LOCALIZING DIFFERENTIALLY EVOLVING COVARIANCE STRUCTURES VIA SCAN STATISTICS

RONAK MEHTA,

Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

HYUNWOO J. KIM,

Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

SHULEI WANG,

Department of Statistics, Columbia University, New York, NY 10027, USA

STERLING C. JOHNSON,

Alzheimer's Disease Research Center, University of Wisconsin-Madison, Madison, WI 53792, USA

MING YUAN,

Department of Statistics, Columbia University, New York, NY 10027, USA

VIKAS SINGH

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, USA

Abstract

Recent results in coupled or temporal graphical models offer schemes for estimating the relationship structure between features when the data come from related (but distinct) longitudinal sources. A novel application of these ideas is for analyzing group-level differences, i.e., in identifying if *trends* of estimated objects (e.g., covariance or precision matrices) are different across disparate conditions (e.g., gender or disease). Often, poor effect sizes make detecting the *differential* signal over the *full* set of features difficult: for example, dependencies between only a *subset of features* may manifest differently across groups. In this work, we first give a parametric model for estimating trends in the space of SPD matrices as a function of one or more covariates. We then generalize scan statistics to graph structures, to search over distinct subsets of features (graph partitions) whose temporal dependency structure may show statistically significant group-wise differences. We theoretically analyze the Family Wise Error Rate (FWER) and bounds on Type 1 and Type 2 error. Evaluating on US census data, we identify groups of states with cultural and legal overlap related to baby name trends and drug usage. On a cohort of individuals with risk factors for Alzheimer's disease (but otherwise cognitively healthy), we find scientifically interesting group differences where the default analysis, i.e., models estimated on the *full graph*, do not survive reasonable significance thresholds.

ronakrm@cs.wisc.edu .

1. Introduction.

Multivariate data analysis exploiting the conditional independence structure between features or covariates using undirected graphical models is now standard within any data analysis toolbox. When the data are multivariate Gaussian, the zeros in the inverse covariance (precision) matrix give conditional independences among the variables [36]. Further, if the precision matrix is sparse, we can derive dependencies between features when the data are high-dimensional and/or the number of measurements are small. The estimation of a graphical model has been extensively studied and a rich literature is available describing its statistical and algorithmic properties [34, 31]. For instance, the so-called *graphical lasso* formulation uses an l_1-norm penalty on the precision matrix and is widely used, and consistency properties in the large p regime [7, 19, 73] are now well understood. These formulations have also been extended to various transformations of Gaussian distributions (e.g., non-paranormal) using rank statistics [41, 71, 42].

Coupled and Temporal Graphical Models.

Often, data come from two (or more) disparate sources or multiple timepoints. Within the last few years, a few proposals have described strategies for linking the sparsity patterns of multiple graphical models, e.g., using a fused lasso penalty [10] [72]. Observe that if the data sources correspond to *longitudinal* acquisitions, we should expect the 'structure' to gradually evolve. Several authors have offered generalizations to address this problem: [75] removes the assumption that each graph is independent and structurally 'close'. Instead, [75] can be thought of as a growth model [43] defined on these structures: they show how non-identically distributed graphs can be learned over time. Recently, the nonparametric procedure in [48] extends these ideas to handle multiple sources, each with multiple samples.

The ideas in the literature so far to "couple" multiple graphical model estimation modules are mostly nonparametric. While such a formulation offers benefits, in many estimation problems, parametric models may be more convenient for downstream statistical analysis, particularly for hypothesis testing [24, 20, 51]. Given that the topic of *coupled* graphical models, by itself, is fairly recent, algorithms for *parametric estimation* of temporal or coupled Gaussian graphical models have not yet been heavily studied. This will involve parameterizing *trends* in the highly structured nature of the 'response' variable (SPD matrices). We find that parametric formulations for manifold-valued data *have* been proposed recently [33, 9]. Because SPD matrices form a Riemannian manifold, algorithms that estimate a parametric model respecting the underlying Riemannian metric are more suitable in many applications as opposed to assuming a Euclidean metric on positively or negatively curved spaces [70, 17, 29]. We will make a few simple modifications (for efficiency purposes) to such algorithms and make use of the estimated parameters for follow-up analysis.

Finding Group-wise Differences.

Assuming that we have a black-box procedure to estimate a parametric model on the SPD manifold available, in many tasks, such an estimation is merely a segue to other analyses

designed to answer scientifically meaningful questions. For example, we are often interested in asking whether the temporally coupled model estimated using the procedure above differs in meaningful ways *across* groups induced by a stratification or dichotomous variable (e.g., gender or disease). For instance, is the 'slope' in structured response space statistically different across education level or body mass index? While the body of work for graphical model estimation is mature, the literature describing hypothesis tests in this regime [58, 6] is sparse at best. Given that such questions are simpler to answer with alternative schemes (with assumptions on the distributional properties of the data), e.g., structural equation modeling, latent growth models and so on [64, 43], it seems that the unavailability of such tools is limiting the adoption of such ideas in a broader cross-section of science. We will seek to address this gap.

Needles in Temporal Haystacks.

If we temporarily set aside the potential value of a hypothesis test framework for temporal trajectories in graphical models, we see that from an operational viewpoint, such procedures are most effective when a practitioner already has a precise scientific question in mind. In reality, however, many data analysis tools are deployed for exploratory analyses to inform an investigator as to which questions to ask. Being able to "localize" which parts of the model are different across groups over the entire time window can be very valuable. This ability actually benefits statistical power as well. Notice that when the stratified groups are not very different to begin with, e.g., healthy individuals with presence or absence of a genetic mutation, the effect sizes are likely to be poor. Here, while the trends identified on the *full* precision matrix may still be different (i.e., there may be a *real* signal associated with a grouping variable), they may not be strong enough to survive significance thresholds. Ideally, what we need here are analogs of the widely used "scan statistics" for our hypothesis testing formulations for temporal graphical models — to identify which *parts of the signal* are promising. Then, even if only a small subset of features were different across groups over all time, we may be able to identify these differential effects efficiently. This benefits Type 2 error, provides a practical turnkey product for an experimental scientist, and makes up the key technical results of our work.

Foundations of our work can be traced back to fundamental developments made by Ulf Grenander in a breadth of fields. Early work with Rosenblatt on the analysis of stochastic processes and time series first brought to light the fundamental issues of linear modeling in Euclidean space, and demonstrated that in many cases it is necessary to develop methods that take explicit advantage of the inherent structure within data [23]. Further pioneering work on the statistical analysis on Lie groups [21] provides the basis of the Riemannian statistics mentioned above. Modern hypothesis testing of these structured, manifold-valued data in image analysis is built upon the his joint work [22]. Here, we marry modern developments in these areas, using recent strides in linear model fitting on manifolds and statistical testing of structured data to develop groupwise testing procedures for longitudinal covariances. Concurrent to our work, [59, 74] have developed similar methods of analyzing the statistical properties of trajectories on the SPD(n) manifold via the transported square-root vector field. While here we focus on a simple approach to enable localization, these developments can be incorporated into our construction.

Briefly, we provide **(i)** a simple and efficient parametric procedure for modeling temporally evolving graphical models, **(ii)** a hypothesis test for identifying differences between group-wise estimated models, and **(iii)** a scan algorithm to identify *those subsets of the features which contribute to the group-wise differences*. Together, these ideas offer a framework for identifying group-wise differences in temporally coupled graphical models. From the experimental perspective, we find scientifically plausible results on a unique longitudinally tracked cohort of middle-aged (and young elderly) persons at risk for Alzheimer's disease due to family history, but who are otherwise completely cognitively healthy.

The rest of the paper is organized as follows. In Section 2 we present an efficient manifold regression procedure for modeling covariance trajectories, which serves as a blackbox module in our hypothesis testing framework. In Section 3, we define our main hypothesis test for group difference analysis over covariance trajectories. In Section 4, we present a set of technical results describing our localization procedure based on scan statistics, as well as derive suitable size corrections to compare across feature subsets. Sections 5, 6, and 7 conclude with empirical evaluations of our model on synthetic data, various types of demographics/behavior data collected longitudinally in the United States from publicly available resources, and finally, our analysis on a unique longitudinal dataset (followed since 2001) from a preclinical Alzheimer's disease study involving approximately 1500 individuals.

2. Characterizing Covariance Trajectories.

Our main statistical testing framework, to be described shortly, needs an efficient means for calculating a "trajectory" of the feature-by-feature interaction graphs over time for the given longitudinal data. We now describe a scheme which offers this capability. Let $X_t \in \mathbb{R}^{n_t, p}$ be the design matrix of all n_t samples at time t, where $t \in \{1,...,T\}$, and T is the total number of distinct timepoints. We wish to capture the trends in the relationships between the features as a function of t. To evaluate the groupwise differences in changes of such interactions, we make use of the fact that these interactions are commonly captured by correlation or conditional independence, represented by the covariance matrix (with normalized features) and the precision matrix (the inverse of covariance matrix).

Here we simply use the covariance matrix for each timepoint t to denote the interaction between features, $C_t = cov(X_t)$. Our goal now is to estimate the parameters of the function, $t \rightarrow C_t$. We may vectorize the covariance matrix and apply a linear model; its parameters will give the trajectory in "vectorized covariance space" as we scan through t. But these predictions are *not* guaranteed to be valid SPD matrices and even if a projection is performed to obtain a covariance estimate, distortions introduced by the process may be significant [16]. It is well known that classical vector space models tend to be suboptimal in the manifold setting (covariance matrices live on the SPD manifold) since they use Euclidean metrics which are defined in the ambient space. For manifold-valued data, Riemannian metrics are shown to be superior in many applications [17, 5, 29, 63], and are increasingly being deployed in machine learning/statistics. We will utilize an appropriate statistical model informed by the manifold-structure of the data and then derive a hypothesis testing procedure to detect groupwise difference in the changes of interactions between

features in longitudinal analysis. To do so, we first summarize basic differential geometry notations [12, 38] and then describe our models. If desired, any other (efficient) manifold-valued linear model [16] can be substituted in; no change in the workflow is needed. A reader familiar with manifold regression algorithms may consider this module as a black-box and skip ahead to Section 3 which uses the parameter estimates from this procedure.

2.1. Riemannian Geometry.

Let $\mathcal{M}$ be a *differentiable (smooth) manifold* in arbitrary dimensions. A differentiable manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure. A *Riemannian manifold* is a differentiable manifold $\mathcal{M}$ equipped with a smoothly varying inner product. The *geodesic curve* is the locally shortest path, analogous to straight lines in $\mathcal{R}^p$ — this geodesic curve will be the object that defines the trajectory of our covariance matrices in SPD space. Unlike the Euclidean space, note that there may exist multiple geodesic curves between two points on a curved manifold. So, the *geodesic distance* between two points on $\mathcal{M}$ is defined as the length of the *shortest* geodesic curve connecting two points. The geodesic distance helps in measuring the error of our trajectory estimation (analogous to a Frobenius or ℓ_2 norm based loss in the Euclidean setting). The geodesic curve from y_i to y_j is parameterized by a tangent vector in the tangent space anchored at y_i with an exponential map $\mathrm{Exp}(y_i, \cdot): T_{y_i}\mathcal{M} \to \mathcal{M}$. The inverse of the exponential map is the logarithm map, $\mathrm{Log}(y_i, \cdot): \mathcal{M} \to T_{y_i}\mathcal{M}$. These two operations move us back and forth between the manifold and the tangent space. For completeness, Table 1 shows corresponding operations in the Euclidean space and Riemannian manifolds. Separate from the above notation, matrix exponential (and logarithm) are simply $\exp(\cdot)$ (and $\log(\cdot)$). Finally, *parallel transport* is a generalized parallel translation on manifolds. Given a differentiable curve $\gamma: \mathcal{I} \to \mathcal{M}$, where $\mathcal{I}$ is an open interval, the parallel transport of $v_0 \in T_{\gamma(t_0)}\mathcal{M}$ along curve γ can be interpreted as the parallel translation of v_0 on the manifold preserving its length and the angle between $v(t)$ and γ. The parallel transport of v from y to y' is $\Gamma_{y \to y'}v$.

2.2. Riemannian Manifold Regression.

Several regression models for manifold-valued data have been proposed recently, a majority of which are nonparametric [29, 5]. Because of the longitudinal nature of our dataset (and recruitment considerations in neuroimaging studies), sample sizes do not exceed a few hundred participants (typically much smaller). We have found that generally, in this regime, parametric methods are better suited and also offer other benefits for downstream applications. Next, we will give a simple parametric model for this problem. Let x and y be vectors in $\mathbb{R}^p$ and $\mathbb{R}^p$, respectively.

DEFINITION 2.1.—(Standard GLM.) The Euclidean multivariate multilinear model is

$$y = \beta^0 + \beta^1 x^1 + \beta^2 x^2 + \ldots + \beta^p x^p + \epsilon \tag{2.1}$$

where β^0, β^i and the error ϵ are in $\mathbb{R}^{p'}$ and $x = [x^1 \ldots x^p]^T$ are the predictor variables.

Henceforth, we will use the terms *covariate* and *predictor* interchangeably to describe those specific features we wish to control for in our model (e.g., time-points in our experiments). For manifold-valued data, we adapt the formulation proposed by [33].

DEFINITION 2.2.—The Manifold Multivariate General Linear Model (MMGLM) is defined as

$$\min_{b \in \mathcal{M},\, \forall j,\, V^j \in T_b \mathcal{M}} \frac{1}{2} \sum_{i=1}^{N} d(\mathrm{Exp}(b, \boldsymbol{V} x_i), y_i)^2, \qquad (2.2)$$

where $\boldsymbol{V} x_i := \sum_{j=1}^{n} V^j x_i^j$ and $d(\cdot, \cdot)$ is the geodesic distance between $\hat{y}_i := \mathrm{Exp}(b, \boldsymbol{V} x_i)$ and y_i.

This formulation generalizes (2.1), by replacing the intercept β^0 and each vector β^j for a covariate with a base point $b \in \mathcal{M}$ and a geodesic basis $V^j \in T_b \mathcal{M}$ respectively. The geodesic basis V^j at b parameterizes a geodesic curve $\mathrm{Exp}(b, V^j x^j)$. Intuitively, this model is a 'generalized' linear model with the inverse exponential map Exp^{-1} (or logarithm map Log) as a 'link' function [33, 9]. When the covariate/predictors are univariate, we will obtain a single geodesic curve, modeled via the so-called Geodesic Regression [16].

2.3. Efficient Estimation of Trajectories.

The objective in (2.2), can be solved by both gradient descent [16, 33] and MCMC methods [9]. Unfortunately, these schemes can be expensive, especially when the dimension of the manifold is large. Further, if the algorithm needs to be run a large number of times, the computational footprint quickly becomes prohibitive. Motivated by these considerations, we use a so-called log-Euclidean approximate algorithm introduced in [33] with some adaptations, which requires mild assumptions on the manifold-valued data.

Recall that in classical ordinary least squares (OLS), the regression curve goes through the mean of covariates and response variables, i.e., $y - \bar{y} = \beta(x - \bar{x})$. Similarly, we assume that geodesic curves go through the mean of response variables on the manifold. Then, the base point, or intercept, "b" in (2.2) can be approximated by the *manifold-valued mean of the sample points*, the Karcher mean [32]. The propositions derived from [33] lead directly to the following.

PROPOSITION 2.3.—Let $\bar{C}$ be the unique Karcher mean of a sufficiently close set of covariance matrices that lie on a curve Ω. Then $\bar{C} \in \Omega$, and for some tangent vector $V \in T_{\bar{C}} \mathcal{M}$ and each C, there exists $x \in \mathbb{R}$ such that $C = \mathrm{Exp}(\bar{C}, V x)$.

This allows us to bypass the fairly involved variational procedure to estimate the base point b.

With this approximation of $\hat{b}$ via $\bar{y}$, the remaining variables to optimize are the tangent vectors $\boldsymbol{V}$. We do so by taking advantage of log-Euclidean schemes. Once the base point is established as the Karcher mean, each data point on the manifold is projected into the tangent space at that point: $\mathrm{Log}(\bar{y}, y)$. These "centered" points $\tilde{y}$ are now Euclidean, and if

the covariates are centered as well ($\tilde{x}$), a closed form solution exists in the standard form of $V = \tilde{y}\tilde{x}^\top(\tilde{x}\tilde{x}^\top)^{-1}$.

In this setting, it is often assumed that two points y_1, y_2 have a distance defined as $d(y_1, y_2) := \|\text{Log}(y_1, y_2)\|_{y_1} \approx \|\text{Log}(b, y_1) - \log(b, y_2)\|_b$. However, on SPD manifolds with an affine invariant metric, each tangent space has a different inner product varying as a function of the base point b, i.e., $\langle u, v \rangle_b := \text{tr}(b^{-1/2}ub^{-1}vb^{-1/2})$. This makes comparison of trajectories difficult without moving to tangent bundle formulations. This issue is discussed in some detail in [46, 27]. However, note that

REMARK 2.4.—When the base point b is the identity I, then the inner product is exactly the Euclidean metric $\langle u, v \rangle_b := \text{tr}(b^{-1/2}ub^{-1}vb^{-1/2}) = \text{tr}(uv) = \text{tr}(u^T v)$.

This follows from the fact that u and v are symmetric matrices on SPD(p). We take advantage of this property through *parallel transport*. Specifically, we can bring all of the data to $T_I\mathcal{M}$ which will allow for a meaningful comparison of two tangent vectors from different base points. Similar schemes have been used for projection on submanifolds in [70] and other problems [55]. With a fast algorithm to compute (2.2) available, we can now accurately model longitudinal trajectories of covariances matrices. Our statistical procedure described next simply assumes the availability of some suitable scheme to solve the manifold-regression as defined in (2.2) efficiently and does not depend on particular properties of the foregoing algorithm.

3. Test Statistics for SPD(p) Trajectories.

With an algorithm to construct a regression model for covariance matrix responses in hand, we can now describe a key component of our contribution: a test statistic which allows addressing the main question of interest: *Is the progression/trajectory of covariance matrices (over time) different across two groups?* In the standard two-sample testing problem, a hypothesis test is set up to check if the parameters of each group are significantly different:

$$H_0: \theta_1 = \theta_2 \quad vs. \quad H_A: \theta_1 \neq \theta_2 \tag{3.1}$$

Recall that in a general linear model (GLM), when testing for mean group differences, the test parameters are the regression slopes from a standard GLM fit. In our setting, the parameters of interest are the population covariance trajectories estimated from the manifold regression in (2.2), see Fig. 1. While the trajectories and the slopes are related, note that our parameters are estimated *on the manifold*. Two unique manifold trajectories, when projected as simple multivariate responses in Euclidean space, may not be significantly different under the GLM hypothesis testing framework, as has been observed by [14]. Returning to our longitudinal trajectory formulation, we have the following naïve Covariance GLM:

DEFINITION 3.1.

Let $vec(C_{g,t})$ be the vectorized covariance matrix at timepoint t for group $g \in \{1, 2\}$. Then the naïve Covariance GLM is defined as

$$vec(C_{g,t}) = \beta_g^0 + \beta_g t + \epsilon \qquad (3.2)$$

with the slope $\theta = \beta$ in the hypothesis test in (3.1), and $vec(\cdot)$ is the vectorized form of the input matrix.

With this model, hypothesis testing reduces to a simple difference of slopes, which is well-studied in classical statistics literature.

DEFINITION 3.2.

[54] Let β_1, β_2 be the multivariate slopes calculated from estimating (3.2). Then an α-level hypothesis test rejects the null hypothesis $\beta_1 = \beta_2$ when $L > \chi_p^2|_{1-\alpha}$, where

$$L = (\hat{\beta}_1 - \hat{\beta}_2)\Sigma^{-1}(\hat{\beta}_1 - \hat{\beta}_2) \qquad (3.3)$$

Knowing that the response space is structured, i.e., our covariance matrices lie on the SPD manifold, we seek a more appropriate test and corresponding test statistic which adequately captures this knowledge.

Observe that we can directly apply the manifold regression in §2 to solve for a linear model on the manifold. That is, we construct the manifold GLM as

DEFINITION 3.3.

Let $C_{g,t}$ be the covariance matrix at timepoint t for group $g \in \{1, 2\}$. Then the Longitudinal-Covariance GLM (LCGLM) is defined as

$$C_{g,t} = \text{Exp}(b_g, V_g t) \qquad (3.4)$$

with b_g and V_g being the base point and tangent vector respectively, as described in §2.

But instead of solving $p(p-1)/2$ independent regressions, now we must concurrently solve for the entire manifold-valued response variable. In this case, we cannot directly compare our trajectories because they lie in *different* tangent spaces. To accurately compare two tangent vectors, we must parallel transport both vectors to the same tangent space. Once they are both in the same space, we can construct a simple test statistic for the trajectory difference.

$$L = \|\Gamma_{b_1 \to I}V_1 - \Gamma_{b_2 \to I}V_2\|_I^2 \qquad (3.5)$$

Recall that the inner product at the Identity I coincides with the Euclidean metric. This can now be naturally interpreted as a difference of slopes, and together with a standard Euclidean Normal noise assumption yields the following hypothesis test.

PROPOSITION 3.4.

Assume that $\Gamma_{b \to I} V$ is normally distributed $N(0, I)$. Then the statistic defined in (3.5) follows a χ_p^2 distribution with p degrees of freedom, and the threshold test in (3.2) is an α-level hypothesis for the covariate trajectory group difference.

3.1. Incorporating First-Order Differences.

In many real world situations, first-order information in the data is often valuable in identifying group differences. Restricting our analysis to only the second-order interactions, i.e., covariances, may be inefficient (or suboptimal) when the mean signal difference between groups is large. Our construction easily extends to these cases. Particularly, the *product space* over both means and covariances is in $\mathbb{R}^p \times \mathrm{SPD}(p)$.

REMARK 3.5.—The typical GLM on the first order information is defined in the standard Euclidean space. So, computing the regression in the product space $\mathbb{R}^p \times \mathrm{SPD}(p)$ amounts to simply computing the regression on the first and second order statistics (mean and covariance) separately.

The above statement suggests that by applying the manifold regression to the covariances and the standard regression model for the means, we are directly solving the product space regression problem, incorporating both first and second order statistics. However, in these cases, the statistic defined above in (3.5) does *not* directly take into account the potential difference in means. However, given our Normal noise assumption we can easily invoke the standard Gaussian multivariate likelihood statistic for group differences.

DEFINITION 3.6.—Let $\hat{\mu}_t, \hat{\Sigma}_t$ be the estimated mean and covariance from the standard linear model and our manifold-covariance GLM respectively. Then the Gaussian likelihood of our data X is

$$P(X \mid \hat{\mu}, \hat{\Sigma}) \prod_{t=1}^{T} \prod_{i=1}^{n_t} P(X_t \mid N(\hat{\mu}_t, \hat{\Sigma}_t)), \tag{3.6}$$

where X_t is the subset of our data collected at timepoint t. Additionally, we can define a standard likelihood ratio test statistic as:

$$L_{prod} = \frac{P(X_1 \mid \hat{\mu}_1, \hat{\Sigma}_1) P(X_2 \mid \hat{\mu}_2, \hat{\Sigma}_2)}{P(X_{1,2} \mid \hat{\mu}_{1,2}, \hat{\Sigma}_{1,2})} \tag{3.7}$$

This statistic is again χ_p^2-distributed [54], and an α-level hypothesis test for group difference analysis can be defined in the same way as above. While our manifold regression modeling is focused on the case of centered data (where the mean signal may not be significantly different between the groups), we use the product space construction, wherever appropriate, in experimental evaluations.

4. Localizing Group Differences for SPD(p) Trajectories.

The above procedure provides a precise mechanism to derive a statistic from the group-wise covariance matrix trajectories. However, when the effect sizes are poor, any scheme operating on the trajectories of the *full covariance matrix* may still fail to identify group differences (as is the case in our experiments). To improve statistical power, localizing the process of computing the trajectories *only to the relevant features* is critical. To this end, we consider the following global hypothesis testing problem

$$H_0 : \forall R, \beta_1^R = \beta_2^R \quad vs. \quad H_1 : \exists R, \beta_1^R \neq \beta_2^R,$$

where β denotes the slope and R is the region of the covariance matrix *which only includes the relevant features*, see Fig. 2. It turns out that by adapting *Scan statistics* [15, 3], we will be able to exclude the effect of irrelevant regions of the covariance matrix in the calculated trajectories. By extending this concept to graphs, we obtain an algorithm to identify *subsets of features* of the covariance matrix which show group differences that are otherwise unidentifiable, in a statistically rigorous way.

4.1. Scan Statistics.

Scan statistics are a valuable tool for structured multiple testing. In its simplest form, we can consider a setting where we place a window (or box) over a region R in an image and calculate a local statistic L_R, e.g., an average or a response to a convolution filter. Then, the window can be raster scanned at various locations in the image ($\mathscr{R}$) and the maximum over the set of local statistics can be called the scan statistic. Intuitively, if the image is assumed to be a Gaussian random field, we can set up a null hypothesis using a critical value and finding a statistically significant signal (i.e., regions) corresponds to comparing the local region-wise statistic with the critical value. Of course, there is flexibility in terms of specifying properties of the regions as described next.

DEFINITION 4.1.—Let $\mathscr{R}$ be the collection of all possible structured regions, and L_R be some statistic over region R, a structured subset of $\mathscr{R}$. The scan statistic is defined as $L^* = \max_{R \in \mathscr{R}} L_R$.

Recent results in scan statistics show how *size corrections* can be used to increase detection power in multi-scale analysis with nice guarantees [65, 66]. To utilize these ideas for our hypothesis test, we must extend scan statistics and these size corrections to a graph setting where the graph is induced by a sparse estimation of the precision matrix, e.g., graphical lasso (or any other algorithm of choice) over the features. To do so, structured regions R and a statistic L_R on each region must be defined on the graph. Intuitively, in our case, L_R must capture the "difference" in group-wise covariance trajectories. As we will describe shortly, it is in the context of this statistic where we utilize the LCGLM (3.4), which will be invoked at the *level of individual regions R*, one by one.

Let $G := (\mathscr{V}, E)$ be a graph over the features (represented in the covariance matrix) with vertex set $\mathscr{V}$ and edge set E. We define the structured region $R \subseteq G$ as a connected

subgraph of G corresponding to the selection of those vertices as our feature subset (block of the covariance matrix, see Fig. 2). A natural question is whether such an enumeration is tractable if the number of connected subgraphs $\mathscr{R}$ is exponential. It turns out that if we make a mild assumption on the graph, the number of induced regions can be shown to be polynomially bounded. Further, it then naturally provides a *size correction*, the analog for a multiple testing adjustment.

Remarks.: In our motivating application, the group differences we seek to identify will involve a cohesive set of features that will be connected to each other, by definition (large changes in covariances indicate dependent features). Based on this observation, we assume that the true localized subgraph is a "ball" subgraph.

Definition 4.2.—A ball subgraph consists of nodes with a given radius r from a particular node (see Fig. 2). The collection of ball subgraphs is defined as

$$\mathscr{R} = \{ B(v, r) : v \in \mathscr{V} \text{ and } r \in \mathbb{N} \} \tag{4.1}$$

where the ball subgraph $B(v, r) := \{v' \in V : d(v, v') \leq r\}$, and $d(v, v')$ is the minimum length path connecting v and v'.

With this assumption, it can be verified that we now only need to search a polynomially bounded set of regions.

Remark 4.3.—The number of unique ball subgraphs in any graph G is bounded above by $D|\mathscr{V}|$, where D is the diameter (longest chain) of the graph G.

On these regions (i.e., blocks of covariance matrix), we will invoke LCGLM to provide us a statistic L_R. This is just the difference in slopes of the calculated manifold regression across groups in (3.5). We will iteratively obtain this statistic for distinct regions R and find subgraphs that differ in their trajectories across groups using a size correction for hypothesis tests.

Let us revisit the standard linear model setting and assume that our slopes β_g^R correspond to the subset of slopes from features in R, and $\hat{\beta}_g^R$ is an estimate of that slope. In this case, we have the following statistic (see e.g. [54]),

$$\left(\hat{\beta}_1^R - \hat{\beta}_2^R \right) \Sigma_R^{-1} \left(\hat{\beta}_1^R - \hat{\beta}_2^R \right) \sim \chi^2_{|E(R)|}, \tag{4.2}$$

where Σ_R^{-1} is the covariance matrix of $\hat{\beta}_1^R - \hat{\beta}_2^R$. With a normal noise assumption, this covariance will be identity and the statistic would simply be the ℓ_2-norm difference as in the classical analysis. To make the statistics comparable across *different sizes*, we use the standardized version of a $\chi^2_{|E(R)|}$ distribution,

$$L_R = \frac{(\overset{\wedge R}{\beta}_1 - \overset{\wedge R}{\beta}_2)\Sigma_R^{-1}(\overset{\wedge R}{\beta}_1 - \overset{\wedge R}{\beta}_2) - E(R)}{\sqrt{E(R)}}. \tag{4.3}$$

We can extend this analysis to our manifold setting.

DEFINITION 4.4.—For a given structured region R, the region-based LCGLM is written as

$$(b_g^R, V_g^R) = \underset{(b^R, V^R) \in T\mathcal{M}^R}{\text{argmin}} \mathbb{E}[d(\text{Exp}(b^R, V^R t_g), C_g^R)^2] \tag{4.4}$$

where C_g^R is the covariance matrix subblock defined by features included in R for group g (t_g is our univariate predictor, i.e., time).

To compare the group trajectories, we first parallel transport each tangent vector to the identity as described in §2 and then compute the statistic in (3.5) given as $\|\Gamma_{b_1^R \to I} V_1^R - \Gamma_{b_2^R \to I} V_2^R\|_I^2$. In the case of the product space construction, we apply the test in (3.4) to the data subset corresponding to the features in region R, with the same correction as in (4.3).

Summary.: We now have a region-based statistic for the manifold regression setting that is approximately normally distributed $N(0, 1)$, allowing effective comparison across differently-sized regions.

4.2. Size Correction.

A final unresolved yet important issue is that we must correct L_R based on the number of edges $E(R)$ in R. This has a direct consequence on detection power. Observe that the normalization for size correction should be determined by the null distribution of L_R, i.e., when there is no slope difference in the trajectories between groups. In order to derive a correction, we need to characterize the behavior of scan statistics within roughly similar regions, $\max_{R \in \mathcal{R}(A)} L_R$, where $\mathcal{R}(A)$ is the collection of region Rs with similar size as $E(R)$,

$$\mathcal{R}(A) = \{R \in \mathcal{R} : A/2 < |E(R)| \leq A\}. \tag{4.5}$$

Clearly, the behavior of $\max_{R \in \mathcal{R}(A)} L_R$ depends on the "complexity" of $\mathcal{R}(A)$. A clear understanding of how similar subgraphs relate to each other leads directly to a correction tied to their relative sizes.

To investigate the complexity of $\mathcal{R}(A)$, we define the following quantities.

DEFINITION 4.5.—The distance between subgraphs R_1 and R_2 can be given as

$$d(R_1, R_2) = 1 - \frac{|E(R_1) \cap E(R_2)|}{\sqrt{|E(R_1)\|E(R_2)|}} \tag{4.6}$$

DEFINITION 4.6.—Let the ϵ-covering number of $\mathcal{R}(A)$, denoted by $N(A, \epsilon)$, be the smallest integer such that there is a subset $\mathcal{R}_{approx}(A, \epsilon)$ of $\mathcal{R}$ such that

$$\sup_{R_1 \in \mathcal{R}(A)} \inf_{R_2 \in \mathcal{R}_{approx}(A, \epsilon)} d(R_1, R_2) \leq \epsilon \tag{4.7}$$

where $|\mathcal{R}_{approx}(A, \epsilon)| = N(A, \epsilon)$.

We can verify that all regions in $\mathcal{R}(A)$ can be approximated by regions in $\mathcal{R}_{approx}(A)$ with reasonably small error. From the definitions, notice that the complexity of $\mathcal{R}(A)$ is reflected by $N(A, \epsilon)$. If $N(A, \epsilon)$ is nicely bounded (as is the case here), scan statistics can be calculated very efficiently (Lemma 4.8).

Before stating this result, we make a mild assumption on our graph. For any ball subgraph, the edges around its center are not too sparse, compared to the edges in the outer region of the ball subgraph, i.e., hard on the inside, soft on the outside. This yields,

ASSUMPTION 4.7.—(Avocado) There exist constants S and H such that, for any $r/2 \leq r' \leq r$ and $v \in \mathcal{V}$,

$$\frac{|E(B(v, r'))|}{|E(B(v, r))|} \geq H\left(1 - \frac{|E(B(v, r - r'))|}{|E(B(v, r))|}\right)^S. \tag{4.8}$$

We see that this assumption holds for many classes of graphs: a ring graph satisfies this condition when $H = 1$ and $S = 1$ and the 2-d lattice satisfies this condition when $H = 1/4$ and $S = 2$ (see Fig. 3). With this assumption, we have the following result for the ϵ-covering number $N(A, \epsilon)$.

LEMMA 4.8.—Let $|E|$ be the total number of edges in G. If (4.8) holds and A is given, then, for a constant $C_{H,S}$ which only depends on H and S in (4.8),

$$N(A, \epsilon) \leq C_{H,S} \frac{|E|}{A}\left(\frac{1}{\epsilon}\right)^{S+1}. \tag{4.9}$$

The proof of this result follows from our ball-subgraph construction and our Avocado assumption and provided in the Appendix.

Intuitively, this result upper bounds the number of graphs that are necessary to search over to completely exhaust the search space of subgraphs. With this result, we can now construct a suitable size correction. Following the work of [65] and [66], we can increase the power of our test by using the following statistic:

$$T^* = \max_{R \in \mathcal{R}}\left(L_R - 2\sqrt{\log\frac{|E|}{|E(R)|}}\right). \tag{4.10}$$

The significance of this size correction is that we now have *a single critical value for each candidate subgraph, regardless of the subgraph size*. Our final test is defined as $\mathbb{I}[T^* > q_\alpha]$, where q_a is the a-level quantile of T^* under the null hypothesis (that no region is truly significant across groups). By construction, we can control the type 1 error at a specified a-level.

Under the alternative hypothesis of this framework, it is important to note that in many cases, large subgraphs that subsume smaller significant graphs may also have large test statistics, and our hypothesis test only indicates the existence of *some* significant region. To identify or localize the smaller subsets, we follow the procedure from [30], by beginning with the subgraph with the largest test statistic and iteratively removing overlapping subsets from the total set of subgraphs. This requires testing each regional/local statistic, $\left(L_R - 2\sqrt{\log(|E|/|E(R)|)}\right)$ against q_a. Under this procedure, we can control the weak family-wise error rate (wFWER) if we view our problem via the lens of multiple testing. The weak FWER is the probability of false discovery under the null hypothesis. To see that this is inherently controlled, note

$$\mathbb{P}(FN \geq 1 \mid H_0) = \mathbb{P}(T^* > q_\alpha \mid H_0) \leq \alpha, \tag{4.11}$$

where *FN* is the number of false discoveries under the null hypothesis. With this correction at the group difference level, we completely avoid any multiple comparisons issues that would arise in the case of a test for each subgraph. In addition to controlling the false positive rate, we have the following guarantee on *identifying truly significant regions* under the normal noise assumption.

THEOREM 4.9.—If (4.8) holds and the number of edges in the candidate subgraph is larger than $\log^2 |E|$, i.e.,

$$|E(R)| \gg \log^2 |E| \quad \forall R \in \mathscr{R}, \tag{4.12}$$

then the critical value q_a satisfies

$$q_\alpha = O(1). \tag{4.13}$$

Moreover, as $|E| \to \infty$, if a subgraph R_0 obeys

$$\frac{(\beta_1^{R_0} - \beta_2^{R_0})^T \Sigma_{R_0}^{-1}(\beta_1^{R_0} - \beta_2^{R_0})}{\sqrt{|E(R_0)|}} \gg 2\sqrt{\log\frac{|E|}{|E(R_0)|}}, \tag{4.14}$$

then as $|E| \to \infty$,

$$\mathbb{P}\left(L_{R_0} - 2\sqrt{\log\frac{|E|}{|E(R_0)|}} > q_\alpha\right) \to 1. \tag{4.15}$$

The full proof of this result follows a generic chaining argument (see, e.g. [60]) along with application of concentration inequalities and union bounds, and can be found in the Appendix.

Summary.: At a high level, this result directly characterizes the behavior of T^* under the null hypothesis H_0 and the alternative hypothesis H_1, respectively. We see that (4.13) implies that T^* can roughly be seen as a constant under the null hypothesis, and under the alternative hypothesis when (4.14) is satisfied, the test based on T^* is consistent, see (4.15).

4.3. Workflow for conducting hypothesis tests on temporal trends of graphs.

With these guarantees, our full workflow is as follows. First, we use an oracle procedure to generate a graph over our features that roughly captures the conditional independences. Any procedure that provides a conditional independence graph is sufficient. Next, for each ball subgraph over this graph, we compute the Longitudinal-Covariance GLM over these features for both groups, and compute the statistics outlined in §3. We then compute the size-corrected statistic, and compare against the single critical value. For all regions that pass this threshold, we apply the procedure from [30]. This workflow shows how to conduct hypothesis tests on temporal trends of large covariance matrices, with improved power and bounded Type 1 error. Additional implementation details can be found in the Appendix.

5. Localization Evaluation: Trends of Tobacco Usage Across Gender.

We begin our empirical analysis of the model by first applying the subgraph localization procedure by itself (standalone), separate from our manifold regression scheme. In this case, our statistic is derived from *only* Generalized Linear Models (GLM) constructions, where the $\hat{\beta}_g^R$ in equation (4.3) is the slope estimated from fitting standard first order linear models. Identifying the differentially varying subgraphs across groups in this way is similar to a simpler version of the planted clique identification problem [4], where the clique we are trying to identify corresponds to those nodes whose slopes vary significantly across groups.

Data.

The Center for Disease Control (CDC) provides extensive statistics regarding tobacco and alcohol usage across the US. This data has been collected systematically for the last few decades and is publically available (includes demographic information and gender). As a simple application of our proposed framework, we may pose the following question: which "sub-groups" of states tend to evolve differently in their correlation (pertaining to tobacco/alcohol usage) over time? Our framework extends easily to answer this question. In this setup, the oracle graph is simply the adjacency graph of the continental US naturally which will be used directly in our scanning procedure. For this dataset, we have direct observations of node measures: the percentage of males and females who reported smoking or drinking heavily in each state. Using gender as the group, we fit standard linear models for each candidate subgraph, and compute the difference of gender-wise slopes statistic as described above. In Figure 4, we see the regions identified using our method, and interpret some of the tobacco usage findings here.

In the northeast, we see that women have reduced their tobacco usage at a significantly faster rate than men compared to the rest of the country. We suspect that this may be at least partly tied to the development of women's cigarette brands in the late 1960s and 1970s followed by subsequent aggressive public policy campaigns in the 1990s and 2000s to highlight health risks beyond pulmonary or cardiovascular diseases for women (e.g., infertility, reduced bone-density in post-menopausal women). We also see that state-wide indoor smoking bans were put in place in the Northeast ahead of many other states in the union. In the South, the trends among men and women also seems to differ significantly. (see Fig. 4). Apart from health factors, the group-wise differences in the group-wise trends may also be explained by a few reasons identified in a study in 2007 [57] which found that as the state sales tax on cigarettes changed (increased), women were significantly more price elastic than men. Between 2006 and 2008, the cigarette tax increased dramatically for all of the 4 states identified *except for Louisiana*, whose tax rate has remained constant. Additionally, while Arkansas did increase their cigarette tax in 2009, they did *not increase taxes in locations near borders shared with higher taxing states*. These intricate relationships among states lend credibility to the fact that our scan statistics framework is indeed identifying interesting sub-regions, and suggests that the full covariance-trajectory pipeline may be more appropriate if effects beyond the means are relevant within an analysis.

6. Pipeline Evaluation on Simulations and Baby Name Trends Over Time.

We next evaluate the ability of our entire analysis pipeline to identify group differences across temporally evolving *covariance* trajectories. In many existing analyses, the effect of the mean differences may be stronger than the effect of the interaction matrix. However, in cases where the *mean signal is weak*, we expect that the covariance effect will be important. To evaluate our model in this regime, we perform a set of simulation studies and also analyze a publicly available longitudinal dataset.

Simulations.

We randomly generate SPD matrices from a 'path' of 4 discrete points along the manifold, and use these data as population covariance matrices to generate 0-mean sample data. Table 6 shows the results of the hypothesis testing procedure with 50 features averaged over 100 runs, where both the true number of features with covariance trajectory differences, p_t, and the number of samples per group, n, were varied. As expected, our recovery rate increases nicely as a function of the number of samples n and decreases as the size of region of change p_t is increased when n is held constant.

We compare our model to baseline methods that may be used in practice for the foregoing group difference hypothesis test. In standard applications, general linear models (GLMs) are often the first line of attack. When the covariates are assumed to be independent, a simple linear model as in (3.2) may be suitable. However, when the group difference is influenced by specific interactions between covariates, such linear models require additional care. A typical solution is to introduce pairwise interaction terms into the model – a choice between all possible interactions or *specific interactions specified by an expert*. The first model has problems since the number of samples $n \ll p^2$. In the second model, we depend

completely on the user's choice of interactions, and must correct for multiple testing when testing different models, at least partly reducing the power of the final test. Figure 5 shows the value of our method over these models. For the interaction GLM case, we randomly select interaction terms to include in the GLM, with size p_t (the ground truth number of variables in the interaction). In this way, we approximate the effect of an oracle specifying to the GLM which terms may describe the underlying interaction. We report the fraction of significance tests where a significance threshold of p 0.05 was found for each model, averaged over 100 runs. We see that our proposed scheme consistently achieves near-perfect results in terms of the percentage of null hypotheses that were correctly rejected (i.e., there was a significant group-difference signal). The power of scan statistics on graphs is particularly evident in the needle in haystack setting where the true differential signal is small (p_t 8) and the sample size is small to medium. When the sample size is large and p_t is also large, the standard linear model with additional interaction terms starts to approach the statistical performance of our algorithm.

Longitudinal trends in Baby Names.

In addition to the simulations above, we report results from a simple analysis of how male/female baby names evolve over time over the last century. The United States Social Security Administration provides a publicly available dataset listing the frequency of the top 1000 baby names in each state for the last 106 years. We evaluate our model in this context to examine which "sub-group" of states tend to evolve (or change) in their "name agreement" (or correlation) over time between boy names and girl names. Here, rather than calculating a sample covariance at each timepoint, we calculate a rank correlation matrix instead. For example, if two neighboring Gulf Coast states, say Georgia and Alabama, substantially agreed on both boys and girls names in the period following the second World War, but gradually this agreement declined over time for girls (but not boys), we expect that our scan statistics on graphs hypothesis test will segment out this differential signal (in slope trends) from the planar graph induced by the states sharing a border. Shown in Figure 6 are the regions identified using our method, applied on only the rank correlations for the top 10 names for both genders per state per year. Each highlighted region indicates a sub-group where their "trends of correlation (or agreement/disagreement)" in preferred baby names over the last century varies between boys and girls. For states not identified by our model (in gray), we can conclude that the state-to-state name preference-interactions may have still evolved over time but we have insufficient statistical evidence to conclude that such trends (slopes) are different between boys and girls.

7. Identifying Differentially Covarying Features in Preclinical Alzheimer's Disease.

We now describe experiments and results focused on the key motivation of this work — to facilitate analysis of a longitudinal study of individuals at risk for Alzheimer's disease (AD) where the statistical signal is weak (with small to medium sample sizes). We describe the dataset details followed by the analysis and then interpret our conclusions in the context of scientific results that have been published in the literature in aging and dementia.

Study background.

We analyzed data from a cohort of individuals who have been longitudinally tracked for at least three visits over multiple years, as part of an ongoing study (since 2001) to understand the disease processes in the brain *before* an individual exhibits signs of cognitive decline due to Alzheimer's Disease (AD) [52]. The study, Wisconsin Registry for Alzheimer's Prevention (WRAP) is among the largest of its kind in existence, focused on "preclinical" AD, i.e., when the individuals are still cognitively healthy, offering a window into the early disease processes where treatments, drugs and interventions are likely to be most effective. WRAP and its ancillary studies acquire neuroimaging data (MRI, PET with different tracers, diffusion MRI) and various clinical test scores, genetic and demographic data as well as clinical measures such as Cerebrospinal Fluid (CSF). Our analysis seeks to understand subtle group-wise differences in longitudinal patterns of dependencies between these measures at this early stage of the disease.

Dataset.

The dataset consisted of 114 subjects with imaging data from at least two types of imaging modalities: Positron emission tomography and diffusion weighted Magnetic Resonance (MR) images. Positron emission tomography (PET) images were used to calculate, using well-validated pre-processing pipelines, the mean amyloid-plaque load (an important biomarker for AD) in 16 different anatomical regions of interest in the brain. Amyloid plaque is known to be an AD-related pathology and generally *precedes* onset of cognitive symptoms. Separately, diffusion tensor MR imaging (DTI) data were processed and used to calculate both Fractional Anisotropy (FA) and Mean Diffusivity (MD) in 48 distinct regions [44]. DTI images provide information about structural connectivity between gray matter regions in the brain. In addition to these 108 (48×2+16) image-derived features, we also included in the analysis the participant's scores on a battery of cognitive tests, known to be correlated with various neuropsychological functions [39]. Differences were evaluated on various groupings of the subjects which were, for the most part, based on known results in the literature. Specifically, gender, APOE (Apolipoprotein E) genotype and amyloid positivity (based on thresholding the amyloid plaque summaries) have all been evaluated as significant in AD studies [49] but often such analyses involve a population covering a broader disease spectrum where the signal is much stronger.

Is analysis of second order statistics necessary?—In Figure 7, we present histograms detailing the distribution of two critical cognitive tests, stratified across various groups of scientific interest. Evaluating these distributions were the key motivation for our exploration into the methods described in the paper. Small differences in means across groups *regardless of grouping selection (i.e., stratification variable)*, and the saturation that occurs at the ceiling of cognitive test scores and other preliminary experiments conducted by us suggest that standard analyses are not sensitive enough to identify subtle higher-order differences.

7.1. Results for Group difference analysis for individuals with imaging data.

We now describe, one by one, the components of the largest feature subset discovered for each stratification scheme and highlight the main scientific findings. In most cases, we provide a brief scientific interpretation of the results for the interested reader. Additional details and results are available in the appendix.

A) Graph Scan Statistics on slope differences across gender.—The most significant (based on region-score) subset identified by the gender grouping was between the FA DTI measurement in the left cingulum gyrus as well as the scores on the Rey Auditory Verbal Learning Test (RAVLT). In recent AD research, gender has been identified as a factor in the progression of various pathology measures (e.g., incidence and prevalence of AD is higher in women [18, 50]), and has contributed to a formal NIH notice (NOT-OD-15–102). However, we note that previous work in the field has *not* identified gender-related differences when looking *only* at diffusion measures in the cingulum [40]. Our algorithm successfully identified longitudinal changes in *interaction* between these variables which supports the earlier results, and provides some evidence that as men and women age, their cognitive decline as measured by RAVLT manifests differently in relation to the cingulum gyrus.

B) Graph Scan Statistics on slope differences across genotype.—Next, we stratified the cohort based on the genotype known to be most closely linked with AD, i.e., the APOE (Apolipoprotein E) gene [8] — we inherit one APOE allele from each parent; having one or two copies of the e4 allele increases a person's risk of getting AD whereas the rarer e2 allele is associated with a lower risk of AD. Using this stratification, we obtain a low-risk and an at-risk group of individuals. Here, we identified amyloid-load regions within the medial and lateral parietal lobes and find that in the "low-risk" group, the covariances between Digit Span and Stroop Color-Word scores (attention and concentration scores) and amyloid load moves from strongly negative towards 0 as a function of age (Table 3). In the "at-risk" group (APOE4), however, we find that as a function of age, the features become more and more positively correlated. Existing studies have shown that the accumulation of amyloid is significantly different across APOE4 gene expression [45], and our results provide some evidence that the expression of the genotype may interact with cognitive scores as well, *even at this early stage of the disease*, when the individuals in our cohort are cognitively healthy. The sets of features showing a differential signal are presented in Table 3.

C) Graph Scan Statistics on slope differences across amyloid load positivity.—As briefly described above, amyloid load is an important biomarker for AD. For our analysis, amyloid (or PiB) positivity is calculated using the mean amyloid PiB measures across all brain regions using a PiB PET image scan of the participant. When we used this measure for stratification (threshold was set at 1.18, following [11]), our model identified fifteen of the sixteen PiB regions that were input to the model when the density of the oracle graph was set to be high. This result is as expected, but interestingly we find that controlling for the linear combination of the features (through centering), the residual error *still* has significant signal with the PiB positivity measure, indicating that amyloid burden

interactions across brain regions plays a very important role in AD progression [25, 26, 61, 28]. When the sparsity of the oracle graph was increased, however, four neighboring regions, the left and right corticospinal tract and the left and right cerebral peduncle were identified on both PiB and DTI measures (supported by the literature [13]), together with Part A of the Trail Making Test (see Table 4) which happens to be used in AD diagnosis [1]. This suggests that changes in atrophy within these regions, as measured by DTI, co-occur with changes in amyloid burden. Additionally, because these regions are highly correlated with rough and fine motor ability [47], it seems plausible that amyloid positivity will lead to higher 'covariation' in the regions associated with a measure of fine motor speed, i.e., the Trail Making Test.

7.2. Results for for Group difference analysis for individuals with Cognitive Testing data.

In addition to the dataset presented above, we apply our method to a much larger dataset consisting of approximately 1500 individuals with only cognitive testing data collected in a longitudinal manner. Each individual was administered these tests for between two and three time-points, yielding approximately $n = 4000$ samples for our model. For each assessment, a conference of experts applied a diagnostic label indicating normal cognition or mild cognitive impairment. Using this binary classification, we can stratify our population for group difference analysis. We find that among many different significant subsets, the covariance trajectory among the scores on both parts of the Trail-Making Test and on all trials of the RAVLT test explain a significant group difference. These have previously been shown to be the *most sensitive tests* for early cognitive decline [2]. Table 5 displays the other tests identified by our algorithm, and additional experiments on this larger cohort can be found in the appendix.

7.3. Baseline.

In various experiments on this dataset, when the MMGLM procedure is performed for the entire feature set in totality (*not* utilizing any of the proposed ideas based on scan statistics), and the null distribution derived using permutation testing, the procedure *yields no significance across any scientifically interesting group stratifications*. This implies that the ability to search over different blocks of the covariance matrix is critical in identifying meaningful group differences in the trajectories, unavailable using alternate schemes. For instance, simpler strategies work well enough for datasets such as ADNI – which includes diseased subjects as well as controls – where the signal is stronger and even temporal modeling may be unnecessary. While the scientific results need to be interpreted with caution and reproducibility experiments on other similar datasets (both within the US and internationally) are in the planning phase, we believe that the ability to localize differences in these interaction patterns in a statistically rigorous manner is valuable and these findings can be investigated standalone, via more classical schemes (e.g., structural equation modeling).

8. Conclusions.

The analysis of datasets to identify where clinically disparate groups differ is pervasive in biology, neuroscience, genomics and epidemiological studies. We find that graphical

models are an ideal tool to analyze high-dimensional data in these areas but have been sparingly used for the analysis of group-wise differences, especially in a longitudinal setting. Motivated by an application related to longitudinal analysis of imaging and clinical/cognitive data from otherwise healthy individuals who are at risk for Alzheimer's disease (AD), we show how a combination of manifold regression with a generalization of scan statistics to the graph setting yields tools that can be directly deployed. We present an efficient algorithm and develop the theoretical results showing the regimes where its application is appropriate. In various experiments, while the standard schemes are not sufficiently powered to detect the signal, our proposed formulation is able to detect meaningful group difference patterns, many of which have a clear scientific interpretation. We believe that these results are promising for the neuroimaging application described and other regimes where group-wise analysis is desired but the number of features is large.

Acknowledgements.

Appendix A.: Technical Proofs.

A.1. Proof of Lemma 4.8.

To remind the reader, this result was necessary in order to allow us to reduce the number of subgraphs (regions) that need to be evaluated over the graph. By bounding the covering number we have a guarantee that we do not need to consider an exponential number of subgraphs in order to find a localization.

Proof.

To upper bound $N(A, \epsilon)$, we first construct the ϵ-covering set of $\mathcal{R}(A)$ under metric d. To this end, we decompose $\mathcal{R}(A)$ into several disjoint sets

$$\mathcal{R}_j(A) = \left\{ B(v, r) \in \mathcal{R}(A) : \left(1 - \frac{(j+1)\epsilon}{2}\right) A < |E(B(v, r))| \leq \left(1 - \frac{j\epsilon}{2}\right) A \right\},$$

for $j = 0, 1, ..., \left\lceil \frac{1}{\epsilon} \right\rceil$. Our strategy is to construct ϵ-covering set for each set $\mathcal{R}_j(A)$.

We only construct ϵ-covering set for $\mathcal{R}_0(A)$; $\mathcal{R}_j(A)$ $(j \geq 1)$ can be treated similarly. To construct the ϵ-covering set for $\mathcal{R}_0(A)$, we denote by $d_{v,r}$ the largest positive number such that

$$\frac{|E(B(v, r - d_{v,r}))|}{|E(B(v, r))|} \geq 1 - \frac{\epsilon}{2}, \tag{A.1}$$

for every $v \in V$ and $r \in \mathbb{N}$. Let $\mathcal{D}_1$ the collection of $d_{v,r}$ such that $B(v, r) \in \mathcal{R}_0(A)$, i.e.

$$\mathcal{D}_1 = \left\{ d_{v,r} : B(v,r) \in \mathcal{R}_0(A) \right\},$$

and $\mathcal{V}_1$ the collection of nodes such that $B(v,r) \in \mathcal{R}_0(A)$, i.e.

$$\mathcal{V}_1 = \{v : B(v,r) \in \mathcal{R}_0(A)\}\,.$$

We pick up the largest number in $\mathcal{D}_1$, denoted by d_{v_1,r_1}, i.e. $d_{v_1,r_1} \geq d_{v,r} \,\forall d_{v,r} \in \mathcal{D}_1$ and define $\widetilde{\mathcal{V}}_1$ as

$$\widetilde{\mathcal{V}}_1 = \left\{ v \in \mathcal{V}_1 : v \in B(v_1, d_{v_1,r_1}/2) \right\}\,.$$

After defining $\widetilde{\mathcal{V}}_1$, $\mathcal{D}_2$ and $\mathcal{V}_2$ can be defined as

$$\mathcal{D}_2 = \mathcal{D}_1 \backslash \{d_{v,r} : v \in \widetilde{\mathcal{V}}_1\} \quad \text{and} \quad \mathcal{V}_2 = \mathcal{V}_1 \backslash \widetilde{\mathcal{V}}_1\,.$$

Then we can pick up the largest number in $\mathcal{D}_2$, denote by d_{v_2,r_2} and $\widetilde{\mathcal{V}}_2$ can be defined similarly. We can repeat the above process until $\mathcal{D}_M$ and $\mathcal{V}_M$ are empty for some M. We actually obtain a partition of $\mathcal{V}_1$,

$$\bigcup_{i=1}^{M} \widetilde{\mathcal{V}}_i = \mathcal{V}_1 \quad \text{and} \quad \widetilde{\mathcal{V}}_{i_1} \cap \widetilde{\mathcal{V}}_{i_2} = \varnothing \quad 1 \leq i_1 < i_2 \leq M\,.$$

Based on $d_{v_1,r_1}, \ldots, d_{v_M,r_M}$, we are ready to prove the set

$$\mathcal{R}_0(A, \epsilon) = \{B(v_i, r_i) : 1 \leq i \leq M\}$$

is actually an ϵ-covering set for $\mathcal{R}_0(A)$. To this end, it is equivalent to show that for arbitrary $B(v', r') \in \mathcal{R}_0(A)$, we have

$$d(B(v', r'), B(v_i, r_i)) \leq \epsilon \tag{A.2}$$

when $v' \in \widetilde{\mathcal{V}}_i$. To show (A.2), we consider two cases where $r' > r_i - d_{v_i,r_i}/2$ and $r' \leq r_i - d_{v_i,r_i}/2$. When $r' > r_i - d_{v_i,r_i}/2$, then

$$B(v_i, r_i - d_{v_i,r_i}) \subset B(v', r')\,.$$

Combining above result, (A.1), and the definition of $\mathcal{R}_0(A)$ yields

$$\frac{|E(B(v', r')) \cap E(B(v_i, r_i))|}{\sqrt{|E(B(v', r'))||E(B(v_i, r_i))|}}$$

$$\geq \frac{|E(B(v_i, r_i - d_{v_i, r_i}))|}{\sqrt{|E(B(v', r'))||E(B(v_i, r_i))|}}$$

$$\geq \sqrt{1 - \frac{\epsilon}{2}} \frac{|E(B(v_i, r_i - d_{r_i}))|}{|E(B(v', r'))|}$$

$$\geq 1 - \epsilon.$$

On the other hand, if $r' \leq r_i - d_{v_i, r_i}/2$, then

$$B(v', r') \subset B(v_i, r_i). \tag{A.3}$$

By definition of $\mathcal{R}_0(A)$, we can get

$$\frac{|E(B(v', r')) \cap E(B(v_i, r_i))|}{\sqrt{|E(B(v', r'))||E(B(v_i, r_i))|}} \geq \sqrt{\frac{|E(B(v', r'))|}{|E(B(v_i, r_i))|}} \geq 1 - \epsilon.$$

Therefore, (A.2) is proved and $\mathcal{R}_0(A, \epsilon)$ is an ϵ-covering set for $\mathcal{R}_0(A)$.

The rest of the proof is to bound the cardinality of $\mathcal{R}_0(A, \epsilon)$, i.e. M. Note that (4.8) implies there exists some constant $D_{H,S}$ only depending on H and S such that, for any $v \in V$ and $r \in \mathbb{N}$,

$$|E(B(v, r/2))| \geq D_{H, S}|E(B(v, r))|.$$

By the definition of d_{v_i, r_i}, we can ensure $B(v_i, d_{v_i, r_i}/4)$ are disjoint. Hence, this implies

$$|E(\widetilde{\mathcal{V}}_i)| \geq |E(B(v_i, d_{v_i, r_i}/4))| \geq D_{H, S}^2 |E(B(v_i, d_{v_i, r_i}))| \geq D_{H, S}^2 H A \epsilon^S / 2^{S+1}.$$

The last inequality is suggested by (4.8) and (A.1). The volume argument yields

$$M \leq \frac{|E|}{D_{H, S}^2 H A \epsilon^S / 2^{S+1}} \leq \frac{2^{S+1}}{D_{H, S}^2 H} \frac{|E|}{A} \left(\frac{1}{\epsilon}\right)^S$$

(4.9) is obtained upon application of the above to each $\mathcal{R}_j(A)$.

A.2. Proof of Theorem 4.9.

Before we are ready to prove Theorem 4.9, we need the following result:

Lemma A.1.

Let $Y_1, \ldots, Y_d$ be i.i.d. standard Gaussian variable, i.e. $N(0, 1)$ and $a_1, \ldots, a_d$ be a sequence of numbers. If

$$Z = \sum_{i=1}^{d} a_i(Y_i^2 - 1), \tag{A.4}$$

then

$$\mathbb{P}(|Z| \geq 2|a|_2\sqrt{x} + 2|a|_\infty x) \leq 2\exp(-x) \tag{A.5}$$

where $|a|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}$ and $|a|_\infty = \max_{i=1, \ldots, d} |a_i|$.

Proof.—This is a direct extension of lemma 1 in [35] to the negative case. We follow arguments similar to theirs. Let $\varphi(x)$ be the the logarithm of the Laplace transform of $Y_i^2 - 1$. For any $-1/2 < x < 1/2$,

$$\phi(x) = \log(\mathbb{E}(\exp(x(Y_i^2 - 1)))) = -x - \frac{1}{2}\log(1 - 2x) \leq \frac{x^2}{1 - 2|x|}.$$

This leads to

$$\begin{aligned}
\log(\mathbb{E}(e^{xZ})) &= \sum_{i=1}^{d} \log\left(\mathbb{E}\left(\exp\left(a_i x\left(Y_i^2 - 1\right)\right)\right)\right) \\
&\leq \sum_{i=1}^{d} \frac{a_i^2 x^2}{1 - 2|a_i|x} \\
&\leq \frac{|a|_2^2 x^2}{1 - 2|a|_\infty x}
\end{aligned}$$

With the same arguments in [35], we could prove that

$$\mathbb{P}\left(Z \geq 2|a|_\infty x + 2|a|_2\sqrt{x}\right) \leq \exp(-x).$$

The other direction can be proved if we apply the same argument for $-Z$. □

With this in hand we proceed to prove Theorem 4.9.

Proof.—In the following proof, C always refers to some constant, although its value may change from place to place. First, we prove (4.13). To this end, we prove concentration inequalities for L_R for some R and $L_{R_1} - L_{R_2}$ for some R_1 R_2. Since we assume the noise follows normal distribution, we have

$$(\hat{\beta}_1^R - \hat{\beta}_2^R)^T \Sigma_R^{-1}(\hat{\beta}_1^R - \hat{\beta}_2^R) = \frac{\sum X_i^2 - (\sum X_i)^2}{2}\|\hat{\beta}_1^R - \hat{\beta}_2^R\|^2 \sim \chi_{|E(R)|}^2.$$

By tail bound for χ^2 random variables (see e.g. [35]), we can yield

$$\mathbb{P}\left(L_R > 2t + \frac{2t^2}{\sqrt{|E(R)|}}\right) \le \exp(-t^2).$$ (A.6)

By definition, $L_{R_1} - L_{R_2}$ can be written as

$$L_{R_1} - L_{R_2} = \frac{\sum_{i \in R_1 \setminus R_2} Z_i}{\sqrt{|E(R_1)|}} + \left(\frac{1}{\sqrt{|E(R_1)|}} - \frac{1}{\sqrt{|E(R_2)|}}\right)_{i \in R_1 \cap R_2} \sum Z_i - \frac{\sum_{i \in R_2 \setminus R_1} Z_i}{\sqrt{|E(R_2)|}}$$

where Z_i are independent random variable following distribution $\chi_1^2 - 1$.. Lemma A.1 implies

$$\mathbb{P}\left(|L_{R_1} - L_{R_2}| > 2\sqrt{2d(R_1, R_2)}t + \frac{2t^2}{\min(|E(R_1)|, |E(R_2)|)}\right) \le 2\exp(-t^2).$$ (A.7)

We now proceed to prove (4.13) by applying a chaining argument (See [60]) and concentration inequalities (A.6) and (A.7). Recall $\mathscr{R}_{app}(A, \epsilon)$ is the smallest ϵ-covering set of $\mathscr{R}(A)$ and $N(A, \epsilon)$ is the covering number of $\mathscr{R}(A)$. For any subgraph candidate R, we denote by

$$\pi_l(R) = \operatorname{argmin}_{R' \in \mathscr{R}_{app}(A, e^{-l})} d(R, R').$$

For any $l^* > l_*$, which will be specified later, we write $\max_{R \in \mathscr{R}(A)} L_R$ into three parts

$$\max_{R \in \mathscr{R}(A)} L_R \le \max_{R \in \mathscr{R}(A)} |L_R - L_{\pi_{l^*}(R)}| + \sum_{l = l_*}^{l^* - 1} \max_{R \in \mathscr{R}(A)} |L_{\pi_{l+1}(R)} - L_{\pi_l(R)}| + \max_{R \in \mathscr{R}(A)} L_{\pi_{l_*}(R)}.$$

Now, we bound these three terms above separately.

Term 1.

Let $l^* = 2\log|E|$. By concentration inequality (A.7) and union bound, we have

$$\mathbb{P}\left(\max_{R \in \mathscr{R}(A)} |L_R - L_{\pi_{l^*}(R)}| > \frac{2\sqrt{2(x + \log|E|)}}{|E|} + \frac{4x + 8\log|E|}{A}\right)$$
$$\le |\mathscr{R}(A)|\mathbb{P}\left(|L_R - L_{\pi_{l^*}(R)}| > \frac{2\sqrt{2(x + \log|E|)}}{|E|} + \frac{4x + 8\log|E|}{A}\right)$$
$$\le 2\frac{|\mathscr{R}(A)|}{|E|^2}\exp(-x) \le 2\exp(-x)$$

for $x < \log|E|$. Therefore, we have

$$\mathbb{P}\left(\max_{R \in \mathcal{R}(A)} |L_R - L_{\pi_{l*}(R)}| > \frac{C(x + \log|E|)}{A}\right) \le \exp(-x),$$

for $x < \log|E|$.

Term 2.

Let $l_* = \log \log(|E|/A)$. Recall that the Avocado assumption (4.8) suggests that

$$N(A, \epsilon) \le C_{H,S} \frac{|E|}{A}\left(\frac{1}{\epsilon}\right)^{S+1}. \tag{A.8}$$

Applying concentration inequality (A.6) along with

$$t = \sqrt{\log\left(\frac{|E|}{A}\right) + (S+1)\log \log\left(\frac{|E|}{A}\right) + x + C} \tag{A.9}$$

and the union bound, we have

$$\mathbb{P}\left(\max_{R \in \mathcal{R}(A)} L_{\pi_{l*}}(R) > 2t + \frac{2t^2}{\sqrt{A}}\right)$$

$$\le N\left(A, \frac{1}{\log(|E|/A)}\right)\mathbb{P}\left(L_{\pi_{l*}}(R) > 2t + \frac{2t^2}{\sqrt{A}}\right)$$

$$\le C_{H,S} \frac{|E|}{A}\left(\log\frac{|E|}{A}\right)^{S+1}\mathbb{P}\left(L_{\pi_{l*}}(R) > 2t + \frac{2t^2}{\sqrt{A}}\right)$$

$$\le \exp(-x)$$

for $x < \log|E|$. Here we also apply condition (4.12). Therefore, we obtain

$$\mathbb{P}\left(\max_{R \in \mathcal{R}(A)} L_{\pi_{l_*}}(R) > 2\sqrt{\log\left(\frac{|E|}{A}\right) + (S+1)\log\log\left(\frac{|E|}{A}\right) + x + C}\right) \le \exp(-x)$$

for $x < \log|E|$.

Term 3.

For any given l, application of concentration inequality (A.7), covering number condition (A.8), and the union bound yields,

$$\mathbb{P}\left(\max_{R \in \mathcal{R}(A)} |L_{\pi_{l+1}}(R) - L_{\pi_l}(R)| > \sqrt{\frac{C(\log(|E|/A) + l + x)}{e^l}} + \frac{C(\log(|E|/A) + l + x)}{A}\right)$$

$$\le C_{H,S} \frac{|E|}{A} e^{(l+1)(S+1)}\mathbb{P}\left(|L_{\pi_{l+1}}(R) - L_{\pi_l}(R)| > \sqrt{\frac{C(\log(|E|/A) + l + x)}{e^l}} + \frac{C(\log(|E|/A) + l + x)}{A}\right)$$

$$\le \frac{\exp(-x)}{l^2}.$$

for any $x < \log|E|$. With another standard application of the union bound, we have

$$\mathbb{P}\left(\sum_{l=l_*}^{l^*-1} \max_{R \in \mathcal{R}(A)} |L_{\pi_l+1}(R) - L_{\pi_l}(R)| > \sqrt{\frac{C(\log(|E|/A)+x)}{\log(|E|/A)}} + \frac{\log^2|E| + x\log|E|}{A}\right)$$

$$\leq \sum_{l=l_*}^{l^*-1} \mathbb{P}\left(\max_{R \in \mathcal{R}(A)} |L_{\pi_l+1}(R) - L_{\pi_l}(R)| > \sqrt{\frac{C(\log(|E|/A)+l+x)}{e^l}} + \frac{C(\log(|E|/A)+l+x)}{A}\right)$$

$$\leq \sum_{l=l_*}^{l^*-1} \frac{\exp(-x)}{l^2}$$

$$\leq 2\exp(-x).$$

Putting the three terms above together yields

$$\mathbb{P}\left(\max_{R \in \mathcal{R}(A)} L_R > 2\sqrt{\log\left(\frac{|E|}{A}\right)} + C(x+1)\right) \leq \frac{4}{\log(e|E|/A)}\exp(-x),$$

where we apply $A \gg \log^2|E|$ and the inequalities $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{a+b} \leq \sqrt{a} + b/\sqrt{a}$.

Now, we apply this bound to $A = |E|2^{-k}$, $k \geq 0$ yielding

$$\mathbb{P}\left(\max_{R \in \mathcal{R}}\left(L_R - 2\sqrt{\log\frac{|E|}{|E(R)|}}\right) > C(x+1)\right) \leq 8\exp(-x).$$

This immediately suggests that $q_a = O(1)$.

Now, let's turn to the case when a subgraph is significant, that is to prove (4.15). Assume the significant region is R_0. Using standard statistics we calculate the mean and variance of L_{R_0}

$$\mathbb{E}(L_{R_0}) = \frac{(\beta_1^{R_0} - \beta_2^{R_0})^T \Sigma_{R_0}^{-1}(\beta_1^{R_0} - \beta_2^{R_0})}{\sqrt{|E(R_0)|}} \quad \text{and} \quad Var(L_{R_0}) = 2 + 4\frac{(\beta_1^{R_0} - \beta_2^{R_0})^T \Sigma_{R_0}^{-1}(\beta_1^{R_0} - \beta_2^{R_0})}{|E(R_0)|}.$$

By Chebyshev's inequality, we have

$$\mathbb{P}\left(\frac{|L_{R_0} - \mathbb{E}(L_{R_0})|}{\sqrt{Var(L_{R_0})}} > x\right) \leq \frac{1}{x^2}. \tag{A.10}$$

If $\left(\beta_1^{R_0} - \beta_2^{R_0}\right)^T \Sigma_{R_0}^{-1}\left(\beta_1^{R_0} - \beta_2^{R_0}\right) \geq |E(R_0)|$, then (A.10) suggests

$$\mathbb{P}\left(L_{R_0} > \sqrt{|E(R_0)|}\right) \to 1, \quad |E| \to \infty$$

by taking x as a sequence (e.g., $\log\log(|E(R_0)|)$) which increases slow enough in (A.10). This leads to (4.15). If $\left(\beta_1^{R_0} - \beta_2^{R_0}\right)^T \Sigma_{R_0}^{-1}\left(\beta_1^{R_0} - \beta_2^{R_0}\right) < |E(R_0)|$, then $Var(L_{R_0}) < 6$. Then (4.14) and (A.10) imply

$$\mathbb{P}\left(L_{R_0} - 2\sqrt{\frac{|E|}{|E(R_0)|}} > q_\alpha\right) \to 1, \quad \text{as } |E| \to \infty.$$

□

Appendix B.: Implementation Details.

The workflow below describes one run of our model given a sparsity is specified for the oracle graph procedure.

(1) **Oracle Graph.** As noted in the main paper, we use *graphical lasso (glasso)* to generate an *oracle graph*, which allows to define structured regions (subgraphs) for scan statistics on graphs. Each element of the input matrix C in (B.1) for glasso is generated by calculating the slope for each position of the covariance matrix across the predictors for each group, and then taking the difference between the groups. The following inverse covariance estimation problem, *glasso*, is then solved using existing MATLAB interfaces to fast C implementations.

$$\Theta = \underset{\Theta \succeq 0}{\arg\min} \quad -\log|\Theta| + tr(C\Theta) + \lambda\|\Theta\|_1 \tag{B.1}$$

With sparsity parameter λ, this procedure generates a reasonably sparse *oracle graph*.

(2) **Candidate Subgraphs.** With the oracle graph in hand, we then construct the set of all ball subgraphs, as defined in Section 4 of our main paper. By limiting ourselves to only a few $(D|V|)$ subgraphs, we can perform scan statistics more efficiently.

(3) **Characterizing the Null Distribution.** In the case where we have few samples, we cannot directly apply the χ^2 result. In these cases, the null distribution is then characterized using permutation testing over all candidate subgraphs. For each subgraph the input data is permuted a number of times to generate a good representation of the distribution at that subgraph. All normalized (but not size-corrected) scan statistics are then calculated for all permutations across all subsets and then combined in order to create the null distribution.

(4) **Calculating the Test Statistic** For a specific subset of the data, the scan statistic is calculated and corrected as described in Section 4 of the main paper, over the original grouping of the data. For each group, the logitudinal-covariance GLM (3.3) is computed using the procedures in §2.3.

(5) **Region Identification.** We first identify all subsets whose statistic falls above the α-level threshold specified. Then the subset-collection procedure outlined in the main paper, developed by [30], is applied, and the non-overlapping critical regions are output.

Numerical Considerations.

In practice, our empirical covariance matrices calculated on the sample data may not be positive definite. The matrix can be rank deficient when we do not have enough linearly independent samples. In addition, we may use a rank correlation matrix in its place, which also may not be PD. To resolve this issue, we *project* the empirical covariance matrix onto the symmetric-positive definite SPD(n) manifold. We first apply a standard procedure for transforming a symmetric matrix into a symmetric positive semidefinite (SPSD) one. As described in [69], the standard eigenvalue thresholding, or clipping, $\lambda_{SPSD} = \max(0, \lambda)$ is sensible since it provides the optimal projection of any matrix onto the SPSD manifold. Let $\Sigma = U \Lambda U^\top$ be the eigenvalue decomposition of the matrix Σ. The SPSD projection of Σ is then $\text{proj}_{SPSD}(\Sigma) = U\text{diag}(\max(\lambda_1, 0), ..., \max(\lambda_n, 0))U^\top$. And so to project to the SPD(n) manifold we can simply add some epsilon to each element of the diagonal:

$$\text{proj}_{SPD}(\Sigma) = U\text{diag}(\max(\lambda_1, 0), ..., \max(\lambda_n, 0))U^\top + \epsilon I \qquad \text{(B.2)}$$

A remark on the term ϵI will be useful here. We find that in experiments, numerical problems can arise if the smallest eigenvalue of the projected matrix is too small. By iteratively adding a small ϵ until the smallest eigenvalue is above our threshold, we ensure that the matrix is positive definite for the exponential and logarithmic maps. They are necessary for moving back and forth between the manifold and the tangent space.

A note on localization accuracy.

In addition to simply checking whether or not we were able to correctly answer the hypothesis test group difference, it is important that if a significance is found, that it is found in the features that were originally used to generate the data. Using the same simulation setup as previous, we take the union of all subsets returned to be significant and check if each of the truly changing features p_t are contained within the superset.

In this particular case we find that our localization is only dependent on the graphical lasso procedure we use to generate the oracle graph. As long as the sparsity specified is large enough to include at least p_t edges, we find that in *every* simulation where we find a significant difference, the features that express the difference are a superset of the true features.

Appendix C.: Preclinical AD Extended Details and Results.

## Data and Variable Descriptions.

In our neuroimaging experiments, a large number of our features describe specific and localized regions of the brain across multiple imaging modalities. Below we list and describe each of regions for each modality, and give a brief background on each of methods used to acquire the data. We also include the list of cognitive scores used in our analysis.

PET Imaging.

Positron emission tomography has become an increasingly popular method of imaging the brain, specifically in the areas where cognitive decline can be strongly correlated with the specific matter being imaged. Pittsburgh compound B (PiB) was used as the tracer for these images, and the 16 mirrored (Left and Right) regions labeled below were selected as strongly correlated with the development and progression of Alzheimer's Disease.

(1) PiB Angular L/R

(2) PiB Cingulum Ant L/R

(3) PiB Cingulum Post L/R

(4) PiB Frontal Med Orb L/R

(5) PiB Precuneus L/R

(6) PiB SupraMarginal L/R

(7) PiB Temporal Mid L/R

(8) PiB Temporal Sup L/R

The average of the voxel values in each ROI (region of interest) of the brain are used for imaging features. The 16 regions are highlighted in Figure 8.
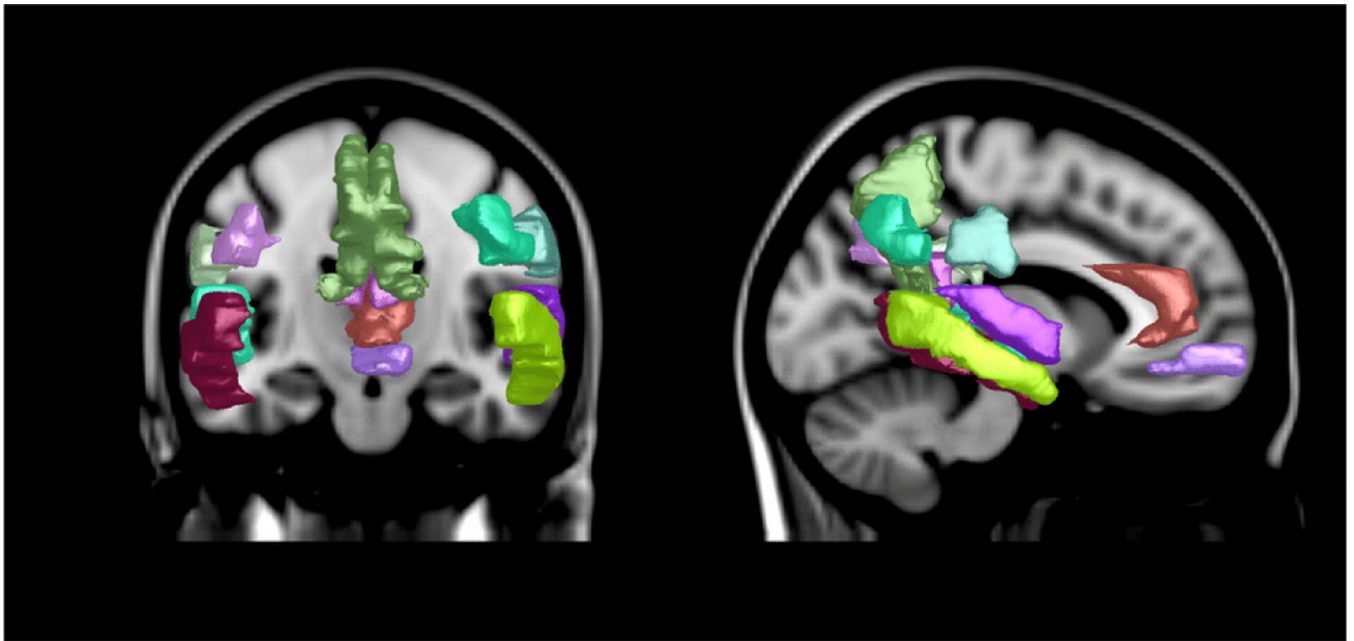
FIG. 8.
16 Positron Emission Tomography (PET) regions.
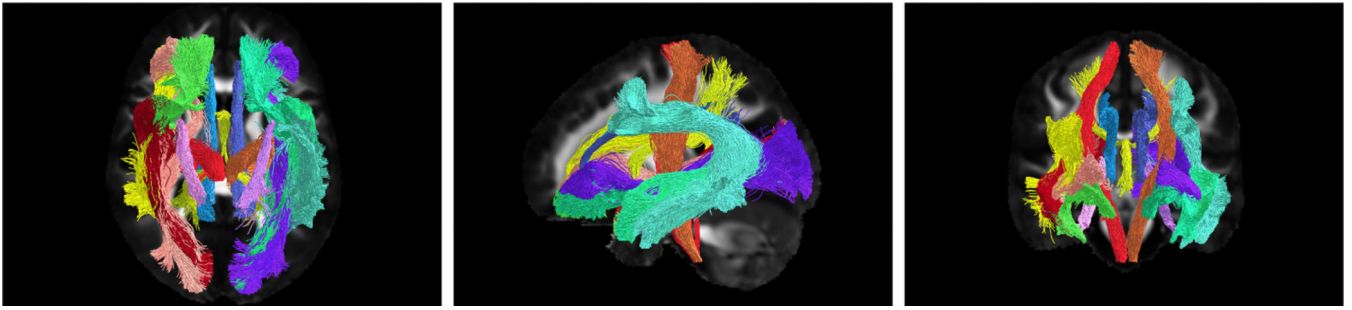
FIG. 9.
17 major DTI fiber bundles measured using Fractional Anisotropy (FA). The 48 selected for our analysis include a subset of these, which have been identified as critical regions that signal the beginnings of cognitive impairment.

DTI Imaging.

Diffusion tensor imaging is used to measure the restricted diffusion of water through and about regions of the brain. The 48 regions here are the aggregated measurements of total rates of diffusion for each voxel in that region. The two measurements, Fractional Anisotropy (FA) and Mean Diffusivity (MD) collectively well describe the diffusion in a specific region. The following is the full list of regions used in our analysis. Regions that spanned across both the left and right sides of the brain are indicated as such, and were treated as separate and independent in our analyses.

(1) Middle cerebellar peduncle

(2) Pontine crossing tract (a part of MCP)

(3) Genu of corpus callosum

(4) Body of corpus callosum

(5) Splenium of corpus callosum

(6) Fornix (column and body of fornix)

(7) Corticospinal tract R/L

(8) Medial lemniscus R/L

(9) Inferior cerebellar peduncle R/L

(10) Superior cerebellar peduncle R/L

(11) Cerebral peduncle R/L

(12) Anterior limb of internal capsule R/L

(13) Posterior limb of internal capsule R/L

(14) Retrolenticular part of internal capsule R/L

(15) Anterior corona radiata R/L

(16) Superior corona radiata R/L

(17) Posterior corona radiata R/L

(18) Posterior thalamic radiation (include optic radiation) R/L

(19) Sagittal stratum (include inferior longitidinal fasciculus and inferior fronto-occipital fasciculus) R/L

(20) External capsule R/L

(21) Cingulum (cingulate gyrus) R/L

(22) Cingulum (hippocampus) R/L

(23) Fornix (cres) / Stria terminalis (can not be resolved with current resolution) R/L

(24) Superior longitudinal fasciculus R/L

(25) Superior fronto-occipital fasciculus (could be a part of anterior internal capsule) R/L

(26) Uncinate fasciculus R/L

(27) Tapetum R/L

Cognitive Evaluations.

The battery of cognitive test scores in our analysis included a breadth of evaluations chosen specifically for their coverage of various measures of cognition. Among all tests given to the cohort, the following 17 were selected by expert clinicians and researchers in the field for their coverage and their potential value in understanding trends across groups.

(1) WAIS-III Digit Span Forward Raw Score

(2) WAIS-III Digit Span Backward Raw Score

(3) WAIS-III Letter-Number Sequencing Raw Score

(4) COWAT CFL Score

(5) Boston Naming Test Total Score

(6) RAVLT Learning Trial A1 Raw Score

(7) RAVLT Learning Trial A2 Raw Score

(8) RAVLT Learning Trial A3 Raw Score

(9) RAVLT Learning Trial A4 Raw Score

(10) RAVLT Learning Trial A5 Raw Score

(11) RAVLT Learning Trial A6 Raw Score

(12) RAVLT Delayed Recall Raw Score

(13) Stroop Word/Color-Word Scaled Score

(14) Trail-Making Test Part A

(15) Trail-Making Test Part B

(16)   Clock Drawing Test Score

(17)   Center for Epidemiologic Studies Depression Scale Score

WAIS-III.

This is the most widely used IQ test. The Digit Span examination is specifically meant to evaluate the working memory of an individual. Participants are required to attempt to recall a series of numbers in order, both forwards and backwards. Letter-Number sequencing reflects a similar idea, but with a mix of both numbers and letters in increasing and alphabetical order, and is meant to be an indicator of more complex mental control [67].

Rey Auditory Visual Learning Test.

This test is specifically meant to evaluate all aspects of memory. Each trial evaluates a different type of memory, ranging from short-term and working memory to procedural and episodic memory. [53].

Trail-Making Test.

This is a very popular test in providing information about executive function in the brain. The test consists of drawing lines among a randomly generated set of points in a square, where each point is labeled with a number. In Part A, participants must 'connect the dots' in increasing numerical order, and in Part B in increasing numerical and alphabetical order. The score on the test is primarily dictated by the time in seconds it takes to complete the task for 25 of these 'dots.' More background information and normative analyses can be found in [62].

Other tests similarly measure various cognitive function. While the Depression Scale Score did not crop up in any of our analyses here, it has been shown that depression is strongly associated with AD-related decline [68].

Detailed Imaging with Cognitive Tests Results.

In the following tables we provide additional details of the statistical test we performed on the preclinical AD cohort. Each set contains a set of features found to display significant group difference (at the p 0.05 level) along the covariance trajectory divided by the group variable indicated.

While some of these associations are well-known, few have been indicated as novel by AD researchers and clinicians, and to be of interesting value for further analysis.

Detailed results on larger cohort with only cognitive scores.

We also applied our method to a larger cohort consisting of approximately 1500 subjects with varying temporal measurements on the battery of cognitive tests. Each individual had approximately 3 visits worth of data, and so our total number of measurements was approximately

TABLE 6.

Group difference across Amyloid Load (PiB Positivity)

Amyloid Load (PiB Positivity)		
Set 1	PiB Angular L/R	PiB Cingulum Ant L/R
	PiB Cingulum Post L/R	PiB Frontal Med Orb L/R
	PiB Precuneus L/R	PiB Temporal Sup L/R
	PiB Temporal Mid L/R	**PiB SupraMarginal L**
Set 2	FA Cerebral peduncle R	FA Cerebral peduncle L
	MD Corticospinal tract R	MD Corticospinal tract L
	Trail-Making Test Part A Score	MD Cerebral peduncle R
	PET Cingulum Post R	

TABLE 7.

Group difference in gender

Gender		
Set 1	Rey Audio and Verbal Learning Test	FA Cingulum L
Set 2	FA Medial lemniscus L	FA Cingulum (hippocampus) L
	FA Posterior thalamic radiation (include optic radiation) L	
Set 3	FA Corticospinal tract R	FA Superior fronto-occipital fasciculus R

TABLE 8.

Group difference across Genotype APOE4 expression

Genotype: APOE4		
Set 1	Digit Span Backward Raw	Score Stroop Color-word
	PiB Cingulum Post L	PiB Cingulum Post R
	PiB Frontal Med Orb L	PiB Frontal Med Orb R
	PiB Precuneus L	PiB Precuneus R
	PiB SupraMarginal	PiB Temporal Mid R

TABLE 9.

Group difference across Expert MCI Diagnosis

Consensus Conference		
Set 2	Digit Span Backward Raw Score	Stroop Color-word
	PiB Cingulum Post L	PiB Cingulum Post R
	PiB Frontal Med Orb L	PiB Frontal Med Orb R
	PiB Precuneus L	PiB Precuneus R

Consensus Conference	
PiB SupraMarginal	PiB Temporal Mid R

$n = 4000$. In addition to the groupings used above, we were able to use an algorithmic cognitive impairment (ACI) measure to further evaluate the model against a factor which is known to be group-separating. Below are the tabulated feature sets identified by our model for each of the group separations described in the main paper. In this case to increase interpretability of the results we limited our search to groups of 3–6 features.

When grouped by genotype, the most indicative subset as shown in Table 11. These tests are most closely associated with memory, and we see that no tests of executive function or spatial ability (Trail-Making or Clock Drawing) were included.

In addition to an algorithmic measure of impairment, a conference of expert clinicians and researchers have given each individual a clinical impairment diagnosis for each time they underwent the cognitive battery. Using this as a group separator, we found a large number of overlapping subsets that displayed significant group difference at the $p = 0.05$ level. These are shown in Table 12. Trail-Making Test Parts A and B appeared in all identified subsets.

TABLE 10.

Group Difference Localization Across Algorithmic Impairment

Algorithmic Cognitive Impairment		
Set 1	Boston Naming Test Total Score	RAVLT Learning Trial A1 Raw Score
	RAVLT Learning Trial A6 Raw Score	

TABLE 11.

Group Difference Localization Across ApoE4 Genotype

Genotype: ApoE4		
Set 1	WAIS-III Digit Span Backward Raw Score	RAVLT Learning Trial A3 Raw Score
	RAVLT Learning Trial A4 Raw Score	RAVLT Learning Trial A5 Raw Score

TABLE 12.

Group Difference Localization Across Expert Clinical Diagnosis

Expert Consensus Measure	
WAIS-3 Letter-Number Sequencing Raw Score	Boston Naming Test Total Score
RAVLT Learning Trial A2 Raw Score	RAVLT Learning Trial A3 Raw Score
RAVLT Learning Trial A4 Raw Score	RAVLT Learning Trial A5 Raw Score
RAVLT Learning Trial A6 Raw Score	RAVLT Delayed Recall Raw Score
Trail-Making Test Part A	Trail-Making Test Part B
Clock Drawing Test Score	Center for Epidemiologic Studies Depression Scale Score

FIG. 10.

Histograms of the Delayed Recall Scores for all time points for the ~ 4000 individual measurements across different group separations. We note in particular that the results found from the genotype separation above would have been hard to identify since given the distributions are extremely overlapping (top left) for this particular separation.

Appendix D.: Differential Geometry Basics and Notes.

We briefly introduce notions that we used in the main paper. For more details, we refer the reader to [12, 37, 56].

Differentiable manifold.

A *differentiable (smooth) manifold* of dimension n is a set $\mathcal{M}$ and a maximal family of *injective* mappings $\varphi_i : U_i \subset \mathbf{R}^n \to \mathcal{M}$ of open sets U_i of $\mathbf{R}^n$ into $\mathcal{M}$ such that:

(1) $\cup_i \varphi_i(U_i) = \mathcal{M}$

(2) for any pair i, j with $\varphi_i(U_i) \cap \varphi_j(U_j) = W \neq \phi$, the sets $\varphi_i^{-1}(W)$ and $\varphi_j^{-1}(W)$ are open sets in $\mathbf{R}^n$ and the mappings $\varphi_j^{-1} \circ \varphi_i$ are differentiable, where $\circ$ denotes function composition.

3. The family $\{(U_i, \varphi_i)\}$ is maximal relative to the conditions (1) and (2).

Roughly speaking, a differentiable (smooth) manifold $\mathcal{M}$ is a topological space that is locally similar to Euclidean space and has a globally defined differential structure.

Tangent space $(T_p\mathcal{M})$.

The *tangent space* at $p \in \mathcal{M}$ is the vector space, which consists of the tangent vectors of *all* possible curves passing through p.

Tangent bundle $(T\mathcal{M})$.

The *tangent bundle* of $\mathcal{M}$ is the disjoint union of tangent spaces at all points of $\mathcal{M}, T\mathcal{M} = \mathrm{I}_{p \in \mathcal{M}} T_p\mathcal{M}$. The tangent bundle is equipped with a natural *projection map* $\pi : T\mathcal{M} \to \mathcal{M}$.

Riemannian manifold.

A *Riemannian manifold* is equipped with a smoothly varying metric (inner product), which is called *Riemannian metric*.

Various geometric notions, e.g., the angle between two curves or the length of a curve, can be extended on the manifold.

Geodesic curves.

A geodesic curve on a Riemannian manifold is the locally shortest (distance-minimizing) curve. These are analogous to straight lines in Euclidean space and a main object to generalize linear models to Riemannian manifolds.

Geodesic distance.

The *geodesic distance* between two points on $\mathcal{M}$ is the length of the shortest *geodesic* curve connecting the two points. More generally, distance between two points on Riemannian manifolds is defined by the infimum of the length of all differentiable curves connecting the two points. Let γ be a continuously differentiable curve $\gamma : [a, b] \to \mathcal{M}$ between p and q in $\mathcal{M}$ and g be a metric tensor in Mc. Then, formally, the distance between p and q is defined as

$$\mathrm{d}(p, q) := \inf_{\gamma} \int_a^b \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt \qquad \text{(D.1)}$$

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

Exponential map.

An exponential map is a map from a tangent space $T_p\mathcal{M}$ to $\mathcal{M}$, which is usually locally defined due to the existence and uniqueness of ordinary differential equation for the map. The geodesic curve from y_i to y_j can be parameterized by a tangent vector in the tangent space at y_i with an exponential map $\mathrm{Exp}(y_i, \cdot) : T_{y_i}\mathcal{M} \to \mathcal{M}$.

Logarithm map.

The inverse of the exponential map is the *logarithm map*, $\mathcal{M} \to T_{y_i}\mathcal{M}$. For completeness, Table 13 shows corresponding operations in the Euclidean space and Riemannian manifolds. In the main paper, for the readability when operations are multiply nested, exponential map and its inverse logarithm map are denoted by $\mathrm{Exp}(p, x)$ and $\mathrm{Log}(p, v)$ respectively, where $p, x \in \mathcal{M}$ and $v \in T_p\mathcal{M}$. They are usually denoted $\exp_p(x)$ and $\log_p(v)$ in most of differential geometry books.

Separate from the above notations, matrix exponential, i.e, $\exp(X) := \sum \frac{1}{k!} X^k$, where $0! = 1$ and $X^0 = I$ and matrix logarithm are denoted by as $\exp(\cdot)$ and $\log(\cdot)$.

Intrinsic mean.

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

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

which may be an arithmetic, geometric or harmonic mean depending on $d(\cdot, \cdot)$. A Karcher mean is a local minimum to (D.2) and a global minimum is referred as a Fréchet mean.

On manifolds, the Karcher mean satisfies $\sum_{i=1}^{N} \log_{\bar{y}} y_i = 0$.

Algorithm 1 : Karcher mean

Input: $y_1, \ldots, y_N \in \mathcal{M}, \alpha$

Output: $\bar{y} \in \mathcal{M}$

$\bar{y}_0 = y_1$

while $\| \sum_{i=1}^{N} \mathrm{Log}(\bar{y}_k, y_i) \| > \epsilon$ **do**

$\Delta \bar{y} = \frac{\alpha}{N} \sum_{i=1}^{N} \mathrm{Log}(\bar{y}_k, y_i)$

$\bar{y}_{k+1} = \mathrm{Exp}(\bar{y}_k, \Delta \bar{y})$

end while

FIG. 11.
Karcher mean on manifolds

This identity implies the first order necessary condition of (D.2), i.e., $\bar{y}$ is a local minimum with a zero norm gradient [32]. In general, on manifolds, the existence and uniqueness of th.e Karcher mean is not guaranteed unless we assume, for uniqueness, that the data is in a small neighborhood.

Parallel transport.

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

Geometry of SPD manifolds.

Covariance matrices are symmetric positive definite matrices. Let SPD(n) be a manifold for symmetric positive definite matrices of size $n \times n$. This forms a quotient space $GL(n)/O(n)$, where $GL(n)$ denotes the general linear group

TABLE 13.

Basic operations in Euclidean space and Riemannian manifolds.

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

(the group of ($n \times n$) nonsingular matrices) and $O(n)$ is the orthogonal group (the group of ($n \times n$) orthogonal matrices). The inner product of two tangent vectors $u, v \in T_p\mathcal{M}$ is given by

$$\langle u, v \rangle_p = \mathrm{tr}(p^{-1/2}up^{-1}vp^{-1/2}) \tag{D.3}$$

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

The exponential map and logarithm map are given as

$$\mathrm{Exp}(p, v) = p^{1/2}\exp\left(p^{-1/2}vp^{-1/2}\right)p^{1/2}, \quad \mathrm{Log}(p, q) = p^{1/2}\log\left(p^{-1/2}qp^{-1/2}\right)p^{1/2}. \tag{D.4}$$

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

$$\Gamma_{p \to q}(w) = p^{1/2}rp^{-1/2}wp^{-1/2}rp^{1/2} \tag{D.5}$$

where $r = \exp\left(p^{-1/2}\frac{v}{2}p^{-1/2}\right)$ and $v = \mathrm{Log}(p, q)$

REFERENCES

[1]. Albert Marilyn S, DeKosky Steven T, Dickson Dennis, Dubois Bruno, Feldman Howard H, Fox Nick C, Gamst Anthony, Holtzman David M, Jagust William J, Petersen Ronald C, et al. , The diagnosis of mild cognitive impairment due to alzheimers disease: Recommendations from the national institute on aging-alzheimers association workgroups on diagnostic guidelines for alzheimer's disease, Alzheimer's & dementia 7 (2011), no. 3, 270–279.

[2]. Albert Marilyn S, Moss Mark B, Tanzi Rudolph, and Jones Kenneth, Preclinical prediction of ad using neuropsychological tests, Journal of the International Neuropsychological Society 7 (2001), no. 05, 631–639. [PubMed: 11459114]

[3]. Arias-Castro E, Candès EJ, et al. , Detection of an anomalous cluster in a network, The Annals of Statistics (2011), 278–304.

[4]. Arora Sanjeev and Barak Boaz, Computational complexity: a modern approach, Cambridge University Press, 2009.

[5]. Banerjee M, Chakraborty R, et al. , Nonlinear regression on Riemannian manifolds and its applications to neuro-image analysis, MICCAI, 2015, pp. 719–727.

[6]. Belilovsky Eugene, Varoquaux Gaël, and Blaschko Matthew B, Hypothesis testing for differences in gaussian graphical models: Applications to brain connectivity, arXiv preprint arXiv:1512.08643 (2015).

[7]. Cai T, Liu W, et al. , A constrained l1 minimization approach to sparse precision matrix estimation, JASA 106 (2011), no. 494, 594–607.

[8]. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GWet al, Roses AD, Haines JL, and Pericak-Vance M Al, Gene dose of apolipoprotein e type 4 allele and the risk of alzheimers disease in late onset families, Science 261 (1993), no. 5123, 921–923. [PubMed: 8346443]

[9]. Cornea E, Zhu H, et al. , Regression models on Riemannian symmetric spaces, JRSS-B (2016).

[10]. Danaher P, Wang P, et al. , The joint graphical LASSO for inverse covariance estimation across multiple classes, JRSS-B 76 (2014), no. 2, 373–397.

[11]. Darst Burcu F, Koscik Rebecca L, Racine Annie M, Oh Jennifer M, Krause Rachel A, Carlsson Cynthia M, Zetterberg Henrik, Blennow Kaj, Christian Bradley T, Bendlin Barbara B, et al. , Pathwayspecific polygenic risk scores as predictors of amyloid-β deposition and cognitive function in a sample at increased risk for alzheimers disease, Journal of Alzheimer's Disease 55 (2017), no. 2, 473–484.

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

[13]. Douaud Gwenaëlle, Jbabdi Saâd, Behrens Timothy EJ, Menke Ricarda A, Gass Achim, Monsch Andreas U, Rao Anil, Whitcher Brandon, Kindlmann Gordon, Matthews Paul M, et al. , Dti measures in crossing-fibre areas: increased diffusion anisotropy reveals early white matter alteration in mci and mild alzheimer's disease, Neuroimage 55 (2011), no. 3, 880–890. [PubMed: 21182970]

[14]. Du J, Goh A, Kushnarev S, and Qiu A, Geodesic regression on orientation distribution functions with its application to an aging study, NeuroImage 87 (2014), 416–426. [PubMed: 23851325]

[15]. Fan J, Han X, et al. , Control of the fdr under arbitrary covariance dependence, JASA (2012).

[16]. Fletcher PT, Geodesic regression and the theory of least squares on Riemannian manifolds, IJCV 105 (2013), no. 2, 171–185.

[17]. Fletcher TP and Joshi S, Riemannian geometry for the statistical analysis of diffusion tensor data, Signal Processing 87 (2007), no. 2, 250–262.

[18]. Fratiglioni Laura, Grut M, Forsell Y, Viitanen M, Grafström M, Holmen K, Ericsson K, Bäckman L, Ahlbom A, and Winblad B, Prevalence of alzheimer's disease and other dementias in an elderly urban population relationship with age, sex, and education, Neurology 41 (1991), no. 12, 1886–1886. [PubMed: 1745343]

[19]. Friedman J, Hastie T, et al. , Sparse inverse covariance estimation with the graphical LASSO, Biostatistics 9 (2008), no. 3, 432–441. [PubMed: 18079126]

[20]. Geer SA, Empirical processes in m-estimation, vol. 6, Cambridge university press, 2000.

[21]. Grenander Ulf, Probabilities on algebraic structures, Courier Corporation, 2008.

[22]. Grenander Ulf and Miller Michael I, Computational anatomy: An emerging discipline, Quarterly of applied mathematics 56 (1998), no. 4, 617–694.

[23]. Grenander Ulf and Rosenblatt Murray, Statistical analysis of stationary time series, (1957).

[24]. Hardle Wolfgang and Mammen Enno, Comparing nonparametric versus parametric regression fits, The Annals of Statistics (1993), 1926–1947.

[25]. Hardy John and Selkoe Dennis J, The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics, Science 297 (2002), no. 5580, 353–356. [PubMed: 12130773]

[26]. Hardy John A and Higgins Gerald A, Alzheimer's disease: the amyloid cascade hypothesis, Science 256 (1992), no. 5054, 184. [PubMed: 1566067]

[27]. Hong Yi, Singh Nikhil, Kwitt Roland, and Niethammer Marc, Group testing for longitudinal data, International Conference on Information Processing in Medical Imaging, Springer, 2015, pp. 139–151.

[28]. Jack Clifford R Jr, Wiste Heather J, Vemuri Prashanthi, Weigand Stephen D, Senjem Matthew L, Zeng Guang, Bernstein Matt A, Gunter Jeffrey L, Pankratz Vernon S, Aisen Paul S, et al. , Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to alzheimers disease, Brain 133 (2010), no. 11, 3336–3348. [PubMed: 20935035]

[29]. Jayasumana S et al., Kernel methods on the Riemannian manifold of SPD matrices, CVPR, 2013.

[30]. Jeng X Jessie, Cai T Tony, and Li Hongzhe, Optimal sparse segment identification with application in copy number variation analysis, Journal of the American Statistical Association 105 (2010), no. 491, 1156–1166. [PubMed: 23543902]

[31]. Jordan MI, Learning in graphical models, vol. 89, Springer Science & Business Media, 1998.

[32]. Karcher Hermann, Riemannian center of mass and mollifier smoothing, Communications on pure and applied mathematics 30 (1977), no. 5, 509–541.

[33]. Hyunwoo J Kim Nagesh Adluru, et al., Multivariate general linear models on Riemannian manifolds with applications to statistical analysis of diffusion weighted images, CVPR, 2014, pp. 2705–2712.

[34]. Koller D and Friedman N, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[35]. Laurent Beatrice and Massart Pascal, Adaptive estimation of a quadratic functional by model selection, Annals of Statistics (2000), 1302–1338.

[36]. Lauritzen SL, Graphical models, Clarendon Press, 1996.

[37]. Lee John M, Smooth manifolds, Introduction to Smooth Manifolds, Springer, 2003, pp. 1–29.

[38]. Lee John M, Introduction to smooth manifolds, Springer, 2012.

[39]. Muriel Deutsch Lezak, Neuropsychological assessment, Oxford University Press, USA, 2004.

[40]. Lin Yi-Cheng, Shih Yao-Chia, Tseng Wen-Yih I, Chu Yu-Hsiu, Wu Meng-Tien, Chen Ta-Fu, Tang Pei-Fang, and Chiu Ming-Jang, Cingulum correlates of cognitive functions in patients with mild cognitive impairment and early alzheimers disease: a diffusion spectrum imaging study, Brain topography 27 (2014), no. 3, 393–402. [PubMed: 24414091]

[41]. Liu H, Lafferty J, et al., The nonparanormal: Semiparametric estimation of high dimensional undirected graphs, JMLR 10 (2009), 2295–2328.

[42]. Liu Han, Han Fang, et al. , High-dimensional semiparametric gaussian copula graphical models, The Annals of Statistics 40 (2012), no. 4, 2293–2326.

[43]. McArdle JJ and Bell RQ, An introduction to latent growth models for developmental data analysis, (2000).

[44]. Mori S, Oishi K, et al. , Stereotaxic white matter atlas based on diffusion tensor imaging in an icbm template, Neuroimage 40 (2008), no. 2, 570–582. [PubMed: 18255316]

[45]. Mormino Elizabeth C, Betensky Rebecca A, Hedden Trey, Schultz Aaron P, Ward Andrew, Huijbers Willem, Rentz Dorene M, Johnson Keith A, Sperling Reisa A, Alzheimer's Disease Neuroimaging Initiative, et al. , Amyloid and apoe ɛ4 interact to influence short-term decline in preclinical alzheimer disease, Neurology 82 (2014), no. 20, 1760–1767. [PubMed: 24748674]

[46]. Muralidharan Prasanna and Fletcher P Thomas, Sasaki metrics for analysis of longitudinal data on manifolds, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 1027–1034.

[47]. Naidich Thomas P, Duvernoy Henri M, Delman Bradley N, Sorensen A Gregory, Kollias Spyros S, and Haacke E Mark, Duvernoy's atlas of the human brain stem and cerebellum: high-field mri, surface anatomy, internal structure, vascularization and 3 d sectional anatomy, Springer Science & Business Media, 2009.

[48]. Qiu H, Han F, et al., Joint estimation of multiple graphical models from high dimensional time series, JRSS-B (2015).

[49]. M Racine A, Adluru N, et al. , Associations between white matter microstructure and amyloid burden in preclinical AD: a multimodal imaging investigation, NeuroImage: Clinical 4 (2014), 604–614. [PubMed: 24936411]

[50]. Rimol Lars M Ingrid, Djurovic Srdjan, Brown Andrew A, Roddey J Cooper, Kähler Anna K, Mattingsdal Morten, Athanasiu Lavinia, Joyner Alexander H, Schork Nicholas J, et al. , Sex-dependent association of common variants of microcephaly genes with brain structure, Proceedings of the National Academy of Sciences 107 (2010), no. 1, 384–388.

[51]. Roehrig Charles S, Conditions for identification in nonparametric and parametric models, Econometrica: Journal of the Econometric Society (1988), 433–447.

[52]. Sager Mark A, Hermann Bruce, and Rue Asenath La, Middle-aged children of persons with alzheimers disease: Apoe genotypes and cognitive function in the wisconsin registry for alzheimers prevention, Journal of geriatric psychiatry and neurology 18 (2005), no. 4, 245–249. [PubMed: 16306248]

[53]. Schmidt Michael et al., Rey auditory verbal learning test: a handbook, Western Psychological Services Los Angeles, 1996.

[54]. Seber George AF and Lee Alan J, Linear regression analysis. hoboken, NJ: Wiley. doi 10 (2003), 9780471722199.

[55]. Sommer S, Lauze F, et al. , Optimization over geodesics for exact principal geodesic analysis, Adv. in Comp. Math 40 (2014), no. 2, 283–313.

[56]. Spivak M, Comprehensive introduction to differential geometry, Publish or Perish, Inc., 1981.

[57]. Stehr Mark, The effect of cigarette taxes on smoking among men and women, Health Economics 16 (2007), no. 12, 1333–1343. [PubMed: 17335102]

[58]. Stdler Nicolas and Mukherjee Sach, Two-sample testing in high-dimensional models, 2012.

[59]. Su Jingyong, Kurtek Sebastian, Klassen Eric, Srivastava Anuj, et al. , Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance, The Annals of Applied Statistics 8 (2014), no. 1, 530–552.

[60]. Talagrand Michel, The generic chaining: upper and lower bounds of stochastic processes, Springer Science & Business Media, 2006.

[61]. Tanzi Rudolph E and Bertram Lars, Twenty years of the alzheimers disease amyloid hypothesis: a genetic perspective, Cell 120 (2005), no. 4, 545–555. [PubMed: 15734686]

[62]. Tombaugh Tom N, Trail making test a and b: normative data stratified by age and education, Archives of clinical neuropsychology 19 (2004), no. 2, 203–214. [PubMed: 15010086]

[63]. Tuzel Oncel, Porikli Fatih, and Meer Peter, Human detection via classification on Riemannian manifolds, Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.

[64]. Ullman JB and Bentler PM, Structural equation modeling, Wiley Online Library, 2003.

[65]. Walther G, Optimal and fast detection of spatial clusters with scan statistics, The Annals of Statistics 38 (2010), no. 2, 1010–1033.

[66]. Wang S, Fan J, et al. , Structured correlation detection with application to colocalization analysis in dual-channel fluorescence microscopic imaging, arXiv preprint arXiv:1604.02158 (2016).

[67]. Wechsler David, Wechsler adult intelligence scale–fourth edition (wais–iv), 2014.

[68]. Wragg Robin E and Jeste Dilip V, Overview of depression and psychosis in alzheimers disease, Am J Psychiatry 146 (1989), no. 5, 577–587. [PubMed: 2653053]

[69]. Wu G, Chang EY, et al., An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines, ICML, vol. 8, 2005.

[70]. Xie Y, Vemuri BC, et al., Statistical analysis of tensor fields, MICCAI, Springer, 2010, pp. 682–689.

[71]. Xue L and and others Zou H, Regularized rank-based estimation of high-dimensional nonparanormal graphical models, The Annals of Statistics 40 (2012), no. 5, 2541–2571.

[72]. Yang S, Lu Z, et al. , Fused multiple graphical lasso, SIAM J. Opt 25 (2015), no. 2, 916–943.

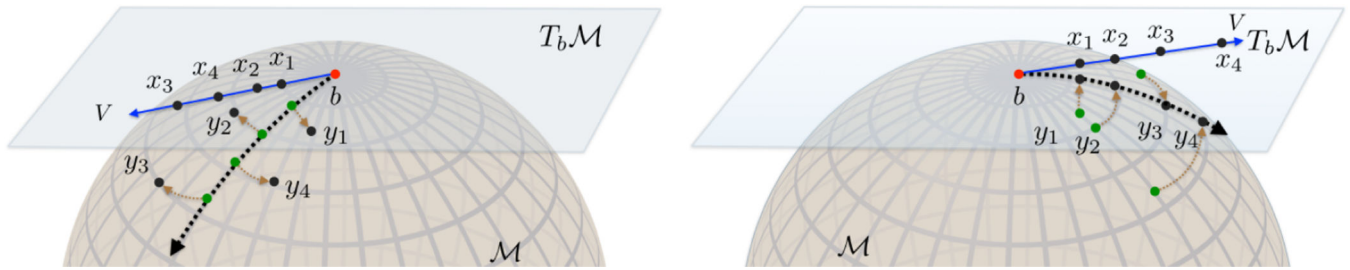[73]. Yuan M, High dimensional inverse covariance matrix estimation via LP, JMRL 11 (2010), 2261–2286.

[74]. Zhang Zhengwu, Su Jingyong, Klassen Eric, Le Huiling, and Srivastava Anuj, Rate-invariant analysis of covariance trajectories, Journal of Mathematical Imaging and Vision (2018), 1–18.

[75]. Zhou S, Lafferty J, et al. , Time varying undirected graphs, ML 80 (2010), no. 2–3, 295–319.

FIG. 1.

Group-wise MMGLM: The left and right figures represent two linear models on the SPD(p) manifold. Points x_i in the tangent space are our covariate or predictor, and points y_i in the manifold space represent SPD(p) matrices. In our regression setting, we wish to minimize the error (brown curves) between the estimation and the sample points. Because each linear model has a different base point, the trajectories cannot be directly compared as in the Euclidean setting.
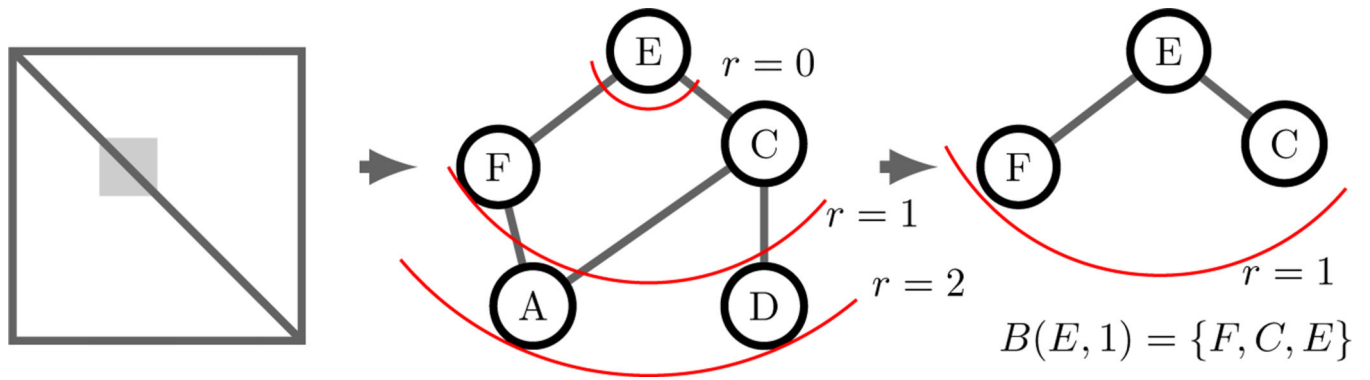
FIG. 2.
(left) A region of the sparse precision matrix, (center) The corresponding subgraph of that region, along with balls of varying radius from the root node *E*, (right) The ball subgraph constructed with *r* = 1. These subgraphs with bounded radius act as the structured regions on which scan statistics can be applied.
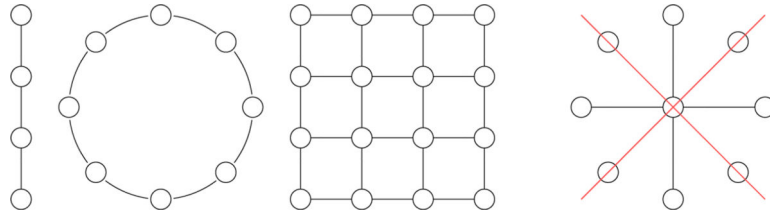
FIG. 3.

(Left) Chain, ring and 2D lattice graphs that satisfy the Avocado Assumption. (Right) Star graph that does not satisfy the property: from the center node the graph is "too dense on the outside."
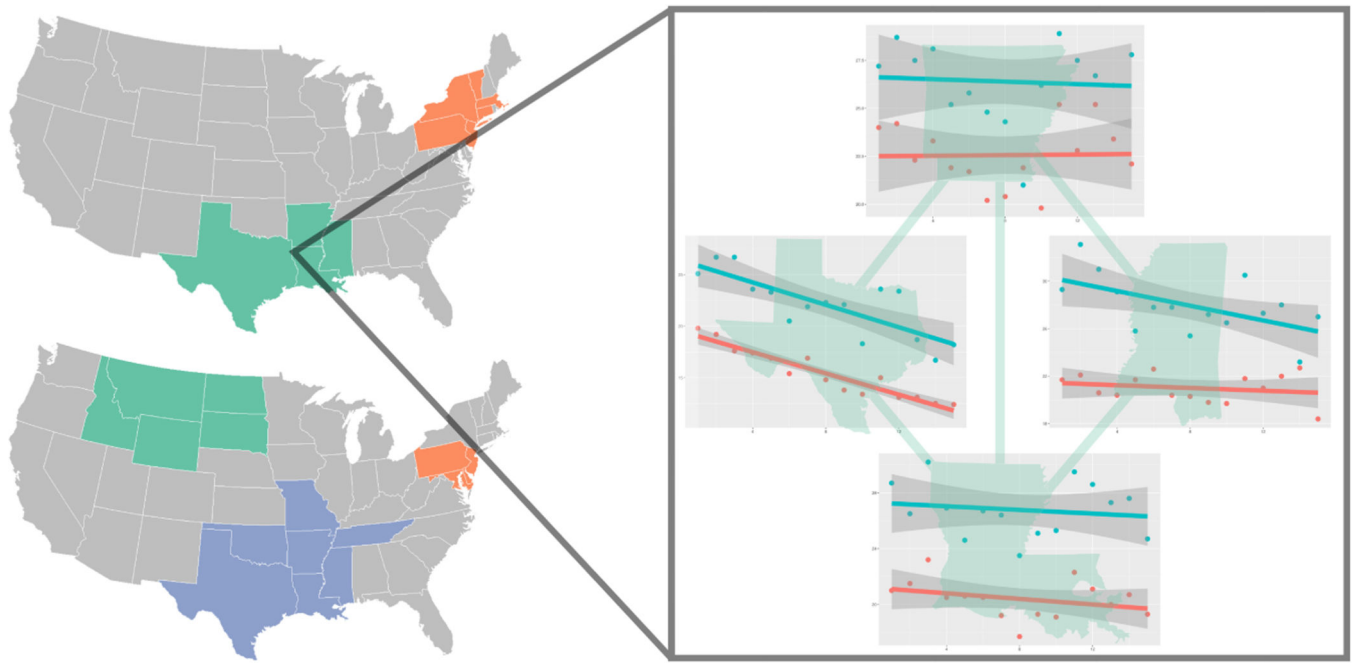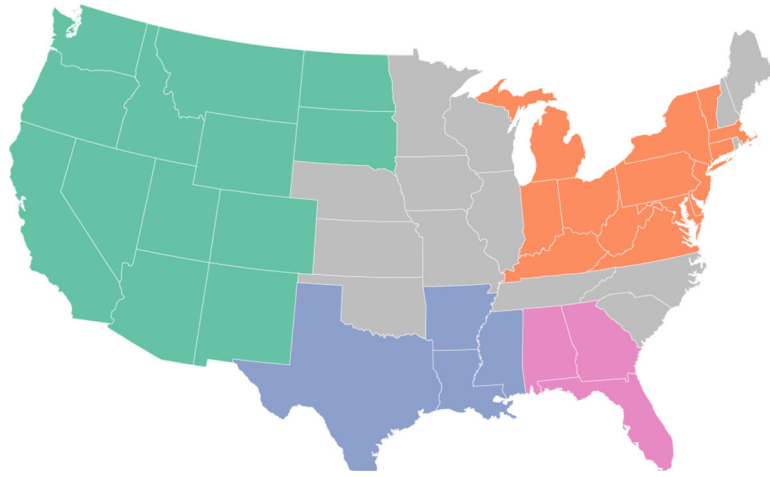
FIG. 4.
(left,top) States identified as having significantly different time-varying tobacco usage across gender from 2001 to 2015. (left,bottom) States identified as having significantly different time-varying heavy drinking use across gender from 2010 to 2015. (right) Linear regressions over tobacco usage fitted to the four states defined by the ball subgraph centered at Louisiana. Best viewed in color.
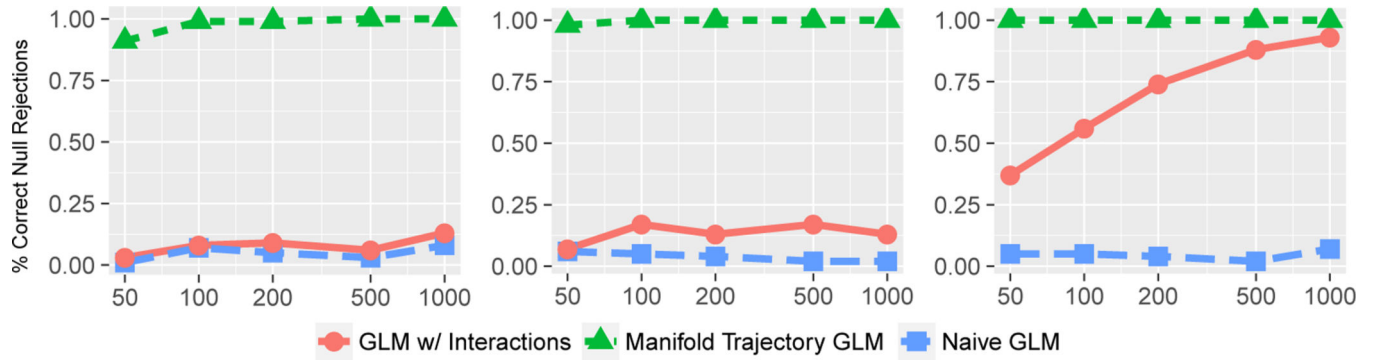
FIG. 5.

Correct null hypothesis rejections over 100 runs for three models. For $p = 50$ features, each plot shows the rejection rate for $p_t \in \{4, 8, 20\}$ (from left to right) respectively as a function of the number of sample points.

FIG. 6.
Contiguous states identified as having significantly different time-varying co-occurrences between boys and girls baby names from 1910 to 2015. Best viewed in color.
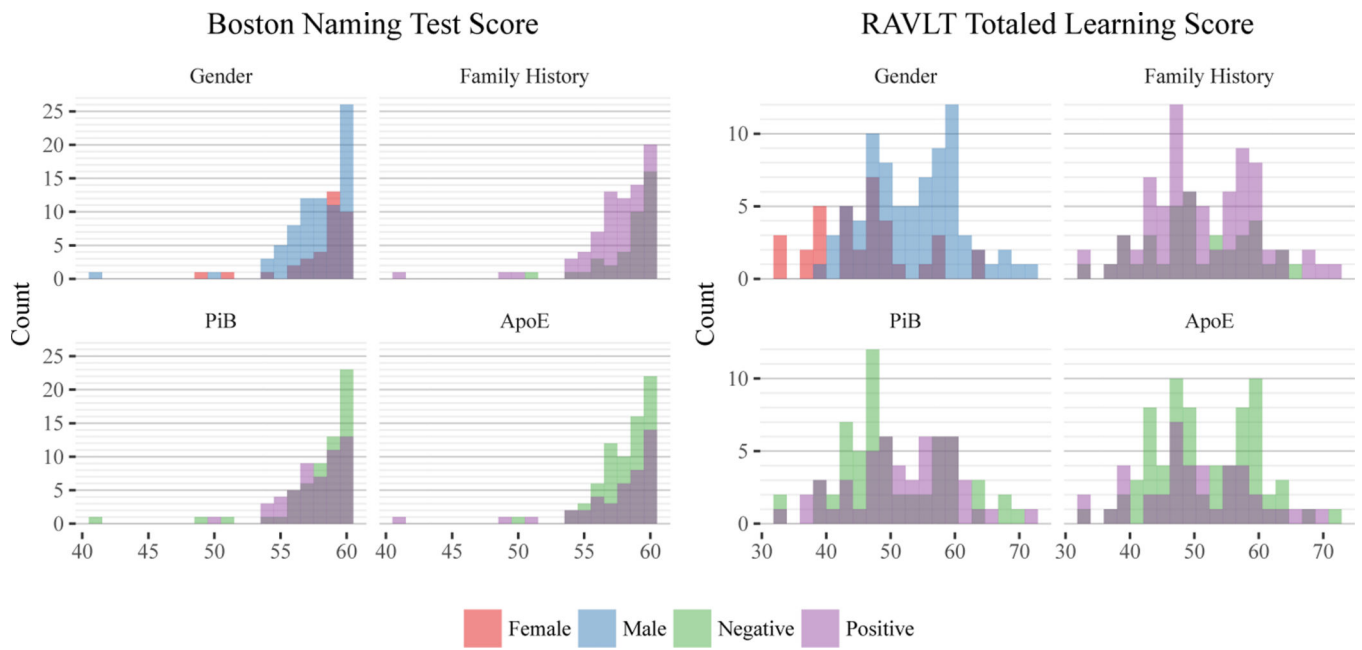
FIG. 7.

Histograms of the Boston Naming Test Scores and RAVLT Total Scores for all time points for the 114 individual measurements across different group separations. The means for each test score is not significantly different across different stratification variable.

TABLE 1.

Basic operations in Euclidean space and Riemannian manifolds.

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

TABLE 2.

Detection Accuracy of hypothesis test scheme (100 runs).

	$p_t = 5$	$p_t = 8$	$p_t = 10$	$p_t = 15$
$n = 10$	0.06	0.02	0.04	0.03
$n = 20$	0.75	0.75	0.53	0.29
$n = 50$	0.99	1.00	1.00	0.80
$n = 100$	1.00	1.00	1.00	0.95
$n = 200$	1.00	1.00	1.00	0.98
$n = 1000$	1.00	1.00	1.00	1.00

TABLE 3.

Group difference across Gender (left) and Genotype APOE4 expression (right). Three disjoint sets of features were identified as coavarying significantly differently among gender, while one larger set was identified in the genotype stratification.

Gender	
Set 1	RAVLT Total (1–5)
	FA Cingulum L
Set 2	FA Medial lemniscus L
	FA Cingulum (hippocampus) L
	FA Post thalamic radiation L
Set 3	FA Corticospinal tract R
	FA Superior O.F. fasciculus R

Genotype: APOE4	
Digit Span Backward Raw Score	Stroop Color-word Score
PiB Cingulum Post L	PiB Cingulum Post R
PiB Frontal Med Orb L	PiB Frontal Med Orb R
PiB Precuneus L	PiB Precuneus R
PiB SupraMarginal	PiB Temporal Mid R

TABLE 4.

Group difference across Amyloid Load (PiB Positivity)

Amyloid Load (PiB Positivity)		
Set 1	PiB Angular L/R	PiB Cingulum Ant L/R
	PiB Cingulum Post L/R	PiB Frontal Med Orb L/R
	PiB Precuneus L/R	PiB Temporal Sup L/R
	PiB Temporal Mid L/R	**PiB SupraMarginal L**
Set 2	FA Cerebral peduncle R	FA Cerebral peduncle L
	MD Corticospinal tract R	MD Corticospinal tract L
	Trail-Making Test Part A Score	MD Cerebral peduncle R
	PET Cingulum Post R	

TABLE 5.

Group difference localization across expert clinical diagnosis. With significantly more samples and a larger set of cognitive tests, those above were identified as significantly different across the expert consensus measure.

Expert Consensus Diagnosis	
WAIS-3 LNS Raw Score	Boston Naming Test Total Score
RAVLT A2 Raw Score	RAVLT A3 Raw Score
RAVLT A4 Raw Score	RAVLT A5 Raw Score
RAVLT A6 Raw Score	RAVLT Delayed Recall Raw Score
Trail-Making Test Part A	Trail-Making Test Part B
Clock Drawing Test Score	CES Depression Scale Score