University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2016 *Paper* 48

Distance-Based Analysis of Variance for Brain Connectivity

Russell T. Shinohara* Haochang Shou† Marco Carone‡
Robert Schultz** Birkan Tunc††
Drew Parker‡‡ Ragini Verma§

*Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, rshi@upenn.edu

†Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, hshou@mail.med.upenn.edu

‡Department of Biostatistics, University of Washington, mcarone@uw.edu

**Center for Autism Research, The Children's Hospital of Philadelphia

††Department of Radiology, Perelman School of Medicine, University of Pennsylvania

‡‡Department of Radiology, Perelman School of Medicine, University of Pennsylvania

§Department of Radiology, Perelman School of Medicine, University of Pennsylvania

http://biostats.bepress.com/upennbiostat/art48

Distance-Based Analysis of Variance for Brain Connectivity

Russell T. Shinohara, Haochang Shou, Marco Carone, Robert Schultz, Birkan Tunc, Drew Parker, and Ragini Verma

Abstract

The field of neuroimaging dedicated to mapping connections in the brain is increasingly being recognized as key for understanding neurodevelopment and pathology. Networks of these connections are quantitatively represented using complex structures including matrices, functions, and graphs, which require specialized statistical techniques for estimation and inference about developmental and disorder-related changes. Unfortunately, classical statistical testing procedures are not well suited to high-dimensional testing problems. In the context of global or regional tests for differences in neuroimaging data, traditional analysis of variance (ANOVA) is not directly applicable without first summarizing the data into univariate or low-dimensional features, a process that may mask salient features of the high-dimensional distributions. In this work, we consider a general framework for two-sample testing of complex structures by studying generalized within- and between-group variances based on distances between complex and potentially high-dimensional observations. We derive an asymptotic approximation to the null distribution of the ANOVA test statistic, and conduct simulation studies with scalar and graph outcomes to study finite sample properties of the test. Finally, we apply our test to our motivating study of structural connectivity in autism spectrum disorder.

Distance-Based Analysis of Variance for Brain Connectivity

Russell T. Shinohara*
Haochang Shou
Department of Biostatistics and Epidemiology
University of Pennsylvania Perelman School of Medicine

Marco Carone
Department of Biostatistics
University of Washington

Robert Schultz
Center for Autism Research
The Children's Hospital of Philadelphia

Birkan Tunc
Drew Parker
Ragini Verma
Department of Radiology,
University of Pennsylvania Perelman School of Medicine

August 27, 2016

Abstract

The field of neuroimaging dedicated to mapping connections in the brain is increasingly being recognized as key for understanding neurodevelopment and pathology. Networks of these connections are quantitatively represented using complex structures including matrices, functions, and graphs, which require specialized statistical techniques for estimation and inference about developmental and disorder-related changes. Unfortunately, classical statistical testing procedures are not well suited to high-dimensional testing problems. In the context of global or regional tests for differences in neuroimaging data, traditional analysis of variance (ANOVA) is not directly applicable without first summarizing the data into univariate or low-dimensional features, a process that may mask salient features of the high-dimensional distributions. In this work, we consider a general framework for two-sample testing of complex structures by studying generalized within- and between-group variances based on distances between complex and potentially high-dimensional observations. We derive an asymptotic approximation

*The authors gratefully acknowledge R01NS085211 from the National Institute of Neurological Disorders and Stroke and R21MH098010 from the National Institute of Mental Health. This work represents the opinions of the researchers and not necessarily that of the granting organizations.

to the null distribution of the ANOVA test statistic, and conduct simulation studies with scalar and graph outcomes to study finite sample properties of the test. Finally, we apply our test to our motivating study of structural connectivity in autism spectrum disorder.

Keywords: Biostatistics, neuroimaging, distance statistics, kernel ANOVA.

1 Introduction

Connectomics, the field of neuroimaging dedicated to mapping connections in the brain, is increasingly being recognized as key for understanding neurodevelopment. As the brain develops, it becomes a complex system with many interconnected networks that coactivate. Complementary imaging modalities provide insight into the structural and functional networks of the brain. Structural networks exist as physically connected regions of the brain, and functional networks are groups of regions that tend to function together. The study of these complex networks is crucial for developing cognitive and pharmaceutical therapies for treating psychiatric, neurological, and developmental disorders. Quantitatively, these networks are often represented using complex structures including matrices, functions, and graphs, and require specialized statistical techniques for estimation and inference about developmental and disorder-related changes.

Principled statistical methods are necessary for analyzing these structures due to the scale of their dimensionality, as simplistic connection-wise methods are hindered in power by the need for multiple comparison correction. Unfortunately, classical statistical testing procedures are not well suited to high-dimensional testing problems. In the context of global or regional tests for differences in neuroimaging data, traditional analysis of variance (ANOVA) is not directly applicable without first summarizing the data into univariate or low-dimensional features, a process that may mask salient features of the high-dimensional distributions. Direct statistical inference on the imaging objects is fundamentally hindered by the lack of a precise definition of variance in high-dimensional data, which in turn hampers the comparison of within- to between-group variability. In this paper, we propose a general framework for two-sample testing of complex structures by studying generalized within- and between-group variances based on distances between high-dimensional observations.

The methods we propose are closely related to the fields of distance statistics and kernel testing, which have been shown to be equivalent in many cases (Sejdinovic et al. 2013). Similarly to our proposed methodology, both of these literatures center on the reduction of the observed data using a distance (or kernel) to describe the dissimilarity between subjects. Kernel tests have been used extensively in statistical genetics, and were pioneered in association studies by Kwee et al. (2008). These score-based tests, which use the high-dimensional genetic data as predictors and scalar outcomes, have been used in the context of common

and rare variant analyses (Wu et al. 2011, Ionita-Laza et al. 2013), and are recognized as an important tool for the analysis of sequencing data. The limiting distribution of kernel test statistics is well understood (Zhang & Lin 2003). Another approach in genetic analyses involves sum tests (Wang & Elston 2007, Pan 2009), which are based on the assumption that all genetic predictors have the same association with the outcome, and sum tests study this common association parameter using weighted sums of the predictors. More recently, much work has centered on developing versions of these tests that adaptively choose these weights and combine kernels to optimize power (Lee et al. 2012, Ionita-Laza et al. 2013, Pan et al. 2014, Zhao et al. 2015) to detect both sparse and dense alternatives. While both kernel and sum tests benefit from the convenience of linear model specification for the sake of adjusting for confounding variables, their performance under model misspecification is not well understood. Finally, distance correlation (Székely et al. 2007, 2009) is an alternate measure of multivariate dependence for high-dimensional random vectors. Leveraging a distance measure to generalize Pearson correlation, distance correlation provides a means for assessing nonlinear complex correlations and testing independence. Distance correlation, which was proposed for the Euclidean random vector observation case, has also been shown to be related to kernel-based maximum mean discrepancy tests popular in the machine learning literature (Sejdinovic et al. 2013).

For two-sample testing in higher dimensions, distance-based ANOVA considers the partitioning of sums of squared distances between subjects. Although pioneered in the ecology literature (McArdle & Anderson 2001), distance-based ANOVA is increasingly being used in genomics (Minas et al. 2011) and neuroimaging (Reiss et al. 2010). This approach uses a pseudo-F statistic that assesses the ratio of the within-group distances to the between-group distances. Unfortunately, the null distribution of this test statistic is not easily approximated and thus inference has been based solely on Monte Carlo approximations of the permutation distribution (PERMANOVA). This is computationally intensive and suboptimal in terms of statistical power. The potential inefficiency of permutation-based testing stems from the flexible estimation strategy for the null distribution. Recent work by Minas & Montana (2014) developed an analytical approximation to the permutation null distribution which promises much improved computational time with similar power. While PERMANOVA tests are closely related to distance correlation under specific choices of distance functions in the case of random vectors, their extension to the case of more complex outcome structures has not yet been formalized. In the remainder of this paper,

we propose an analytical approach for testing for differences in the distribution of complex structures leveraging the PERMANOVA framework. In the next section, we describe this test in detail and derive its limiting distribution. In Section 3, we conduct simulation analyses in two different setting: one with a scalar outcome, and another with a graphical outcome. Finally, we consider the motivating study of structural connectivity in the brains of subjects with autism spectrum disorders and conclude with a discussion.

2 Proposed methodology

For a prototypical patient, the data unit is $X := (M, D)$, where M is an object in some space $\mathscr{M}$ and $D \in \{0, 1, \ldots, K-1\}$ is a disease group indicator. Denote by P_0 the true distribution of X. Suppose that $r : \mathscr{M} \times \mathscr{M} \to [0, +\infty)$ is a symmetric discrepancy function that quantifies in some scientific context-dependent manner how dissimilar two given objects in $\mathscr{M}$ are. Suppose that we observe n independent draws $X_1 := (M_1, D_1), X_2 := (M_2, D_2), \ldots, X_n := (M_n, D_n)$ from P_0 – each of these correspond to measurements taken on a different patient.

Our goal is to use the available data to determine whether the distribution of M is the same across all subpopulations defined by D. This corresponds to the null hypothesis $\mathscr{H}_0$ wherein M and D are independent under P_0. To test this hypothesis, we will use the classical partitioning of variation approach from the standard ANOVA setting. A central quantity in the developments to follow is the $r-$variance of a population of objects $M \in \mathscr{M}$ with distribution P, defined as $\sigma^2(P) := E_{P \times P}\{r(M_1, M_2)\} = \iint r(m_1, m_2) dP(m_1) dP(m_2)$. This generalized variance based on the marginal distribution of P_0 can be estimated consistently using the U-statistic

$$\sigma_n^2 := \binom{n}{2}^{-1} \sum_{i<j} r(M_i, M_j) \ .$$

This definition for the variance of an object is particularly convenient because it generalizes the usual notion of variance in an interpretable fashion and its natural U-statistic estimator lends itself to relatively simple theoretical analysis. We note that although the computation of the above sum may be burdensome in large samples, a Monte Carlo approximation based on randomly sampling pairs of subjects can easily be used. We wish to construct an ANOVA-like test statistic using r-variance. Denoting by P_{0*} and P_{0k}, respectively, the

marginal distribution of M and the conditional distribution of M given $D = k$ implied by P_0, and by π_{0k} the marginal probability that $D = k$ under P_0, we observe that the ANOVA discrepancy

$$T(P_0) := \frac{\sigma^2(P_{0*}) - \sum_{k=0}^{K-1} \pi_{0k}\sigma^2(P_{0k})}{\sum_{k=0}^{K-1} \pi_{0k}\sigma^2(P_{0k})}$$

is identically zero under $\mathscr{H}_0$. A sample version of this discrepancy based on U-statistics can be used as the test statistic. To construct this statistic, we define the within-group sum of squared distances as

$$SSE_n := \sum_{k=0}^{K-1} (n_k - 1)\binom{n_k}{2}^{-1} \sum_{i<j} r(M_i, M_j) I(D_i = D_j = k)$$

and the total sum of squared distances $SS_n := (n-1)\binom{n}{2}^{-1} \sum_{i<j} r(M_i, M_j)$, where we set $n_k := \sum_i I(D_i = k)$. The scaled quantities SSE_n/n and SS_n/n are nonparametric estimators of $\sum_{k=0}^{K-1} \pi_{0k}\sigma^2(P_{0k})$ and $\sigma^2(P_{0*})$, respectively. The difference $SST_n := SS_n - SSE_n$ corresponds to an analogue of the between-group sum of squares. An empirical version of $T(P_0)$ is given by

$$T_n := \frac{1}{n} \cdot \left(\frac{SST_n}{K-1}\right) \Big/ \left(\frac{SSE_n}{n-K}\right),$$

which can be seen as a scaled F-like test statistic. The unscaled counterpart of this statistic, $Q_n := nT_n$, is known as the distance-based pseudo-F statistic. It was first proposed by Anderson (1963), who noted that it reduces to the classical ANOVA F-test in the scalar Euclidean case. However, in the general case the test statistic does not follow an F distribution, and Anderson (1963) suggested randomly permuting the group labels D_i to approximate sampling from the null distribution. Minas et al. (2011) suggested a similar statistic with different degrees of freedom in the numerator and denominator. Recently, Minas & Montana (2014) developed a Pearson type III approximation for the permutation distribution of the F-statistic, which significantly reduces the computational burden of testing. These methods all potentially suffer from a loss of power from the permutation-based approximation – using simulations, we investigate this phenomenon in Section 3. When M is a random vector in $\mathbb{R}^p$ and r is the Euclidean norm, the numerator of the proposed pseudo-F statistic is closely related to the empirical distance covariance (Székely et al. 2009) between M and D observations, which is a weighted combination of the average within- and

between-group distances. However, for more complex structures such as those of interest in connectomics, the ability to use more general distance functions and to accommodate complex data objects is crucial.

Large positive values of the test statistic Q_n are incompatible with the null hypothesis of independence between M and D. To determine appropriate test cutoffs to use in practice, we require a better understanding of the distribution of Q_n under the null hypothesis. The U-statistic form of the sample $d-$variance can be leveraged to obtain a distributional approximation to the various building blocks of the pseudo-F statistic under the null hypothesis. We begin by studying the large-sample behavior of both SSE_n and SST_n. To simplify the presentation, we will only explicitly consider the case $K = 2$ hereon, although all results to follow can be easily generalized for arbitrary $K \in \mathbb{N}$. Below, we denote by π_0 the probability $P_0(D = 1)$.

Theorem. *Suppose that $\sigma_0^2 < +\infty$, that M and D are independent, and that $\pi_0 \in (0, 1)$. Then, SSE_n/n tends to σ_0^2 in probability and*

$$SST_n - \sigma_0^2 - n \cdot \binom{n}{2}^{-1} \sum_{i<j} u(M_i, M_j)$$

tends to zero in probability for some first-order non-degenerate kernel u. As such, in large samples, the distribution of SST_n is approximated by that of the infinite series

$$\sigma_0^2 + \sum_{j=1}^{\infty} \lambda_j \left(Z_j^2 - 1\right),$$

where $Z_1, Z_2, \ldots$ are independent standard normal variates and $\lambda_1, \lambda_2, \ldots$ are eigenvalues of the operator that maps any given function g into $m \mapsto \int u(m, m_2) g(m_2) dP(m_2)$.

Using this description of the behavior of the key building blocks of the pseudo-F statistic Q_n, we are able to describe the large-sample properties of the latter.

Corollary. *Suppose that $\sigma_0^2 < +\infty$, that M and D are independent, and that $\pi_0 \in (0, 1)$. We then have that Q_n tends in distribution to*

$$Z^* := 1 + \sigma_0^{-2} \sum_{j=1}^{\infty} \lambda_j \left(Z_j^2 - 1\right),$$

where $Z_1, Z_2, \ldots$ are independent standard normal variates and $\lambda_1, \lambda_2, \ldots$ are eigenvalues

described in the above theorem.

We note that this result provides an asymptotic approximation to the null distribution of Q_n without any restriction on the complexity of the structure of M except for the symmetry of the discrepancy r. Furthermore, estimates of the eigenvalues $\lambda_1, \lambda_2, \ldots$ can be computed in a relatively straightforward fashion by calculating the eigenvalues of the kernel matrix H_n defined to have $(i,j)^{th}$ element $H_{n,ij} := u(X_i, X_j)$. While somewhat complicated, the exact form of $u(X_j, X_k)$ is provided in the Supplementary Materials. Once estimates $\lambda_{1,n}, \lambda_{2,n}, \ldots$ of the eigenvalues $\lambda_1, \lambda_2, \ldots$ have been obtained, the 95^{th} percentile of Z^* can be approximated by the sample 95^{th} percentile of the collection

$$\left\{ Z^*_{q,J} := 1 + \sigma_n^{-2} \sum_{j=1}^{J} \lambda_{j,n} \left(Z^2_{j,q} - 1 \right) : q = 1, 2, \ldots, Q \right\}$$

for both J and Q very large, $\pi_n := n_1/n$ and $\{Z_{j,q}\}_{j,q}$ a matrix of independent standard normal variates, thereby enabling the construction of an asymptotically valid test cutoff.

3 Simulation studies

To assess the performance of our proposed test, we considered two simulations scenarios. The first is the simple scalar ANOVA case with two groups. That is, we repeatedly ($B = 1000$ times) sample subjects $i = 1, \ldots, n$, with group indicator $D_i \sim Bern(0.5)$ and outcome $M_i \sim N(D_i, \sigma^2)$ for various values of n and σ^2. We then conduct traditional ANOVA as well as distance-based AVOVA using two distance functions: the squared Euclidean distance $d_1(M_j, M_k) = (M_j - M_k)^2/2$, and the absolute distance $d_2(M_j, M_k) = |M_j - M_k|/2$. The results are shown in Table 1. The power of the distance-based ANOVAs were slightly lower than the standard ANOVA, with the Euclidean distance-based ANOVA showing very similar power to the classical test. The absolute distance-based test showed the lowest power, with a loss of up to 15%. All type 1 error rates were around the nominal 5% level, and are shown in a table in Table A2.

We next considered a simple graph-outcome case, with five common nodes labeled A through E in each subject and variation in the presence or absence of edges between the nodes as illustrated in Figure 2. We fixed $\tau_1 = 5\%$ and simulated $B = 1000$ datasets for various values of n and τ_1, and used the number of edge disagreements as the distance func-

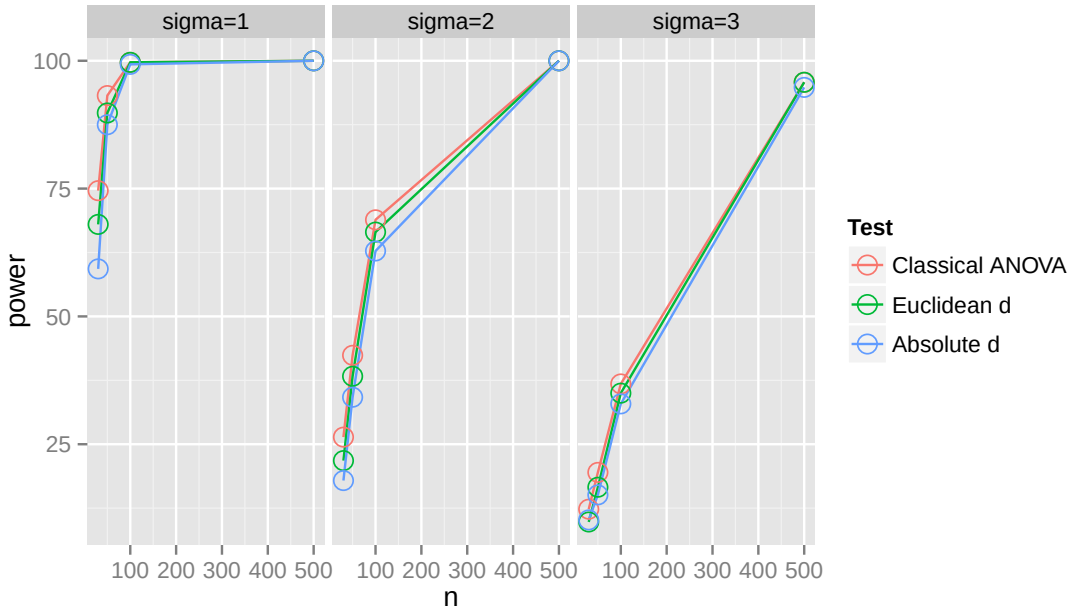

Figure 1: Power (%) estimated using squared Euclidean and absolute distance-based ANOVA in the scalar case, along with the classical ANOVA test. Note the estimated power of the standard ANOVA is highest as expected, and is quite similar to the Euclidean d_1-ANOVA. The absolute d_2-ANOVA shows slightly lower power.

$$\begin{bmatrix} 1 & 1 & 1-\tau_{D_i} & \tau_{D_i} & \tau_{D_i} \\ 1 & 1 & \tau_{D_i} & \tau_{D_i} & \tau_{D_i} \\ 1-\tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \\ \tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \\ \tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \end{bmatrix}$$

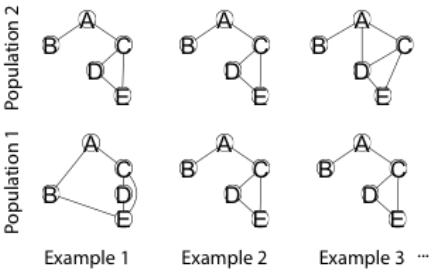

Figure 2: Graph-outcome simulation design. On the left, the adjacency matrix for the simulated graphs is shown, and on the right three example subjects are shown with $\tau_1 = 5\%$ and $\tau_2 = 10\%$.

tion between any two subjects. To compare this with the state-of-the-art PERMANOVA approach, we used the `adonis` function in the `vegan` package (Oksanen et al. 2015) in R (R Core Team 2015) and the results are shown in Table 3. In all cases, the type I error rate was estimated to be less than the nominal 5% rate but the tests were conservative (type I error rate around 1-2%) for the $n = 30$ and $n = 50$ cases (see Supplementary Materials). Note that the PERMANOVA approach yielded very low power in most cases, whereas the proposed distance-based test showed high power for larger sample sizes and effect sizes.

4 Results from the Autism study

In order to study the effect of autism on structural connectivity, we used diffusion tensor imaging (DTI) on 264 subjects aged 6 to 19 including 144 subjects with autism spectrum disorders (ASD) and 120 typically developing controls (TDC). Diagnoses were confirmed using expert consensus by two independent psychologists following the guidelines set by Collaborative Programs of Excellence in Autism (CPEA). 30-direction DTI were acquired and quality assured following de-noising and brain extraction, and a tensor model was fit to identify the direction of water diffusion across the brain. The brain was segmented into 301 regions using a co-registered T1-weighted image, and FSL probtrackx (Behrens et al. 2003) was used to estimated the degree of structural connectivity between each of the 301 regions accounting for the volume of each region. The observed data for each subject were thus symmetric 301 by 301 connectivity matrices, with (i, j)-th entry being a measure of the strength of connection between regions i and j. Figure (4) shows network representation of connectivity matrices, with nodes representing regions and edges being weighted by connection strength for two subjects for illustration.

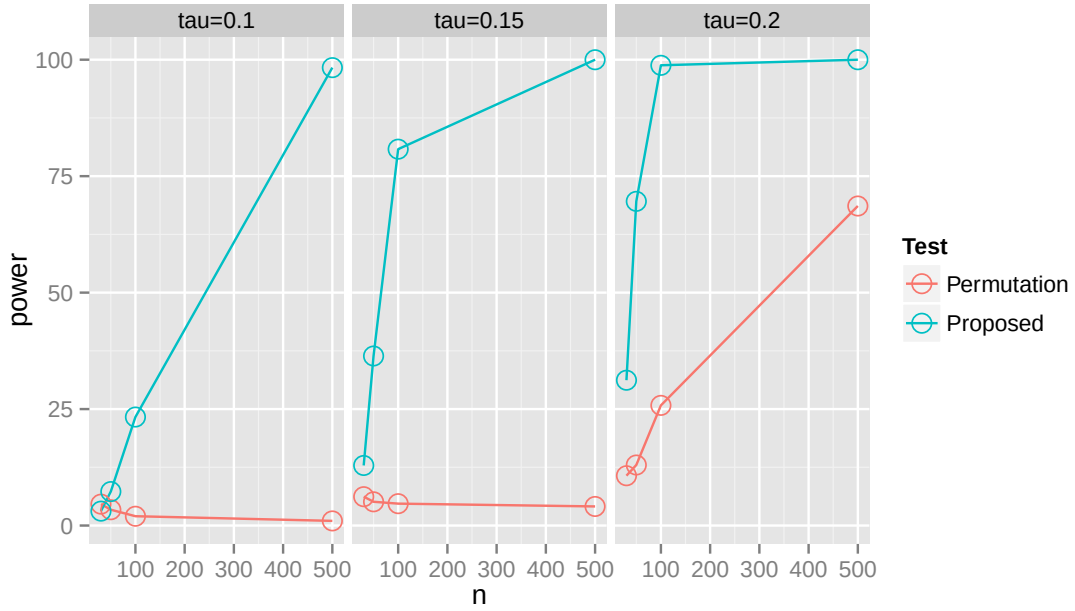

Figure 3: Estimated power (%) from simulations with graph outcomes distributed using the adjacency matrix shown in Figure 2. Note that the permutational ANOVA has no power for the $\tau_2 = 10\%$ and 15% cases and very low power for sample sizes smaller than $n = 500$ for the $\tau_2 = 20\%$ case. The proposed test, on the other hand, shows much higher power in all cases for all sample sizes.

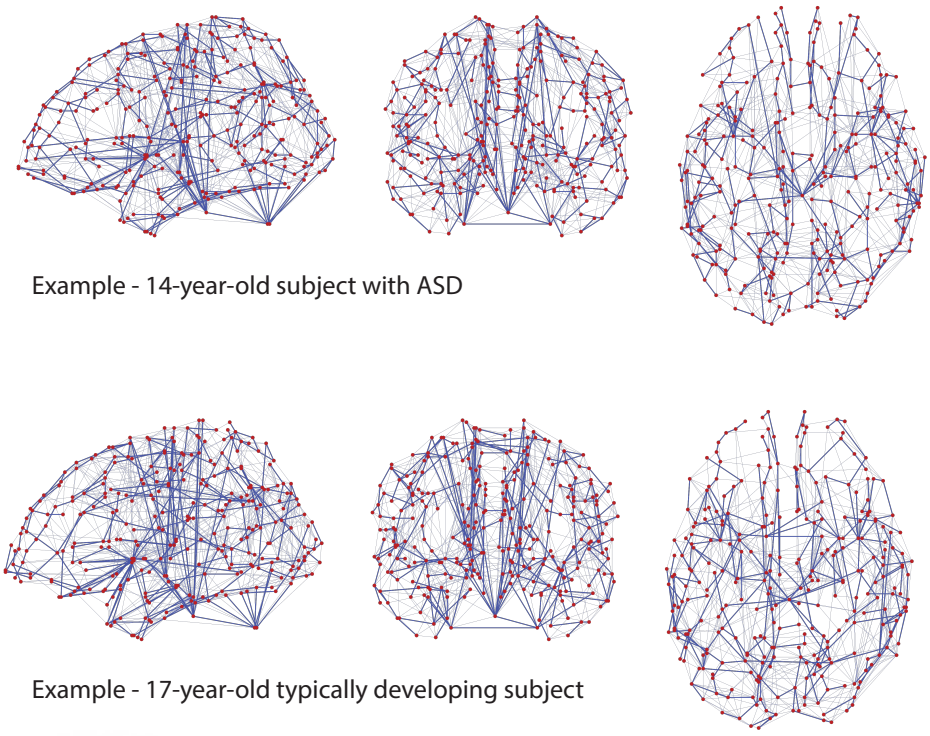

Figure 4: Observed data for two selected subjects, consisting of volume-normalized counts of streamline connections between each pair of regions. Regions are represented spatially in sagittal, coronal, and axial views as red dots, and blue lines are connections. Darker and wider blue lines indicate stronger connections between regions.

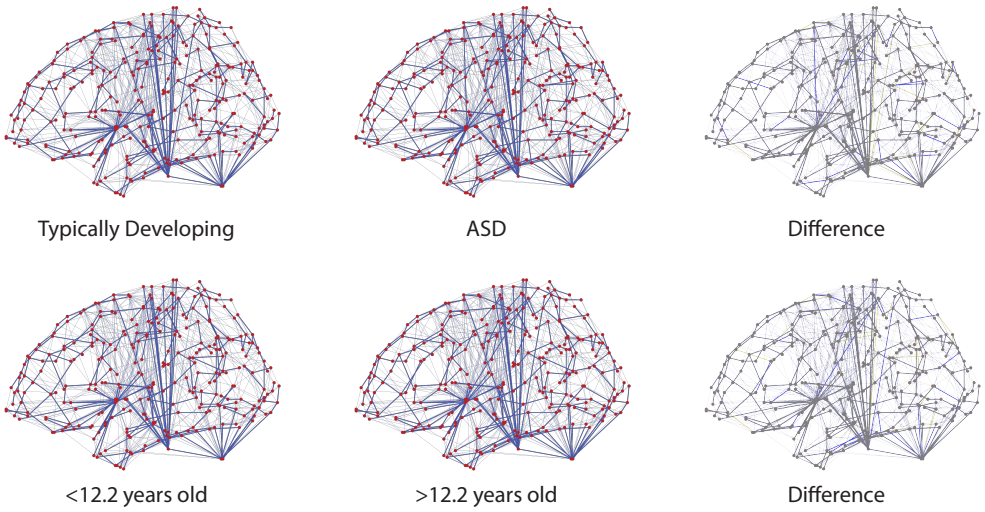

Figure 5: Mean networks (first and second columns) for ASD versus TDC subjects and absolute difference (top), and for subjects younger than the median age (12.2 years) compared to older (bottom). Regions are represented as red dots, and connections are shown in blue lines. Darker and thicker lines indicate stronger connections on average between regions. In the third column, the differences between the maps are shown, with gray lines indicating similar strength of connections in ASD (top) and older (bottom) subjects. Blue lines indicate stronger connections in ASD/older subjects, and yellow lines indicate weaker connections in these groups.

Our experiment included two comparisons which we accomplished using both the proposed test and the traditional permutational ANOVA: first, we tested for differences between the ASD and TDC groups. This difference is likely to be small, and thus our power to detect a global group difference is low. The mean connectivity matrices for the TDC and ASD groups are shown in Figure (4), and appear visually quite similar. Indeed, neither the proposed test ($p = 0.45$) nor PERMANOVA ($p = 0.36$) rejected the null of no difference between groups, likely due to the small effect size, relatively small sample size, and the global nature of the discrepancy measure selected. To assess the performance of our proposed test in the presence of stronger signal, we also examined differences associated with age-related development by dichotomizing age by its median across all subjects (12.2 years), and the average connectivity matrices in the younger and older groups are shown in the bottom row of Figure (4). Both the proposed test ($p = 0.004$) and the PERMANOVA ($p = 0.006$) indicated a significant difference in structural connectivity associated with age. This shows the promise for the proposed method for detecting structural changes in the connectome despite the high dimensionality of the connectomic representation.

5 Conclusion

In this paper, we propose a distance-based analysis of variance technique that allows for fast two-sample testing of connectomes represented by graphs and more complex structures using a subject-to-subject distance or discrepancy measure. We leverage the U-statistic form of the generalized variances studies to find the limiting behavior of the pseudo-F statistic. Our test shows improved power in simulations while maintaining nominal type one error rates. We expect that this test will also be useful in large genomic studies and other biomedical big data settings, in which permutational ANOVA testing are increasingly popular.

We demonstrate the utility of this methodology in a modern connectivity study. As ASD is an elusive disease in which only certain networks are affected, a global test using structural connectivity measures alone may not be most informative nor powerful. Future investigations of ASD using connectivity measures for certain functional systems, and targeted comparisons in pre-specified brain networks will likely be more fruitful for understanding disorder-related dysconnectivity. The proposed testing methodology may be useful for such scenarios, in combination with other approaches that include appropriate

dimension reduction. Additionally, to study global connectivity differences such as those attributable to age, the proposed test is promising for examining group differences.

SUPPLEMENTARY MATERIAL

ADDITIONAL SIMULATION RESULTS

Distance \ n	30	50	100	500
Euclidean r	2	5	4	6
Absolute r	3	4	3	4

Table A1: Type 1 error rates (%) for scalar simulation scenario. Rows indicate different hypothesis tests and columns corresponding to various sample sizes.

τ_2 \ n	30	50	100	500
0.1	1	1	3	4
0.15	1	2	3	4
0.2	2	2	4	5

Table A2: Type 1 error rates (%) for graph simulation scenario. Rows indicate different values for the parameter τ_2, and columns corresponding to various sample sizes.

Technical proofs

We begin by establishing a result on the asymptotic behavior of the product of the estimation error of two asymptotically linear estimators. Specifically, let θ_1 and θ_2 be two parameters, and denote by θ_{10} and θ_{20} their respective true values. Suppose that $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ are asymptotically linear estimators of θ_{10} and θ_{20}, respectively, with influence functions ϕ_{10} and ϕ_{20}.

Lemma. *The mapping* $h : (u, v) \mapsto \frac{1}{2}\{\phi_{10}(u)\phi_{20}(v) + \phi_{10}(v)\phi_{20}(u)\}$ *is a first-order degenerate kernel, and furthermore,*

$$n(\hat{\theta}_{1n} - \theta_{10})(\hat{\theta}_{2n} - \theta_{20}) = E_0\{\phi_{10}(X)\phi_{20}(X)\} + nU_n + o_P(1) \ ,$$

where U_n *is a* U*-statistic with kernel* h*. Additionally, it holds that*

$$\iiint f(x_1, x_2)h(x_1, x_3)dP_0(x_1)dP_0(x_2)dP_0(x_3) = 0$$

for any function $(x_1, x_2) \mapsto f(x_1, x_2)$ *such that* $\iint f^2(x_1, x_2)dP_0(x_1)dP_0(x_2) < \infty$*.*

Proof. In view of asymptotic linearity, we may write

$$\begin{aligned}(\hat{\theta}_{1n} - \theta_{10})(\hat{\theta}_{2n} - \theta_{20}) &= \left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{10}(X_i) + o_P(n^{-1/2})\right\}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{20}(X_i) + o_P(n^{-1/2})\right\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{10}(X_i)\phi_{20}(X_j) + \left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{10}(X_i)\right\}o_P(n^{-1/2}) \\ &\quad + \left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{20}(X_i)\right\}o_P(n^{-1/2}) + o_P(n^{-1})\end{aligned}$$

and so, using the fact that both $\frac{1}{n}\sum_{i=1}^{n}\phi_{10}(X_i)$ and $\frac{1}{n}\sum_{i=1}^{n}\phi_{20}(X_i)$ are $O_P(n^{-1/2})$ by the Central Limit Theorem,

$$n(\hat{\theta}_{1n} - \theta_{10})(\hat{\theta}_{2n} - \theta_{20}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{10}(X_i)\phi_{20}(X_j) + o_P(1) \ .$$

Furthermore, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{10}(X_i)\phi_{20}(X_j)$$

$$
\begin{aligned}
&= n\left(\frac{n-1}{n}\right)\binom{n}{2}^{-1}\sum_{i<j}h(X_i,X_j)+\frac{1}{n}\sum_{i=1}^{n}\phi_{10}(X_i)\phi_{20}(X_i)+o_P(1)\\
&= nU_n+E_0\left\{\phi_{10}(X)\phi_{20}(X)\right\}+o_P(1)\ .
\end{aligned}
$$

Since $E_0\{\phi_{10}(X)\} = E_0\{\phi_{20}(X)\} = 0$, it follows that $E_0\{h(X_i, X_j)\} = 0$ for $i \neq j$ and furthermore that $E_0\{h(X_1, X_2)h(X_1, X_3)\} = 0$. Finally, it is straightforward to verify that $E_0\{f(X_1, X_2)h(X_1, X_3)\} = 0$ for an arbitrary function $(x_1, x_2) \mapsto f(x_1, x_2)$.

□

We now establish the validity of our main theorem and its corollary.

Proof of Theorem. The weak consistency of SSE_n/n to σ_0^2 is a consequence of the weak law of large numbers for U-statistics. Using the fact that $SST_n = SS_n - SSE_n$, we observe that we may represent SST_n as

$$
\begin{aligned}
&(n-1)\binom{n}{2}^{-1}\sum_{j<k}r(M_j,M_k)-\sum_{r=0}^{1}(n_r-1)\binom{n_r}{2}^{-1}\sum_{j<k}r(M_j,M_k)I(D_j=D_k=r)\\
&= \frac{2}{n}\sum_{j<k}r(M_j,M_k)\left\{1-\frac{I(D_j=D_k=0)}{1-\pi_0}-\frac{I(D_j=D_k=1)}{\pi_0}\right\}\\
&\quad -\frac{2}{n}\sum_{j<k}r(M_j,M_k)\left[\left(\frac{n}{n_0}-\frac{1}{1-\pi_0}\right)I(D_j=D_k=0)+\left(\frac{n}{n_1}-\frac{1}{\pi_0}\right)I(D_j=D_k=1)\right]\\
&= A_{1n}-A_{2n}-A_{3n}-A_{4n}+o_P(1)\ ,
\end{aligned}
$$

where we have defined the summands

$$
\begin{aligned}
A_{1n} &:= n\cdot\binom{n}{2}^{-1}\sum_{j<k}r(M_j,M_k)\left\{1-\frac{I(D_j=D_k=0)}{1-\pi_0}-\frac{I(D_j=D_k=1)}{\pi_0}\right\}\\
A_{2n} &:= n\cdot\left(\frac{n}{n_0}-\frac{1}{1-\pi_0}\right)\binom{n}{2}^{-1}\sum_{j<k}\left\{r(M_j,M_k)I(D_j=D_k=0)-(1-\pi_0)^2\sigma_0^2\right\}\\
A_{3n} &:= n\cdot\left(\frac{n}{n_1}-\frac{1}{\pi_0}\right)\binom{n}{2}^{-1}\sum_{j<k}\left\{r(M_j,M_k)I(D_j=D_k=1)-\pi_0^2\sigma_0^2\right\}\\
A_{4n} &:= n\cdot\left(\frac{n}{n_0}-\frac{1}{1-\pi_0}\right)(1-\pi_0)^2\sigma_0^2+n\cdot\left(\frac{n}{n_1}-\frac{1}{\pi_0}\right)\pi_0^2\sigma_0^2\ .
\end{aligned}
$$

We now analyze each of the above four terms under the null hypothesis. The term A_{1n} is

an n-scaled U-statistic with first-order degenerate kernel h_1 given by

$$(x_1, x_2) \mapsto r(m_1, m_2) \left\{ 1 - \frac{I(d_1 = d_2 = 0)}{1 - \pi_0} - \frac{I(d_1 = d_2 = 1)}{\pi_0} \right\},$$

where $x_1 := (m_1, d_1)$ and $x_2 := (m_2, d_2)$ represent two arbitrary realizations of X. To study the term A_{2n}, we first note that

$$\binom{n}{2}^{-1} \sum_{j<k} \left\{ r(M_j, M_k) I(D_j = D_k = 0) - (1 - \pi_0)^2 \sigma_0^2 \right\}$$

is a non-degenerate U-statistic with kernel $(x_1, x_1) \mapsto r(m_1, m_2)\{I(d_1 = d_2 = 0) - (1 - \pi_0)^2 \sigma_0^2\}$, and by the Hájek projection, it is asymptotically linear with influence function

$$\begin{aligned} x \;\mapsto\; & 2 \left[E_0 \left\{ r(M, m) I(D = d = 0) \right\} - (1 - \pi_0)^2 \sigma_0^2 \right] \\ & = \; 2(1 - \pi_0) \left[I(d = 0) E_0 \left\{ r(M, m) \right\} - (1 - \pi_0) \sigma_0^2 \right] \;=:\; \phi_{21}(x) \;, \end{aligned}$$

where $E_0 \{r(M, m)\}$ is the average distance from m to a random draw from the distribution of M. By the delta method, we have that

$$\frac{n}{n_0} - \frac{1}{1 - \pi_0} = (1 - \pi_0)^{-2} \left(\frac{n_1}{n} - \pi_0 \right) + o_P(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} \phi_{22}(X_i) + o_P(n^{-1/2})$$

with $\phi_{22}(x) := \{I(d = 1) - \pi_0\}/(1 - \pi_0)^2$. By the Lemma, we can then write $A_{2n} = B_2 + nU_{2n} + o_P(1)$, where $B_2 := E_0 \{\phi_{21}(X)\phi_{22}(X)\} = -2\pi_0 \sigma_0^2$ and U_{2n} is a first-order degenerate U-statistic with kernel $h_2 : (x_1, x_2) \mapsto \{\phi_{21}(x_1)\phi_{22}(x_2) + \phi_{21}(x_2)\phi_{22}(x_1)\}/2$. We can proceed similarly for the term A_{3n}. Specifically, setting

$$\phi_{31}(x) := 2\pi_0 \left[I(d = 1) E_0 \left\{ r(M, m) \right\} - \pi_0 \sigma_0^2 \right]$$

and $\phi_{32}(x) := -\pi_0^{-2} \{I(d = 1) - \pi_0\}$, we find that $A_{3n} = B_3 + nU_{3n} + o_P(1)$ with $B_3 := E_0 \{\phi_{31}(X)\phi_{32}(X)\} = -2(1 - \pi_0)\sigma_0^2$ and U_{3n} is a first-order degenerate U-statistic with kernel $h_3 : (x_1, x_2) \mapsto \{\phi_{31}(x_1)\phi_{32}(x_2) + \phi_{31}(x_2)\phi_{32}(x_1)\}/2$. Finally, we note that

$$A_{4n} \;=\; \sigma_0^2 \left(\frac{n^2}{n_0 n_1} \right) \left\{ \sqrt{n} \left(\frac{n_1}{n} - \pi_0 \right) \right\}^2 \;=\; \frac{\sigma_0^2}{\pi_0(1 - \pi_0)} \left\{ \sqrt{n} \left(\frac{n_1}{n} - \pi_0 \right) \right\}^2 + o_P(1)$$

and therefore, by the Lemma, we readily find that $A_{4n} = B_4 + nU_{4n} + o_P(1)$, where

$B_4 := [\sigma_0^2/\{\pi_0(1-\pi_0)\}]E_0\left\{I(D=1)-\pi_0\right\}^2 = \sigma_0^2$ and U_{4n} is a first-order degenerate U-statistic with kernel $h_4 : (x_1, x_2) \mapsto [\sigma_0^2/\{\pi_0(1-\pi_0)\}]\{I(d_1=1)-\pi_0\}\{I(d_2=1)-\pi_0\}$. In view of the above facts, we obtain that

$$SST_n = \sigma_0^2 + n \cdot \binom{n}{2}^{-1} \sum_{j<k} u(X_j, X_k) + o_P(1) \ ,$$

where u is the first-order degenerate kernel $(x_1, x_2) \mapsto h_1(x_1, x_2) + h_2(x_1, x_2) + h_3(x_1, x_2) + h_4(x_1, x_2)$. It is easy to verify that $Cov\{f(X_1, X_2), h_j(X_1, X_3)\} = 0$ for any $f : \mathbb{R}^2 \to \mathbb{R}$, and so, in particular, we find that $Cov\{u(X_1, X_2), u(X_1, X_3)\} = 0$. It follows then that SST_n tends in distribution to $\sigma_0^2 + \sum_{j=1}^{\infty} \lambda_j(Z_j^2 - 1)$ as $n \to \infty$, where $Z_1, Z_2, \ldots$ are independent standard normal random variables and $\lambda_1, \lambda_2, \ldots$ are eigenvalues of the operator $g \mapsto \Gamma(g) : m \mapsto \int u(m, m_2)g(m_2)dP(m_2)$ (Lee 1990).

Proof of Corollary. Follows from the theorem by an application of Slutsky's theorem.

References

Anderson, T. W. (1963), 'Asymptotic theory for principal component analysis', *Annals of Mathematical Statistics* pp. 122–148.

Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., Matthews, P., Brady, J. & Smith, S. (2003), 'Characterization and propagation of uncertainty in diffusion-weighted mr imaging', *Magnetic resonance in medicine* **50**(5), 1077–1088.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. (2013), 'Sequence kernel association tests for the combined effect of rare and common variants', *The American Journal of Human Genetics* **92**(6), 841–853.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D. & Epstein, M. P. (2008), 'A powerful and flexible multilocus association test for quantitative traits', *The American Journal of Human Genetics* **82**(2), 386–397.

Lee, J. (1990), 'U-statistics: Theory and practice'.

Lee, S., Wu, M. C. & Lin, X. (2012), 'Optimal tests for rare variant effects in sequencing association studies', *Biostatistics* **13**(4), 762–775.

McArdle, B. H. & Anderson, M. J. (2001), 'Fitting multivariate models to community data: a comment on distance-based redundancy analysis', *Ecology* **82**(1), 290–297.

Minas, C. & Montana, G. (2014), 'Distance-based analysis of variance: Approximate inference', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **7**(6), 450–470.

Minas, C., Waddell, S. J. & Montana, G. (2011), 'Distance-based differential analysis of gene curves', *Bioinformatics* **27**(22), 3135–3141.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. & Wagner, H. (2015), *vegan: Community Ecology Package*. R package version 2.3-0.
URL: *http://CRAN.R-project.org/package=vegan*

Pan, W. (2009), 'Asymptotic tests of association with multiple snps in linkage disequilibrium', *Genetic epidemiology* **33**(6), 497.

Pan, W., Kim, J., Zhang, Y., Shen, X. & Wei, P. (2014), 'A powerful and adaptive association test for rare variants', *Genetics* **197**(4), 1081–1095.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Reiss, P. T., Stevens, M. H. H., Shehzad, Z., Petkova, E. & Milham, M. P. (2010), 'On distance-based permutation tests for between-group comparisons', *Biometrics* **66**(2), 636–643.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K. et al. (2013), 'Equivalence of distance-based and rkhs-based statistics in hypothesis testing', *The Annals of Statistics* **41**(5), 2263–2291.

Székely, G. J., Rizzo, M. L., Bakirov, N. K. et al. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Székely, G. J., Rizzo, M. L. et al. (2009), 'Brownian distance covariance', *The annals of applied statistics* **3**(4), 1236–1265.

Wang, T. & Elston, R. C. (2007), 'Improved power by use of a weighted score test for linkage disequilibrium mapping', *The american journal of human genetics* **80**(2), 353–360.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011), 'Rare-variant association testing for sequencing data with the sequence kernel association test', *The American Journal of Human Genetics* **89**(1), 82–93.

Zhang, D. & Lin, X. (2003), 'Hypothesis testing in semiparametric additive mixed models', *Biostatistics* **4**(1), 57–74.

Zhao, N., Chen, J., Carroll, I. M., Ringel-Kulka, T., Epstein, M. P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. & Wu, M. C. (2015), 'Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test', *The American Journal of Human Genetics* **96**(5), 797–807.