouette.m	cluster validity index of Silhouette score
	corr_CH.m	cluster validity index of Calinski and Harabasz
Hypothesis Testing	**corr_test2bg.m**	two-sample test via Biswas-Ghosh method
	corr_test2energy.m	two-sample test with the energy distance
	corr_test2mmd.m	two-sample test via maximum mean discrepancy
	corr_test2wass.m	two-sample test with the Wasserstein distance

Table 1. Summary of learning algorithm categories for population-level inference using correlation-based functional connectivity. The middle column provides the MATLAB function names included in the CORRbox package.

Methods

This section introduces several categories of algorithms tailored for analyzing populations of FC matrices, which is summarized in Table 1, where each connectivity matrix is treated as an element of $\mathscr{C}^n_{++}$. Throughout this section, the symbol d represents a general distance metric, which may correspond to distances defined by the ECM or LEC geometries, as described in Equations (1) and (2). Whenever these specific metrics are utilized, they will be explicitly indicated.

Exploratory analysis

Consider a random sample of FC representations $\{\Sigma_i\}_{i=1}^m \subset \mathscr{C}^n_{++}$. The initial step in data analysis often involves examining the sample's summary statistics, such as its centroid and dispersion. These are referred to as generalized Fréchet means or L_p centers of mass in the context of manifold-valued data analysis[28].

The L_p center of mass is defined as the minimizer of the functional:

$$F_p(\Sigma) = \frac{1}{m}\sum_{i=1}^{m} d^p(\Sigma, \Sigma_i), \tag{3}$$

where $p \geq 1$. For $p = 2$, the minimizer is known as the Fréchet mean (**corr_mean.m**), which generalizes the notion of the mean to general metric spaces. For $p = 1$, the minimizer of $F_1(\Sigma)$ is called the Fréchet or geometric median (**corr_median.m**), a robust alternative to the Fréchet mean[31]. The Fréchet variation, which measures dispersion, generalizes the concepts of variance and mean absolute deviation for $p = 2$ and $p = 1$, respectively. Denoting the minimizer of Equation (3) as $\hat{\Sigma}$, the Fréchet variation is given by $F_p(\hat{\Sigma})$, representing the value of the functional at its minimum.

We make a remark that both geometries guarantee the existence of a unique Fréchet mean. However, the uniqueness of the Fréchet median requires the assumption that the images of the mappings are not collinear, which is rare in practice.

Regression on scalars

Can individual FC predict phenotypic traits? This question falls under the domain of regression analysis, which investigates the relationship between individuals' brain networks and variables of interest. Given a set of correlation-based connectivity matrices as independent variables and scalar phenotypes as dependent variables, the data pairs $(\Sigma_i, Y_i)_{i=1}^m \subset \mathscr{C}_{++}^n \times \mathbb{R}$ are analyzed to identify a function $f : \mathscr{C}_{++}^n \to \mathbb{R}$ that satisfies

$$Y_i = f(\Sigma_i) + \epsilon_i,$$

where ϵ_i is an additive error term. The function f can be estimated under specific assumptions regarding the functional form and error distribution.

We explore three nonlinear regression models within the framework of kernel methods[32]: (1) Gaussian process regression (**corr_gpreg.m**), (2) kernel regression (**corr_kernreg.m**), and (3) support vector regression (**corr_svmreg.m**). Two main reasons motivate the inclusion of kernel-based approaches in this context.

First, while linear models are simple and interpretable, they often lack the flexibility required for capturing the inherent nonlinear relationships in correlation matrices. Although transformations such as Θ or $\log \circ \Theta$ could be applied to linear models, they still fail to improve interpretability due to the nonlinear nature of such transformations. Additionally, linear models may underperform when modeling the complexities of brain connectivity.

Second, kernel methods built on ECM and LEC geometries offer theoretical and practical advantages. Kernel methods replace inner products in high-dimensional feature spaces with kernel functions, leveraging the Gram matrix $K \in \mathbb{R}^{m \times m}$, where $K_{i,j} = k(x_i, x_j)$ for some kernel function $k(\cdot, \cdot)$. For consistent and generalizable performance, the Gram matrix must be positive semi-definite[33]. A positive-definite kernel is a continuous function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ satisfying

$$\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j k(x_i, x_j) \geq 0$$

for any $x_1, \ldots, x_m \in \mathscr{X}$ and $c_1, \ldots, c_m \in \mathbb{R}$. A commonly used positive-definite kernel is the squared exponential kernel, which is valid under both ECM and LEC geometries.

The following proposition establishes that the squared exponential kernel is positive-definite, the proof of which is available in the Supplementary Information.

Proposition 1 For $C_i, C_j \in \mathscr{C}_{++}^n$ and a non-negative constant $\theta \geq 0$, the squared exponential kernel

$$k(C_i, C_j) = \exp\left(-\theta \cdot d_*^2(C_i, C_j)\right) \tag{4}$$

is a positive-definite kernel when $d_* = d_{\mathrm{ECM}}$ or d_{LEC}.

It is important to note that the distance in quotient geometry does not inherently guarantee positive definiteness of the induced kernels due to the intrinsic curvature of $\mathscr{C}_{++}^n$. We also emphasize that the three regression algorithms discussed, Gaussian process regression, kernel regression, and support vector regression rely on several hyperparameters. To optimize these hyperparameters, we utilized 5-fold cross-validation, ensuring robust fine-tuning for generalization and performance.

Dimensionality reduction

Visual examination of data distribution is an invaluable step in analyzing complex datasets, as it provides intuitive insights into patterns and relationships that might be obscured in high-dimensional representations. The field of dimensionality reduction addresses this challenge by focusing on techniques to represent high-dimensional data in low-dimensional spaces that are more interpretable by humans[34]. From a theoretical standpoint, dimensionality reduction involves finding a mapping f from the original data space onto $\mathbb{R}^d$, where $d = 2, 3$ to facilitate visualization. This mapping can be defined either explicitly, where the transformation is mathematically described, or implicitly, where the relationship is inferred through computational algorithms. Another benefit of dimensionality reduction includes mitigating the curse of dimensionality in a non-trivial setting of correlation manifold as the number of samples is typically at a much smaller scale than the original dimensionality of the manifold induced by the number of variables such as ROI size.

In the context of our work, we leverage the Euclideanization properties of the proposed geometries, ECM and LEC, to incorporate several dimensionality reduction algorithms into our analysis toolkit without developing dedicated algorithms on $\mathscr{C}_{++}^n$. A list of algorithms capable of defining an explicit form of mapping from $\mathscr{C}_{++}^n$ to $\mathbb{R}^d$ includes principal geodesic analysis (**corr_pga.m**)[35] and shallow autoencoders (**corr_ae.m**)[36]. On the other

hand, methods such as classical and metric multidimensional scaling (**corr_cmds.m**, **corr_mmds.m**)[37], and *t*-stochastic neighbor embedding (**corr_tsne.m**)[38] belong to a category of algorithms that do not provide explicit mappings.

It is important to emphasize that except for principal geodesic analysis, the dimensionality reduction methods integrated into our framework are nonlinear. This nonlinearity is crucial for capturing and representing the low-dimensional structures embedded within the correlation manifold. These methods are particularly advantageous for uncovering complex relationships in data, as they adapt to the curvature and geometry of $\mathscr{C}_{++}^{n}$, providing a more faithful representation of the intrinsic structure of the dataset.

Cluster analysis

Cluster analysis focuses on uncovering the inherent subgroup structure within data when the true labels of the data points are unknown[39]. This type of unsupervised learning is particularly valuable in exploratory data analysis, where prior knowledge about the underlying groups is limited. In this context, we emphasize partitional clustering algorithms, including geometry-aware version of *k*-means (**corr_kmeans.m**)[40], *k*-medoids (**corr_kmedoids.m**)[41], and spectral clustering (**corr_specc.m**)[42], all of which are designed to identify a predetermined number of clusters within a set of correlation matrices.

In partitional clustering, the primary objective is to minimize within-cluster variation while maximizing between-cluster separation. Algorithms such as *k*-means and *k*-medoids operate by iteratively refining the assignment of data points to clusters based on a predefined distance metric. In our case, the Euclideanized metrics derived from both geometries are employed. The *k*-means algorithm uses centroids to represent each cluster, which may be unsuitable for non-Euclidean spaces. However, the geometric adaptations in **corr_kmeans.m** ensure compatibility with $\mathscr{C}_{++}^{n}$. On the other hand, the *k*-medoids algorithm selects actual data points as cluster representatives, providing a robust alternative, especially when data distributions are non-convex or include outliers. Spectral clustering further enhances this toolkit by leveraging eigenvalue decomposition on an affinity matrix constructed from pairwise distances, allowing for flexible and effective identification of nonlinearly separable clusters.

In our context, these clustering methods facilitate the discovery of cohesive yet distinct subpopulations of FC patterns. This is particularly valuable in neuroimaging studies, where the data often exhibits high heterogeneity. For instance, in research on autism spectrum disorder, identifying meaningful subgroups can provide insights into the disorder's variability across individuals and may even guide personalized therapeutic approaches[43].

To evaluate clustering performance, we employ two metrics: the Silhouette score (**corr_silhouette.m**)[44] and the Calinski-Harabasz (CH) index (**corr_CH.m**)[45]. The Silhouette score measures the degree of cohesion and separation within clusters, with higher values indicating well-defined and distinct clusters. It is computed as the difference between the average intra-cluster distance and the smallest average inter-cluster distance, normalized by the maximum of the two. This score provides an interpretable measure of how similar each data point is to its assigned cluster relative to other clusters.

The CH index evaluates clustering quality by comparing the dispersion of data points within clusters to the dispersion between clusters. Specifically, it computes the ratio of between-cluster dispersion to within-cluster dispersion, scaled by the number of clusters and the total number of data points. Higher CH values indicate better-defined cluster structures, making it an effective tool for selecting the optimal number of clusters.

Given the limited *a priori* knowledge of subgroup characterization in data-driven studies, these indices are critical for assessing clustering quality and determining the most appropriate or plausible number of clusters. They provide a quantitative basis for validating clustering outcomes, enabling researchers to interpret results with greater confidence.

Hypothesis testing

The final set of algorithms pertains to two-sample hypothesis testing, an essential framework for comparing two groups of FC matrices to identify statistically significant differences. Consider two sets of correlation matrices, $C_1^{(1)}, \ldots, C_{m_1}^{(1)}$ and $C_1^{(2)}, \ldots, C_{m_2}^{(2)}$, sampled from underlying probability distributions $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, respectively. While many two-sample tests focus on differences in means, variances, or other summary statistics, our emphasis lies on testing the equality of entire distributions by formulating the null hypothesis as $H_0 : \mathbb{P}^{(1)} = \mathbb{P}^{(2)}$. This type of test is particularly relevant for studies involving FC matrices, where the data are naturally grouped based on population characteristics such as disease status, cognitive phenotype, or experimental condition.

Central to this framework is the concept of measuring dissimilarity $\mathscr{D}$ between two probability distributions. This defines a class of two-sample testing algorithms that quantify the extent to which $\mathbb{P}^{(1)}$ differs from $\mathbb{P}^{(2)}$[46]. Commonly used measures in the context of correlation matrices under ECM and LEC geometries include:

- Maximum mean discrepancy (**corr_test2mmd.m**) measures differences between distributions in a reproducing kernel Hilbert space, leveraging kernel-based representations of data.
- Wasserstein distance (**corr_test2wass.m**) captures the minimal cost of transforming one distribution into another, often referred to as the "earth mover's distance."
- Energy distance (**corr_test2energy.m**) computes pairwise distances between all samples, focusing on differences in inter-point relationships.

These measures share a crucial property: $\mathscr{D} \geq 0$, where equality implies that $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are indistinguishable under the specific discrepancy measure. Viewing the two sets of observations as empirical measures:

$$\mathbb{P}^{(1)} = \frac{1}{m_1}\sum_{i=1}^{m_1}\delta_{C_i^{(1)}} \quad \text{and} \quad \mathbb{P}^{(2)} = \frac{1}{m_2}\sum_{j=1}^{m_2}\delta_{C_j^{(2)}},$$

where δ represents a Dirac mass, the null hypothesis is equivalently tested by checking whether $\mathscr{D}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) = 0$.

An alternative perspective is provided by inter-point distance-based methods. Testing the equality of distributions can also be formulated by analyzing the distributions of distances $d(C^{(1)}, \tilde{C}^{(1)})$, $d(C^{(2)}, \tilde{C}^{(2)})$, and $d(C^{(1)}, C^{(2)})$, where $\tilde{X}$ represents an independent sample identically distributed as X. This approach underpins the Biswas-Ghosh test (**corr_test2bg.m**)[47].

Despite their theoretical soundness, these four tests face practical challenges. The limiting distributions of their test statistics are either unknown or are only known under restrictive assumptions, limiting their direct application in real-world scenarios. To address this, we adopt a permutation testing framework, a resampling-based approach that establishes a threshold for the test statistic by permuting class labels[48]. This method provides robust control over Type I error rates without requiring strong parametric assumptions.

We outline a generic pipeline for the resampling procedure, applicable to all four tests introduced in this paper. Let $\mathscr{C}_i = \{C_1^{(i)}, \ldots, C_{m_i}^{(i)}\}$ represent the two samples for $i = 1, 2$, and let $\mathscr{C} = \mathscr{C}_1 \cup \mathscr{C}_2$ denote the combined dataset with a total size of $m_1 + m_2$. Denote $T(\cdot, \cdot)$ as the mechanism for computing the test statistic for one of the four tests. The pipeline proceeds as follows:

1. Calculate the observed test statistic $\hat{T}_{m_1,m_2} = T(\mathscr{C}_1, \mathscr{C}_2)$ for the original data.
2. For $n = 1, \ldots, N$ iterations:
   - Randomly permute the combined dataset $\mathscr{C}$.
   - Assign m_1 observations to $\mathscr{C}_1^{(n)}$ and the remaining m_2 observations to $\mathscr{C}_2^{(n)}$.
   - Compute the test statistic $T^{(n)} = T(\mathscr{C}_1^{(n)}, \mathscr{C}_2^{(n)})$.
3. Calculate the permutation p-value using:

$$\hat{p} = \frac{1}{N+1}\left(\sum_{n=1}^{N} I(\hat{T}_{m_1,m_2} \leq T^{(n)}) + 1\right),$$

where $I(\cdot)$ is the indicator function.

Once a significance level $\alpha \in (0, 1)$ is specified, the test based on permutation rejects the null hypothesis of equal distributions if $\hat{p} \leq \alpha$. This approach provides strong theoretical guarantees for controlling false positive rates[49,50].

Permutation-based testing has some advantages. First, it is distribution-free and avoids reliance on asymptotic approximations, making it suitable for small sample sizes or data with non-standard distributions. Second, the framework is easily adaptable to different test statistics and discrepancy measures. It further accommodates complex data structures such as correlation matrices on $\mathscr{C}_{++}^n$.

Results

In this section, we evaluate the proposed methods on real-world neuroimaging tasks, including behavioral score prediction, subject fingerprinting, and two-sample hypothesis testing for group differences. These tasks represent typical use cases in the field, where efficient processing and analysis of FC data are essential. All computations were performed on a consumer-grade laptop (MacBook Air M1 with an 8-core CPU and 8GB of unified memory), illustrating the computational feasibility of our approach even on modest hardware. Benchmark simulation studies assessing the computational gain and estimation accuracy of centroid measures are provided in the Supplementary Information.

To demonstrate the utility and versatility of our framework, we employed two publicly available datasets spanning different imaging modalities. The first dataset was drawn from the 1200-subject release of the Human Connectome Project (HCP) database[51]. From this cohort, we selected 980 subjects (Age: 28.71 ± 3.71 years, range 22–37; Males: 460, Females: 520). Each subject completed two 15-minute resting-state fMRI recordings with left-to-right (LR) and right-to-left (RL) phase encoding, resulting in four sessions per subject. The time-series data were sampled at 0.72 Hz, with 1200 time points per session. We used the version of the extensively processed fMRI data where preprocessing followed the HCP minimal preprocessing pipeline and mapped the data onto cortical surfaces[52]. Additional cleaning was performed using the HCP ICA-FIX pipeline, which regresses out motion-related artifacts and noise components identified via independent component analysis (ICA)[53,54].

For network-level FC analysis, time series data were extracted using the Schaefer atlas[55], which parcellates the cortical surface into 300 regions of interest (ROIs). Principal component analysis (PCA) was applied to the time series within each ROI, with the first principal component used to summarize the BOLD signal in each region. Empirical correlation matrices based on this parcellation were frequently rank-deficient due to the high number of ROIs relative to the available time points. To mitigate this, we applied three covariance matrix estimators: (1) Oracle Approximating Shrinkage (OAS)[56], (2) Ledoit-Wolf (LW) Shrinkage[57], and (3) Ridge Estimation[58], incorporating a regularization term with $\tau = 1.0$, such that $\Sigma_{\text{Ridge}} = \Sigma_{\text{empirical}} + \tau \cdot I$. All resulting covariance matrices were normalized to adhere to the constraints of $\mathscr{C}_{++}^n$, ensuring unit diagonals.

The second dataset was the EEG motor movement and imagery dataset[59], available through the PhysioNet database[60]. This dataset comprises 64-channel EEG recordings from 109 participants, collected using the BCI2000 system[61]. After excluding six participants due to annotation errors, we retained data from 103 subjects. Participants completed motor execution and motor imagery tasks involving fists and feet movements across 14 experimental sessions. Neural activity was recorded at a sampling rate of 160 Hz. For this analysis, we selected a single participant (S001) and focused on the motor imagery tasks.

Preprocessing followed the pipeline outlined in our previous study[25]. First, 32 channels identified as 'bad' (e.g., flat signals or poor signal-to-noise ratios) were removed. A Butterworth IIR band-pass filter with cutoff frequencies at 7 Hz and 35 Hz was applied using a two-pass zero-phase method. The filtered signals were segmented into epochs from each stimulus onset to one-second post-stimulus, resulting in 161 temporal measurements per epoch. This process yielded 45 samples, with 21 corresponding to feet movements and 24 to fists. Empirical correlation matrices computed for this dataset were full-rank, eliminating the need for regularized correlation estimators.

Experiment 1. Predicting Behavior Score

The first experiment focuses on the predictive modeling of behavior scores using correlation matrices within a regression framework. Leveraging the rich data from the Human Connectome Project (HCP), we assessed the effectiveness of nonparametric regression models for correlation-valued covariates in predicting behavioral outcomes, with a particular emphasis on the Penn Matrix Test (PMAT). The PMAT24 is a cognitive assessment tool designed to measure abstract reasoning and problem-solving abilities, serving as a brief version of Raven's Progressive Matrices[62]. We selected two outcome variables from the dataset: the number of correct responses (PMAT_A_CR) and the total number of skipped items (PMAT_A_SI), which serve as indicators of fluid intelligence. After excluding three subjects due to missing scores, the final sample consisted of 977 individuals.

Before predictive modeling, we first took a look at potential heterogeneity within the population of correlation-based FCs. Specifically, we considered two groups of subjects stratified by their PMAT_A_CR scores, selecting individuals from the top and bottom 10% of the score distribution. For each subject, their FC matrix was estimated using the oracle-approximating shrinkage estimator, followed by normalization. In Figure 4, we present the Fréchet means of the correlation networks for the two groups. To highlight prominent connections, each network was binarized by retaining only the top 2% of edges with the largest absolute magnitudes. Additionally, we computed the elementwise difference matrix between the two group-level mean FCs. Rows, and equivalently columns, of this difference matrix were then sorted and grouped based on a granular parcellation of the cerebral cortex into seven functional networks[55]. Notably, this visualization reveals distinct patterns of connectivity differences at a macroscopic level, suggesting that functional architecture varies systematically with fluid intelligence.

We compared the performance of our proposed models against two widely used approaches in neuroimaging-based predictive modeling: connectome-based predictive modeling (CPM)[63] and least absolute shrinkage and selection operator (LASSO) regression[64]. Both methods aim to predict behavioral or cognitive outcomes using FC matrices derived from preprocessed fMRI data. For both CPM and LASSO, we optimized hyperparameters using five-fold cross-validation on the training set, which was constructed by randomly splitting the data into an 80% training subset and a 20% testing subset.

The CPM approach involves identifying significant brain connections by correlating them with the outcome variable, and grouping these connections into positive and negative networks. The summed strengths of these networks are used as predictors in a linear regression model to estimate the behavioral outcome. This model emphasizes interpretability, as the identified networks provide insights into connectivity patterns associated with the behavior in question.

LASSO regression, on the other hand, operates on a design matrix where each row represents the vectorized upper triangular part of a correlation connectome. By applying L_1-regularization, LASSO shrinks the coefficients of less relevant variables toward zero, effectively performing variable selection. The non-zero

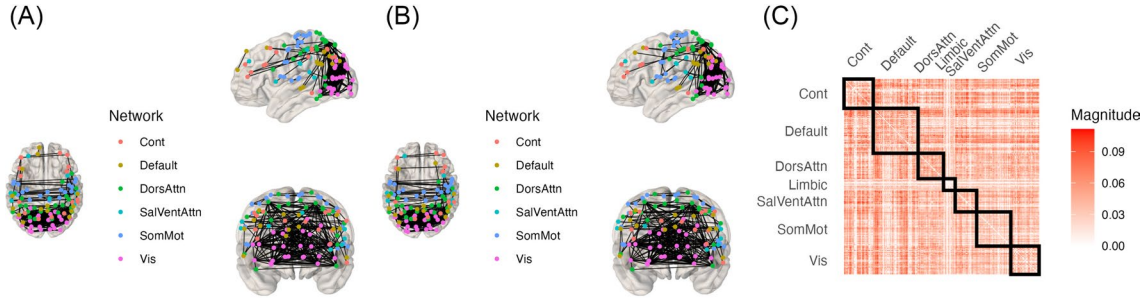

Fig. 4. Visualization of average correlation networks estimated using oracle-approximating shrinkage estimators. The Fréchet means of subjects with (**A**) the top 10% and (**B**) the bottom 10% of PMAT_A_CR scores were computed under the Euclidean-Cholesky metric. Each correlation matrix was binarized to retain only the top 2% of connections with the largest magnitudes. The elementwise difference matrix is shown in (**C**), where each entry represents the absolute difference between the two mean networks. Bounding boxes indicate the seven functional networks defined in[55].

coefficients correspond to the most influential features, which are then used to predict the behavioral outcome. Cross-validation ensures robust hyperparameter tuning and generalizability, with feature selection repeated independently for each fold.

Both CPM and LASSO offer advantages in linking FC patterns to behavior, particularly in terms of simplicity and interpretability. CPM highlights key networks by grouping edges into interpretable positive and negative categories, while LASSO identifies specific connections with predictive significance.

Table 2 summarizes the accuracy results for the predictive modeling task. It is immediately apparent that the nonparametric regression models leveraging the newly introduced geometries outperform the two competing models across all outcomes and estimators. This finding aligns with expectations, as nonlinear methods generally offer greater flexibility in capturing complex relationships between covariates and outcomes, albeit with a trade-off in interpretability. The results, derived from test data errors, provide empirical evidence supporting the superior predictive power of the proposed regression framework utilizing the two alternative geometries of $\mathscr{C}^n_{++}$.

A noteworthy observation is that CPM consistently demonstrated the poorest performance, even when compared to the basic LASSO model applied to half-vectorized covariates. For CPM, we employed default options for training and cross-validation and observed that the results varied significantly based on hyperparameter configurations. Methodologically, CPM identifies covariates highly correlated with the outcome and aggregates them into a single predictive variable. This approach assigns uniform weights to the selected variables while setting the coefficients of all others to zero, akin to a standard linear regression model. However, the effectiveness of this strategy becomes questionable when applied to the large number of highly correlated covariates ($\mathscr{O}(n^2)$) inherent in correlation matrices. This limitation likely stems from the well-documented challenges of constructing linear regression models with highly correlated variables, highlighting potential drawbacks of CPM in this context. We note that the 95% bootstrap confidence interval fo Table 2 is provided in the Supplementary Information.

Experiment 2. Fingerprinting

Next, we evaluate the effectiveness of the novel geometric structures in the task of functional connectome fingerprinting[65], which aims to capture individual variability in FC profiles. The fingerprinting task is formulated as follows: for M subjects, each undergoes two independent brain scan sessions, Session 1 and Session 2, producing corresponding FC representations. Given an individual's FC from Session 2, without identifying information, the goal is to determine which subject it belongs to by comparing it to the FCs from Session 1 based on a measure of similarity. This task can be viewed as a multiclass classification problem where each class contains a single observation. A 1-nearest neighbor (1-NN) classification method is naturally employed to assign the label of the most similar subject to the test sample. Identification accuracy I_{acc} is calculated as the number of correctly identified objects divided by the total number of subjects, and ranges from 0 to 1, with higher values indicating better identification accuracy.

In our experiment, we selected 100 unrelated subjects from the Human Connectome Project (HCP) dataset. Time series data were extracted from all four sessions of resting-state scans, and correlation-based FC representations were constructed for each session using three different estimators: the oracle approximating shrinkage (OAS) estimator, the Ledoit-Wolf (LW) estimator, and the L_2-regularized Ridge estimator. This procedure resulted in four distinct FCs for each subject per estimator. The experimental design considered 12 combinations of session pairs, where one session served as the training data and another as the test data. To streamline the reporting, symmetric pairs, such as (Session 1, Session 2) and (Session 2, Session 1), were treated as equivalent. Their identification accuracy scores were averaged, reducing the number of reportable cases to six.

For comparison, we employed the similarity-based fingerprinting approach proposed by[65]. This method identifies the subject corresponding to a query by finding the individual with the maximum Pearson correlation coefficient between vectorized FCs. From a machine learning perspective, this approach aligns with a 1-NN classification model where vectorized FCs serve as the data and Pearson correlation defines the similarity metric. In our adaptation, we retained the identical experimental pipeline but replaced the definition of data and

Outcome	Correlation estimator	Geometry						LASSO Regression	CPM
		ECM			LEC				
		GP	KERN	SVR	GP	KERN	SVR		
PMAT_A_CR	LW	4.50	4.60	4.41	4.50	4.60	4.44	13.29	32.33
	OAS	4.50	4.56	4.50	4.50	4.43	4.45	12.58	28.00
	Ridge	4.50	4.56	4.50	4.50	4.59	4.50	9.41	28.73
PMAT_A_SI	LW	3.69	3.90	3.66	3.69	3.79	3.69	5.46	28.36
	OAS	3.69	3.91	3.69	3.69	4.13	3.69	11.58	27.06
	Ridge	3.69	3.75	3.69	3.69	3.82	3.69	8.57	26.53

Table 2. Accuracy of fluid intelligence prediction. For each setting, the mean squared error (MSE) between the predicted and actual scores on the test data is reported. Across two correlation geometries - the Euclidean-Cholesky metric (ECM) and the Log-Euclidean Cholesky metric (LEC) - three regression models (Gaussian process regression [GP], kernel regression [KERN], and support vector regression [SVR]) were applied to correlation-valued functional connectomes estimated using the Ledoit-Wolf estimator (LW), the oracle approximating shrinkage estimator (OAS), and the L_2-regularized estimator (Ridge) with a penalty $\tau = 1$.

similarity with the correlation-based FCs and the geometric structures introduced previously. This modification allowed us to directly assess how the proposed geometries impact the accuracy of functional connectome fingerprinting, providing insights into their utility for capturing individual-specific patterns in FC.

Figure 5 summarizes the experimental results. The patterns observed across different pairs of runs are consistent, indicating low heterogeneity in the relative association of FCs across sessions for identification purposes. Across all estimators, the baseline method outperformed naive identification based on the Euclidean distance between FCs, which aligns with prior expectations and the findings of[65]. Notably, the incorporation of appropriate geometric structures into the space of FCs significantly enhanced performance. While the ECM geometry provided marginal improvements (except in the case of the Ridge estimator), the LEC geometry demonstrated substantial gains over the baseline method, with identification rates rising from below 75% to approximately 90% for both the OAS and LW estimators and even higher for the Ridge estimator.

These findings offer strong empirical support for the effectiveness of novel geometries on the correlation manifold in achieving fine-grained classification tasks. Additionally, it is worth noting the near-optimal performance of AIRM, which was comparable to LEC. AIRM, as a geometric structure on $\mathcal{S}_{++}^{n}$, serves as a valid distance metric for correlation matrices and has been shown to perform well in similar tasks[17]. In this study, AIRM's performance closely approached that of LEC, differing only slightly. However, this does not diminish the importance of our proposed framework. For example, AIRM does not preserve the correlation structure during operations such as mean computation, which limits its utility in specific scenarios compared to the proposed geometries.

Experiment 3. Hypothesis Testing

The final experiment focuses on two-sample hypothesis testing to determine whether two sets of correlations share the same underlying distribution, using an EEG dataset as the basis for analysis. For each 32-channel signal, after removing any bad channels, three types of FC representations were computed: the LW estimator, the OAS estimator, and the sample correlation matrix (SCM). After normalization, these representations served as inputs for the hypothesis testing framework, enabling a robust assessment of distributional equivalence between the two sets of correlations.

Before conducting the hypothesis testing, we visualized the data distributions by projecting them into a two-dimensional space using various geometries and algorithms, as illustrated in Figure 6. The resulting visualizations reveal noticeable differences in the data distribution shapes depending on the chosen geometry. With the ECM geometry, all estimators exhibited some degree of separation between the two classes, suggesting distinguishable patterns in the data. In contrast, visualizations based on the LEC geometry showed nearly overlapping distributions for the two sets of correlations, indicating reduced separability. This stark contrast between the geometries underscores their significant influence on data representation and the subsequent interpretability of results.

Next, we performed hypothesis testing to evaluate the equality of distributions using three proposed tests under both geometries for all considered estimators. The number of resampling iterations was set to $10^4 - 1$, which is adequate for the size of our dataset. Table 3 summarizes the attained empirical p-values after the false discovery rate correction[66]. Notably, the Biswas-Ghosh test rejected the null hypothesis of equal distributions for all estimators under the ECM geometry, whereas the LEC geometry showed no statistically significant differences. This result aligns with the prior visualizations, reinforcing the conclusion that the choice of geometry strongly influences both low-dimensional embeddings and statistical outcomes.

For the other tests, based on MMD and Wasserstein distance, the results were mixed, with some cases rejecting the null hypothesis and others failing to do so. While this variability might appear unsatisfactory, it highlights critical considerations for practitioners. The MMD test, which relies on kernel methods, is sensitive

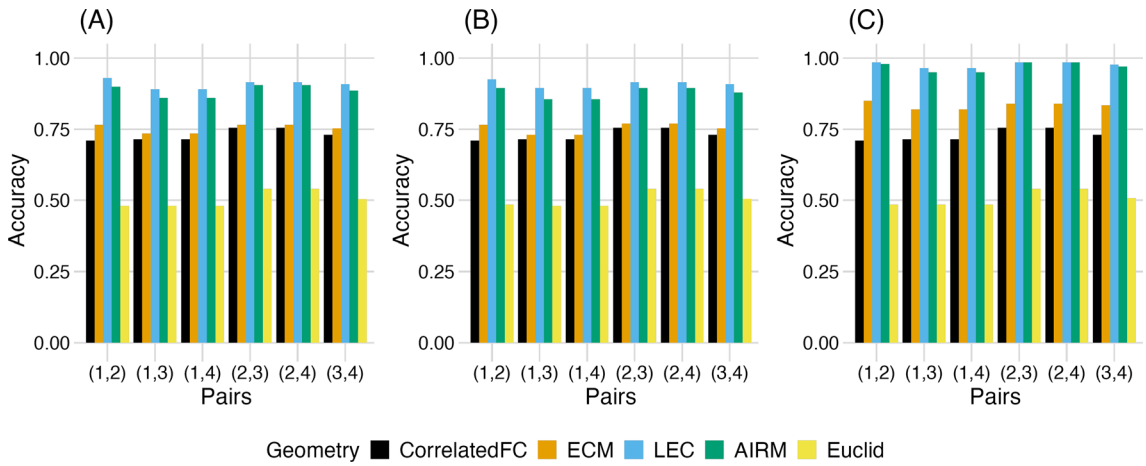

Fig. 5. Results for the fingerprinting example with different estimators: (**A**) Ledoit-Wolf (LW), (**B**) Oracle Approximating Shrinkage (OAS), and (**C**) Ridge. Identification rates for six pairs of runs are reported, where each 'pair' represents the average of two runs with flipped training and test dataset indices.

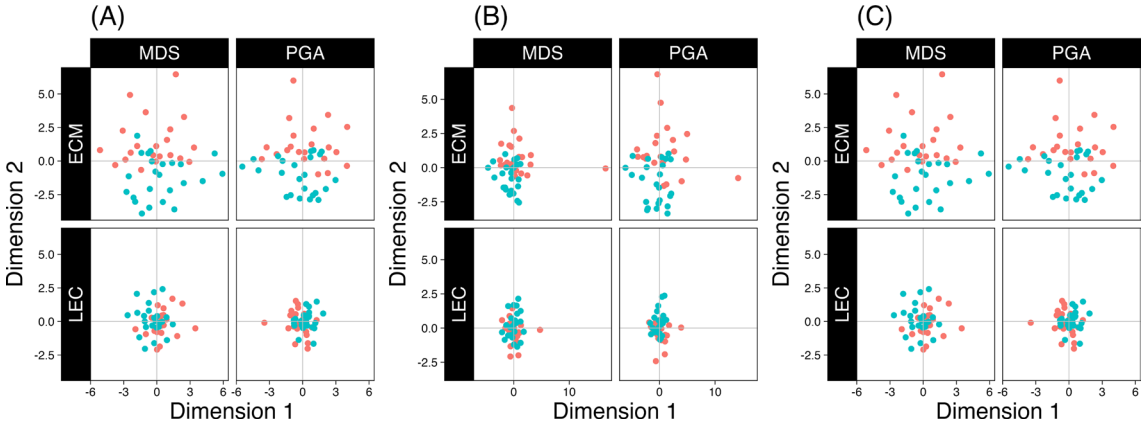

Fig. 6. Low-dimensional embedding of EEG hypothesis testing data with different estimators in $\mathbb{R}^2$: (**A**) Ledoit-Wolf (LW), (**B**) Oracle Approximating Shrinkage (OAS), and (**C**) empirical correlation matrix (SCM). Each subplot presents embeddings generated by a combination of different geometries and algorithms.

Geometry		**ECM**			**LEC**		
Estimator		**LW**	**OAS**	**SCM**	**LW**	**OAS**	**SCM**
Tests	Biswas-Ghosh	0.0335*	0.0394*	0.0335*	0.4794	0.3920	0.6650
	MMD	0.5993	0.3537	1	0.1800	0.3436	0.7204
	Wasserstein	0.0018**	0.0018**	0.0024**	0.0097**	0.0032**	0.0335*

Table 3. Adjusted empirical p-values from the hypothesis testing example with EEG data. (* $p < 0.05$, ** $p < 0.01$)

to the choice of kernel and its parameters. We used the squared exponential kernel, as discussed in Proposition 1, with a default parameter value of $\theta = 1$. This choice likely led to conservative results, as optimal performance requires careful tuning of θ, which controls the penalization of distant observations. Adjusting θ to better suit the data could improve the sensitivity of the test.

In contrast, the Wasserstein distance-based test rejected the null hypothesis for all combinations. However, the p-values obtained from the LEC geometry were notably higher than those from the ECM geometry, indicating that the test detected smaller discrepancies under the LEC geometry. The consistent rejection of the null hypothesis across all combinations may be attributed to challenges in estimating the Wasserstein distance in high-dimensional settings[67]. The EEG dataset used in this experiment consists of signals from 32 channels, resulting in an intrinsic dimensionality of 496 for the correlation FCs, while the sample size is only 45. This imbalance between dimensionality and sample size likely complicates the estimation process and hinders the ability to draw robust statistical inferences, as evidenced by the observed outcomes. For completeness, we also computed raw p-values and the ones adjusted by the Bonferroni correction in the Supplementary Information.

Discussion

The correlation matrix, a fundamental tool in functional network analysis, encapsulates collective information that extends beyond independent pairwise correlation coefficients. As such, treating it as a manifold-valued object with distinct geometric structures is a natural and advantageous perspective.

Our previous work[25] has integrated the quotient geometry of the correlation manifold into machine learning and statistical inference, enabling more robust FC analysis. However, challenges like computational inefficiency and instability in high-dimensional settings with many ROIs have limited its practical use.

To overcome these limitations, we introduced alternative geometric characterizations of the correlation manifold[26]. These alternatives offer dual advantages: they enable the application of well-established learning algorithms in traditional Euclidean settings and provide substantial computational benefits over the quotient geometry. Using these advancements, we implemented a suite of computational operations on the correlation manifold, including measures of central tendency, cluster analysis, hypothesis testing, and low-dimensional embedding. These tools are particularly advantageous for large-scale functional network analyses, where the brain is typically divided into hundreds of regions. Consequently, we proposed new techniques for FC and statistical learning aimed at population-level inference. The efficacy of these algorithms, grounded in the novel geometric structures, was validated using both simulated and real datasets, encompassing a variety of common neuroimaging analysis tasks. From a theoretical point, both ECM and LEC geometries employ diffeomorphism, which is a smooth bijection with a smooth inverse. This ensures that all statistical information encoded in the original correlation matrices, including higher-order interactions, remains fully recoverable.

Despite these contributions, some questions remain unresolved, opening avenues for future exploration. One critical issue is the selection of an appropriate geometry. While our study highlighted the comparative advantages of ECM and LEC geometries over QAM and other ambient geometries, no single geometry emerged as universally superior. In practical applications, it is often impossible to determine the optimal model in advance, and theoretical guarantees are lacking. This makes geometry selection a hyperparameter tuning problem, best addressed through data-driven approaches such as cross-validation. Another point of interest is its plausibility for deep learning. While deep architectures have been successfully adapted to the SPD space such as SPDNet[68] and graph neural networks[69], the diffeomorphic nature of the new geometries on the correlation manifold calls for a mathematical investigation into how standard neural network layers can be effectively extended to correlation matrices while preserving their geometric constraints.

Beyond the coverage of this paper, these contributions can be applied to a broader range of problems. For instance, multi-site harmonization of FC, recently approached from a Riemannian geometric perspective, may benefit from the computational efficiencies introduced by our framework. Techniques such as replacing centroids with the Fréchet mean and translation operations with parallel transport have been demonstrated on the SPD manifold[70,71]. Our work could facilitate the efficient application of such methods to correlation-valued FC. Furthermore, modeling continuous trajectories of dynamic FC[72], a task that usually relies on window-based segmentation, could be improved through manifold-valued regression.

The potential applications of these computational advancements extend well beyond neuroscience. In finance, for example, correlation matrices have long been used to model associations among asset returns, supporting tasks such as portfolio optimization, risk assessment, and change-point detection[73–75]. Climate science could benefit from geometric analysis of correlation matrices to study interdependencies among climate variables over spatial and temporal scales[76,77]. The development of scalable and efficient computational pipelines for correlation matrix analysis thus holds promise for diverse fields.

To encourage broader adoption and foster further exploration, we have consolidated all the algorithms discussed in this paper into a MATLAB toolbox, CORRbox, which is publicly available online. By providing an accessible and optimized platform, we aim to democratize the analysis of functional networks and inspire the integration of geometric approaches to correlation matrix analysis in a wide array of scientific disciplines. Future research could expand this foundation by exploring multi-modal integration, real-time analysis pipelines, and deeper theoretical characterizations of manifold structures to further enhance the capabilities and applications of these tools.

Data availability

The fMRI and EEG datasets used for the real data analysis are publicly available at https://www.humanconnectome.org/ and https://physionet.org, respectively. The CORRbox toolbox is openly accessible via GitHub at https://github.com/kisungyou/corrbox, along with illustrative examples.

Received: 11 April 2025; Accepted: 17 June 2025
Published online: 02 July 2025

References

1. Park, H.-J. & Friston, K. Structural and Functional Brain Networks: From Connections to Cognition. *Science* **342**, 1238411–1238411. https://doi.org/10.1126/science.1238411 (2013).
2. Biswal, B., Zerrin Yetkin, F., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine* **34**, 537–541, https://doi.org/10.1002/mrm.1910340409 (1995).
3. Brookes, M. J. et al. Measuring functional connectivity using MEG: Methodology and comparison with fcMRI. *NeuroImage* **56**, 1082–1104. https://doi.org/10.1016/j.neuroimage.2011.02.054 (2011).
4. Cohen, M. X. *Analyzing neural time series data: theory and practice* (Issues in clinical and cognitive neuropsychology (The MIT Press, Cambridge, Massachusetts, 2014).
5. Dosenbach, N. U. F. et al. Prediction of Individual Brain Maturity Using fMRI. *Science* **329**, 1358–1361. https://doi.org/10.1126/science.1194144 (2010).
6. Leonardi, N. et al. Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage* **83**, 937–950. https://doi.org/10.1016/j.neuroimage.2013.07.019 (2013).
7. Siman-Tov, T. et al. Early Age-Related Functional Connectivity Decline in High-Order Cognitive Networks. *Frontiers in Aging Neuroscience* **8**, https://doi.org/10.3389/fnagi.2016.00330 (2017).
8. Park, B., Kim, D.-S. & Park, H.-J. Graph Independent Component Analysis Reveals Repertoires of Intrinsic Network Components in the Human Brain. *PLoS ONE* **9**, e82873. https://doi.org/10.1371/journal.pone.0082873 (2014).
9. Bhatia, R. *Positive Definite Matrices* (Princeton University Press, 2009).
10. Wang, M., Wang, Y. & Yang, Y. Dynamic and low-dimensional modeling of brain functional connectivity on Riemannian manifolds. *NeuroImage* **314**, 121243. https://doi.org/10.1016/j.neuroimage.2025.121243 (2025).
11. Wang, T., Mao, R., Liu, S., Cambria, E. & Ming, D. Explainable multi-frequency and multi-region fusion model for affective brain-computer interfaces. *Information Fusion* **118**, 102971. https://doi.org/10.1016/j.inffus.2025.102971 (2025).
12. Li, W., Wang, M., Liu, M. & Liu, Q. Riemannian manifold-based disentangled representation learning for multi-site functional connectivity analysis. *Neural Networks* **183**, 106945. https://doi.org/10.1016/j.neunet.2024.106945 (2025).
13. Moghaddam, M. et al. Tangent space functional reconfigurations in individuals at risk for alcohol use disorder. *Network Neuroscience* **9**, 38–60. https://doi.org/10.1162/netn_a_00419 (2025).
14. Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P. & Thirion, B. Detection of Brain Functional-Connectivity Difference in Post-stroke Patients Using Group-Level Covariance Modeling. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, vol. 6361, 200–208 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).
15. Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S. & Kolaczyk, E. D. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics* **11**, https://doi.org/10.1214/16-AOAS1015 (2017).
16. Yamin, A. et al. Comparison Of Brain Connectomes Using Geodesic Distance On Manifold: A Twins Study. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1797–1800, https://doi.org/10.1109/ISBI.2019.8759407 (IEEE, Venice, Italy, 2019).

17. Abbas, K. et al. Geodesic Distance on Optimally Regularized Functional Connectomes Uncovers Individual Fingerprints. *Brain Connectivity* **11**, 333–348. https://doi.org/10.1089/brain.2020.0881 (2021).
18. You, K. & Park, H.-J. Re-visiting Riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage* **225**, 117464. https://doi.org/10.1016/j.neuroimage.2020.117464 (2021).
19. Tropp, J. A. Simplicial Faces of the Set of Correlation Matrices. *Discrete & Computational Geometry* **60**, 512–529. https://doi.org/10.1007/s00454-017-9961-0 (2018).
20. Grubišić, I. & Pietersz, R. Efficient rank reduction of correlation matrices. *Linear Algebra and its Applications* **422**, 629–653. https://doi.org/10.1016/j.laa.2006.11.024 (2007).
21. Nielsen, F. & Sun, K. Clustering in Hilbert's Projective Geometry: The Case Studies of the Probability Simplex and the Elliptope of Correlation Matrices. In Nielsen, F. (ed.) *Geometric Structures of Information*, 297–331 (Springer International Publishing, Cham, 2019).
22. David, P. *A Riemannian Quotient Structure for Correlation Matrices with Applications to Data Science*. PhD Thesis, Claremont Graduate University (2019).
23. Pennec, X. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision* **25**, 127–154. https://doi.org/10.1007/s10851-006-6228-4 (2006).
24. Thanwerdas, Y. & Pennec, X. Geodesic of the Quotient-Affine Metrics on Full-Rank Correlation Matrices. arXiv:2103.04621 [math] (2021). ArXiv: 2103.04621.
25. You, K. & Park, H.-J. Geometric learning of functional brain network on the correlation manifold. *Scientific Reports* **12**, 17752. https://doi.org/10.1038/s41598-022-21376-0 (2022).
26. Thanwerdas, Y. & Pennec, X. Theoretically and Computationally Convenient Geometries on Full-Rank Correlation Matrices. *SIAM Journal on Matrix Analysis and Applications* **43**, 1851–1872. https://doi.org/10.1137/22M1471729 (2022).
27. Hall, B. C. *Lie groups, Lie algebras, and representations: an elementary introduction*. No. 222 in Graduate texts in mathematics (Springer, Cham ; New York, 2015), second edition edn. OCLC: ocn910324548.
28. Afsari, B. Riemannian L_p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society* **139**, 655–655. https://doi.org/10.1090/S0002-9939-2010-10541-5 (2011).
29. Demmel, J. W. *Applied numerical linear algebra* (Society for Industrial and Applied Mathematics, Philadelphia, 1997).
30. Demmel, J., Dumitriu, I. & Holtz, O. Fast linear algebra is stable. *Numerische Mathematik* **108**, 59–91. https://doi.org/10.1007/s00211-007-0114-x (2007).
31. Fletcher, P. T., Venkatasubramanian, S. & Joshi, S. The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage* **45**, S143–S152. https://doi.org/10.1016/j.neuroimage.2008.10.052 (2009).
32. Schölkopf, B. & Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning series (MIT Press, Cambridge, Mass., 2002), reprint. edn.
33. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of machine learning* (Adaptive computation and machine learning series (MIT Press, Cambridge, MA, 2012).
34. You, K. & Shung, D. Rdimtools: An R package for dimension reduction and intrinsic dimension estimation. *Software Impacts* **14**, 100414. https://doi.org/10.1016/j.simpa.2022.100414 (2022).
35. Fletcher, P., Lu, C., Pizer, S. & Joshi, S. Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape. *IEEE Transactions on Medical Imaging* **23**, 995–1005. https://doi.org/10.1109/TMI.2004.831793 (2004).
36. Baldi, P. & Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks* **2**, 53–58. https://doi.org/10.1016/0893-6080(89)90014-2 (1989).
37. Borg, I. & Groenen, P. J. F. *Modern multidimensional scaling: theory and applications* (Springer series in statistics (Springer, New York, 1997).
38. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
39. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics (Wiley, Hoboken, N.J, 2005).
40. MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations. In Cam, L. M. L. & Neyman, J. (eds.) *Proc. of the fifth berkeley symposium on mathematical statistics and probability*, vol. 1, 281–297 (University of California Press, 1967).
41. Kaufman, L. & Rousseeuw, P. J. Partitioning Around Medoids (Program PAM). In *Wiley Series in Probability and Statistics*, 68–125, https://doi.org/10.1002/9780470316801.ch2 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1990).
42. von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17**, 395–416. https://doi.org/10.1007/s11222-007-9033-z (2007).
43. Eaves, L. C., Ho, H. H. & Eaves, D. M. Subtypes of autism by cluster analysis. *Journal of Autism and Developmental Disorders* **24**, 3–22. https://doi.org/10.1007/BF02172209 (1994).
44. Rousseeuw, P. J. & Leroy, A. M. *Robust regression and outlier detection* (Wiley, New York, 1987). OCLC: 219924217.
45. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* **3**, 1–27. https://doi.org/10.1080/03610927408827101 (1974).
46. Ramdas, A., Trillos, N. & Cuturi, M. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy* **19**, 47. https://doi.org/10.3390/e19020047 (2017).
47. Biswas, M. & Ghosh, A. K. A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160–171. https://doi.org/10.1016/j.jmva.2013.09.004 (2014).
48. Pitman, E. J. G. Significance Tests Which May be Applied to Samples From any Populations. *Supplement to the Journal of the Royal Statistical Society* **4**, 119. https://doi.org/10.2307/2984124 (1937).
49. Romano, J. P. & Wolf, M. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association* **100**, 94–108. https://doi.org/10.1198/016214504000000539 (2005).
50. You, K., Kim, I., Jin, I. H., Jeon, M. & Shung, D. Comparing multiple latent space embeddings using topological analysis, https://doi.org/10.48550/ARXIV.2208.12435 (2022). Publisher: arXiv Version Number: 1.
51. Van Essen, D. et al. The Human Connectome Project: A data acquisition perspective. *NeuroImage* **62**, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018 (2012).
52. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127 (2013).
53. Salimi-Khorshidi, G. et al. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90**, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046 (2014).
54. Griffanti, L. et al. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage* **95**, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034 (2014).
55. Schaefer, A. et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex* **28**, 3095–3114. https://doi.org/10.1093/cercor/bhx179 (2018).
56. Chen, Y., Wiesel, A., Eldar, Y. C. & Hero, A. O. Shrinkage Algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing* **58**, 5016–5029. https://doi.org/10.1109/TSP.2010.2053029 (2010).
57. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4 (2004).

58. Mejia, A. F. et al. Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. *NeuroImage* **172**, 478–491. https://doi.org/10.1016/j.neuroimage.2018.01.029 (2018).
59. Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N. & Wolpaw, J. R. EEG Motor Movement/Imagery Dataset, https://doi.org/10.13026/C28G6P (2009).
60. Goldberger, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101**, https://doi.org/10.1161/01.CIR.101.23.e215 (2000).
61. Schalk, G., McFarland, D., Hinterberger, T., Birbaumer, N. & Wolpaw, J. BCI2000: A General-Purpose Brain-Computer Interface (BCI) System. *IEEE Transactions on Biomedical Engineering* **51**, 1034–1043. https://doi.org/10.1109/TBME.2004.827072 (2004).
62. Bilker, W. B. et al. Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test. *Assessment* **19**, 354–369. https://doi.org/10.1177/1073191112446655 (2012).
63. Shen, X. et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols* **12**, 506–518. https://doi.org/10.1038/nprot.2016.178 (2017).
64. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58**, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (1996).
65. Finn, E. S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience* **18**, 1664–1671. https://doi.org/10.1038/nn.4135 (2015).
66. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
67. Panaretos, V. M. & Zemel, Y. *An Invitation to Statistics in Wasserstein Space* (SpringerBriefs in Probability and Mathematical Statistics (Springer International Publishing, Cham, 2020).
68. Huang, Z. & Gool, L. V. A riemannian network for SPD matrix learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 2036–2042 (AAAI Press, 2017). Place: San Francisco, California, USA Number of pages: 7.
69. Ju, C. & Guan, C. Graph Neural Networks on SPD Manifolds for Motor Imagery Classification: A Perspective From the Time-Frequency Analysis. *IEEE Transactions on Neural Networks and Learning Systems* **35**, 17701–17715. https://doi.org/10.1109/TNNLS.2023.3307470 (2024).
70. Simeon, G., Piella, G., Camara, O. & Pareto, D. Riemannian Geometry of Functional Connectivity Matrices for Multi-Site Attention-Deficit/Hyperactivity Disorder Data Harmonization. *Frontiers in Neuroinformatics* **16**, 769274. https://doi.org/10.3389/fninf.2022.769274 (2022).
71. Honnorat, N. et al. Riemannian frameworks for the harmonization of resting-state functional MRI scans. *Medical Image Analysis* **91**, 103043. https://doi.org/10.1016/j.media.2023.103043 (2024).
72. Preti, M. G., Bolton, T. A. & Van De Ville, D. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage* **160**, 41–54. https://doi.org/10.1016/j.neuroimage.2016.12.061 (2017).
73. Mantegna, R. Hierarchical structure in financial markets. *The European Physical Journal B* **11**, 193–197. https://doi.org/10.1007/s100510050929 (1999).
74. Bonanno, G., Caldarelli, G., Lillo, F. & Mantegna, R. N. Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review E* **68**, 046130. https://doi.org/10.1103/PhysRevE.68.046130 (2003).
75. Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J. & Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E* **68**, 056110. https://doi.org/10.1103/PhysRevE.68.056110 (2003).
76. Braunisch, V. et al. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography* **36**, 971–983. https://doi.org/10.1111/j.1600-0587.2013.00138.x (2013).
77. Runge, J., Petoukhov, V. & Kurths, J. Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models. *Journal of Climate* **27**, 720–739. https://doi.org/10.1175/JCLI-D-13-00159.1 (2014).

Acknowledgements

K.Y. was supported by a PSC-CUNY Award (TRADB-55-511), jointly funded by The Professional Staff Congress and The City University of New York. H-J. P. was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00401794).

Author contributions

K.Y. conceptualized the study, developed the methodology and software, and performed data analysis. Y.L. conducted data preparation. H.J.P. initiated the project, contributed to data preparation, and co-conceptualized the methodology. K.Y. and H.J.P. wrote the main manuscript text. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-07703-1.

Correspondence and requests for materials should be addressed to H.-J.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.