Two-Sample Testing for Connectomes

Russell Shinohara, Haochang Shou, Marco Carone,
Drew Parker, Robert Schultz, and Ragini Verma

Department of Biostatistics and Epidemiology
Perelman School of Medicine
University of Pennsylvania

February 4, 2016

Collaborators

Disclaimer

The following is work in progress.

Feedback is very much appreciated.

This work was supported by R01NS085211 (Shinohara) and R21MH098010 (Verma). This work represents the opinions of the researchers and not necessarily that of the granting organizations.

Introduction

Connectomics is the production and study of connectomes: comprehensive maps of connections within an organism's nervous system, typically its brain or eye. (Wikipedia)

Connections are assess in two ways: **structurally** and **functionally**.

Structural connectivity asks

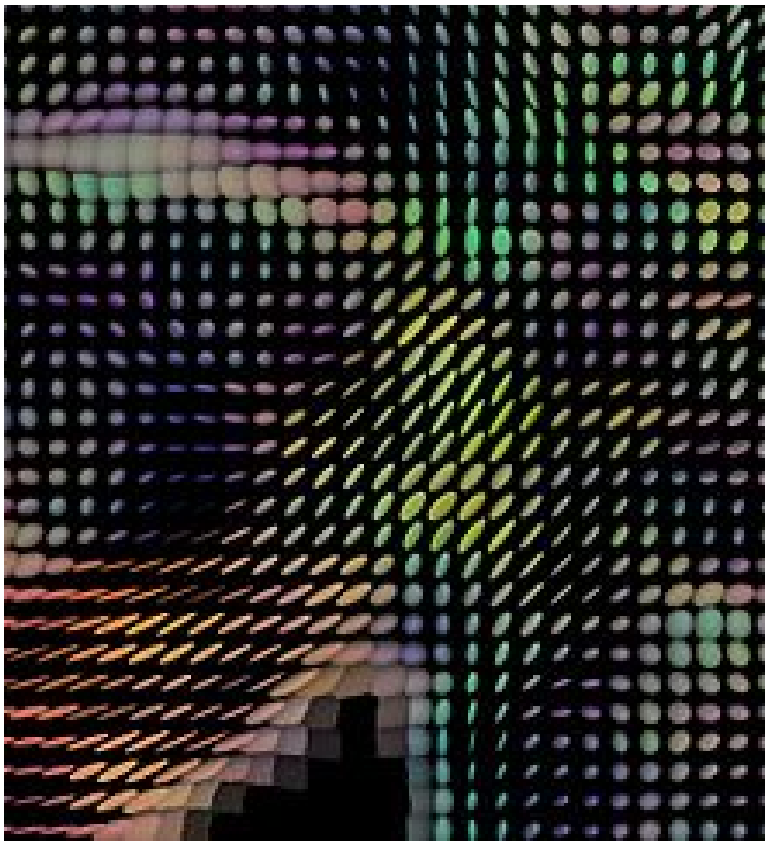
"which regions of the brain have physical connections between them?"

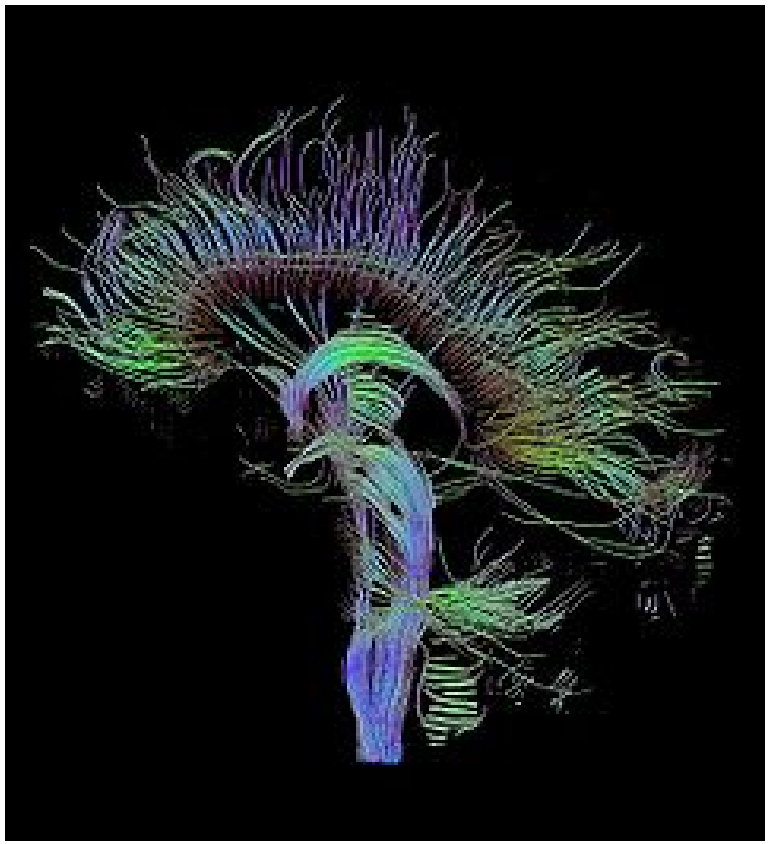
Functional connectivity asks

"which regions of the brain tend to function in related ways?"

Structural Connectivity

Structural connectivity is assessed using diffusion tensor imaging (DTI). This MRI technique measures the direction of water flow in the brain; it allows us to **measure connections** and quantify their **strength**.

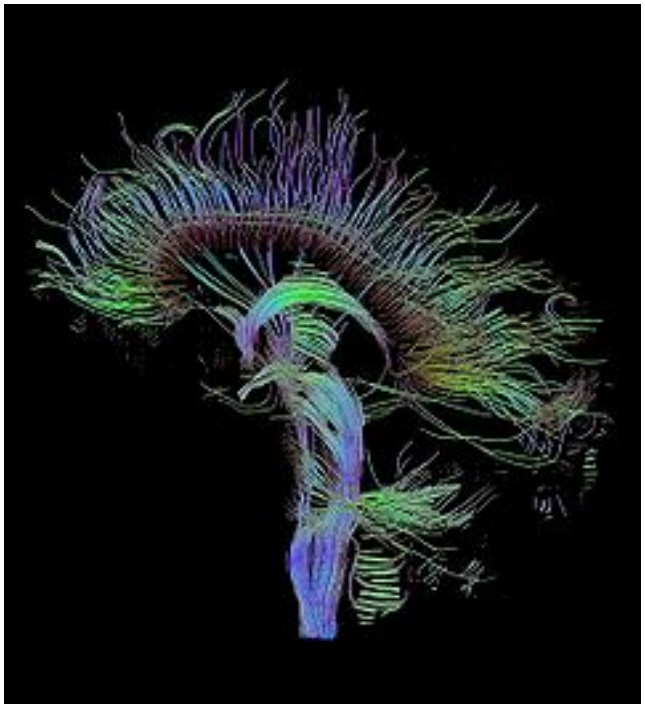




http://www.sci.utah.edu/ gk/DTI-data/

Structural Connectivity

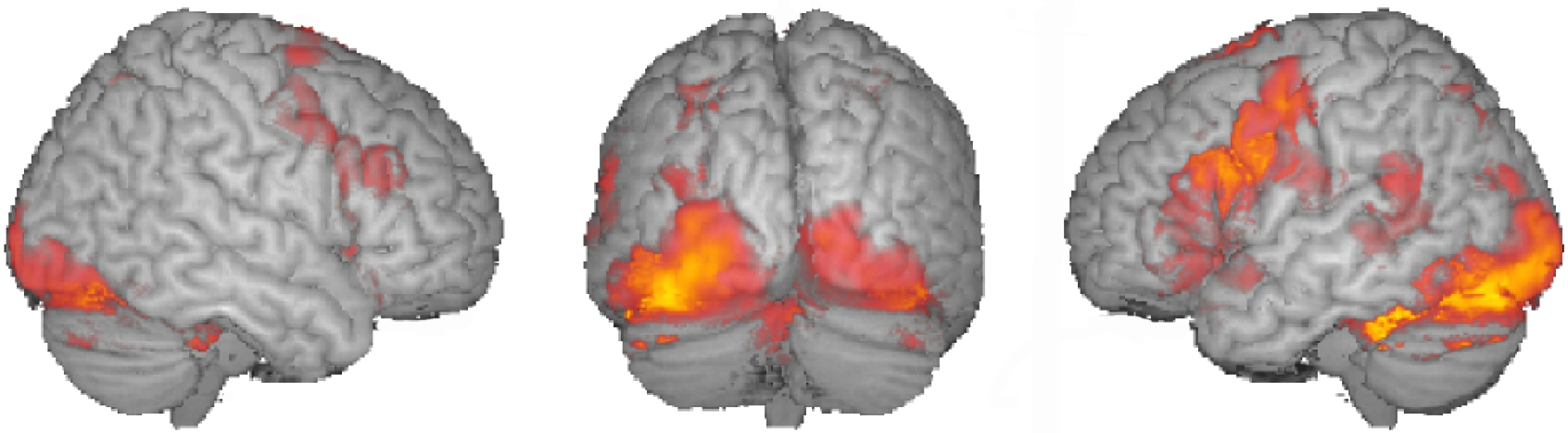
Structural connectivity can be represented (quantified) in many different ways, but one way involves asking: **how many streamlines (connections) are there between any two regions?**



More specifically, the outcome for each subject is a matrix of counts, where the i, j-th entry is the number of connections between locations i and j as defined commonly across subjects.

Functional Connectivity

Functional connectivity can be assessed in a variety of ways. The most common is **fMRI**, which measures blood oxygen level across the brain as a surrogate for brain activity. This can be conducted while a subject is at rest (resting state) or doing something (task). Magnetoencephalograpy (MEG) is an (expensive) alternative.



http://www.med.nyu.edu/thesenlab/research-0/research-functional-magnetic-resonance-imaging-fmri/

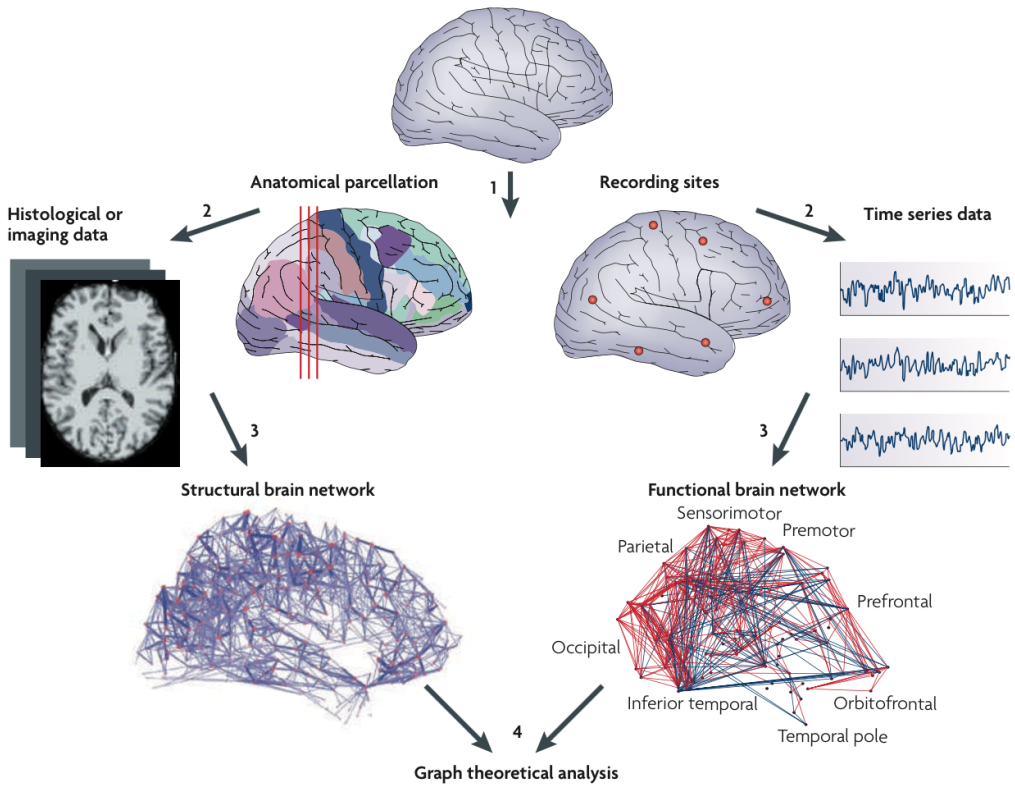
Functional Connectivity

Both MEG and fMRI provide **time series** observed at each location in the brain. In studies of functional connectivity, the goal is to measure similarity in the functional status between locations in the brain. A common approach for analysis is to ask: **how correlated are the time series in pairs of regions?**

In this case, the outcome for each subject is a matrix, where the i, j-th entry is the a measure of connection of the time series observed in locations i and j as defined commonly across subjects.

Graphs and Connectivity

Big (symmetric) matrices are difficult to interpret, so oftentimes investigators will threshold the matrices to produce graphs.

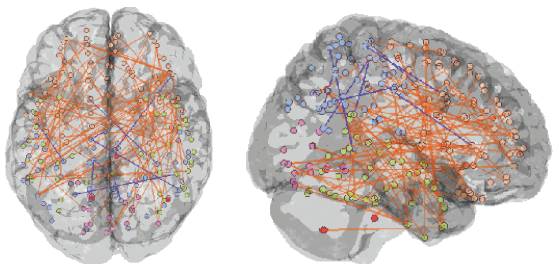


E. Bullmore, O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems Nat. Rev., Neurosci., 10 (2009), pp. 186198

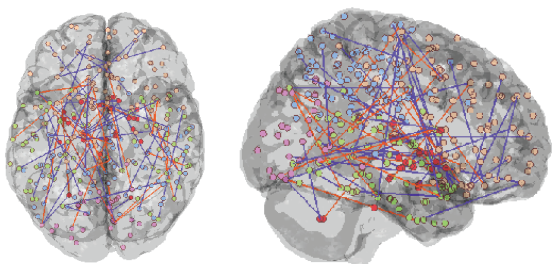
Motivating Example: Autism

Our interest is to ask: is the connectome different in diseased patients compared to healthy subjects? If so, where are they different?

One motivating example is in autism, in a study with 264 subjects aged 6-19 (matched) with autism spectrum disorders (ASD) and typically developing (TD) controls. These subjects have undergone MEG/DTI, and our goal is to ask about differences in the structural connectome of the brain in ASD.



Results of edge-wise comparisons from MEG



Results of edge-wise comparisons from DTI

Figure courtesy of Ragini Verma.

Notation

We consider the observed data to be $X_i = (M_i, D_i)$, where M_i is the DTI connectivity matrix, and $D_i \in \{0, 1\}$ is the indicator of ASD for $i = 1, \ldots, n$. We let F_0 and F_1 denote the distribution functions of the data M_i in the TD and ASD groups respectively, and we wish to test:

$$\mathcal{H}_0 : F_0 = F_1 \text{ versus}$$
$$\mathcal{H}_1 : F_0 \neq F_1$$

But how? This might be especially complex if we represent the connectomes by complex graphs, etc. We propose to do this using distance statistics.

What are distance statistics?

To assess this, we can remember that from the theory of U-statistics, the sample variance

$$\begin{aligned}\hat{\sigma}^2 &= \sum_i (x_i - \bar{x})^2/(n-1) \\ &= \sum_{(i,j)\in\Gamma} (x_i - x_j)^2/2|\Gamma|\end{aligned}$$

where $\Gamma = \{(i,j) : i \neq j\}$. Thus, variance may be written as a scaled version of the average squared Euclidean distance between any two observations. To generalize this, we may replace the squared Euclidean distance with a general measure of discrepancy or a metric, say d, and for computational efficiency we estimated the variance of a set of densities by $\sum_{(l,k)\in\Gamma^*} d(M_k, M_l)du/|\Gamma^*|$ where Γ^* is a randomly chosen sufficiently large subset of Γ (for this study we used $|\Gamma^*| = 2000$).

What are distance statistics?

How does that help us with testing differences in connectomes? Well, if we can define variance, we can do ANOVA.

We can ask: do the distances in connectomes between groups tend to be larger than the distances within groups?

Also, if we want to assess where the connectomes might differ, we can do ANOVA on submatrices/subgraphs of M.

Distance-based ANOVA

Consider the ANOVA test statistic[1]

$$T = \frac{SST/(2-1)}{SSE/(n-1)}$$

where

$$SSE = \sum_{r=1}^{2}(n_r - 1)\binom{n_r}{2}^{-1} \sum_{j<k:D_j=r,D_j=r} d(M_j, M_k),$$

$$SS = (n-1)\binom{n}{2}^{-1} \sum_{j<k} d(M_j, M_k), \text{ and}$$

$$SST = SS - SSE$$

[1] This formulation turns out to be equivalent to McArdle and Anderson (2001).

But what's the null distribution of T?

To estimate the null distribution of T, a simple option[2] is to **permute** the group labels D_i repeatedly to generate the distribution of T under the null hypothesis.

This is **computationally expensive**; to address this, Minas et al. (2014) recently proposed an approximation to the permutation null distribution.

However, the permutation-based null distribution may also result in **suboptimal statistical power**.

But wait- didn't we get here in the first place by invoking the U-statistic representation of sample variance? And don't **U-statistics have nice asymptotic behaviors**? Sort of.

[2] McArdle and Anderson (2001), Reiss et al. (2010), Minas et al. (2011)

But what's the null distribution of T?

We know that

$$\begin{aligned} SS &= (n-1)\binom{n}{2}^{-1} \sum_{j<k} d(M_j, M_k) \\ &= (n-1)U \end{aligned}$$

where U is a U-statistic with kernel d and $\sqrt{n}U \Rightarrow N(\sigma^2, 4\zeta_1)$ and $n \to \infty$, where

$$\begin{aligned} \sigma^2 &= E\{d(M_1, M_2)\} \text{ and} \\ \zeta_1 &= Cov\{d(M_1, M_2), d(M_1, M_3)\} \end{aligned}$$

And a similar argument applies to SSE. So this should be easy?

But what's the null distribution of T?

Not quite. Although we might be tempted to use the delta method after finding the asymptotic distribution of SSE and SS, there is a problem. In particular, SST is a measure of the additional variation between groups compared to that within groups.

Under the null, there is none.

So, we don't get $\sqrt{n}$-asymptotics.

But what's the null distribution of T?

However, we can write:

$$
\begin{aligned}
SST &= (n-1)\binom{n}{2}^{-1}\sum_{j<k} d(M_j, M_k) - \sum_{r=0}^{1}(n_r - 1)\binom{n_r}{2}^{-1}\sum_{j<k:D_j=r,D_j=r} d(M_j, M_k) \\
&= \frac{2}{n}\sum_{j<k} d(M_j, M_k)\left\{1 - \frac{n}{n_1}I(D_j = D_k = 0) - \frac{n}{n_2}I(D_j = D_k = 1)\right\} \\
&= \underbrace{\frac{2}{n}\sum_{j<k} d(M_j, M_k)\left\{1 - \eta^{-1}I(D_j = D_k = 0) - (1-\eta)^{-1}I(D_j = D_k = 1)\right\}}_{\text{(A)}} \\
&\quad + \underbrace{\frac{2}{n}\sum_{j<k} d(M_j, M_k)\left[\left\{\frac{n}{n_1} - \eta^{-1}\right\}I(D_j = D_k = 0) + \left\{\frac{n}{n_2} - (1-\eta)^{-1}\right\}I(D_j = D_k = 1)\right]}_{\text{(B)}}
\end{aligned}
$$

First, it is easy to see that $U = (n-1)^{-1}(\mathrm{A})$ is a U-statistic, with kernel $h_\eta(X_j, X_k)$. It turns our that U is a degenerate U-statistic and the degeneracy is of order 1.

But what's the null distribution of T?

However, we can write:

$$SST = \frac{2}{n}\sum_{j<k} d(M_j, M_k)\left\{1 - \eta^{-1}I(D_j = D_k = 0) - (1-\eta)^{-1}I(D_j = D_k = 1)\right\}$$
$$\underbrace{- \left\{\frac{n}{n_1} - \eta^{-1}\right\}\frac{2}{n-1}\sum_{j<k}\left\{d(M_j, M_k)I(D_j = D_k = 0) - \eta^2\sigma^2\right\}}_{(B1)}$$
$$\underbrace{- \left\{\frac{n}{n_2} - (1-\eta)^{-1}\right\}\frac{2}{n-1}\sum_{j<k}\left\{d(M_j, M_k)I(D_j = D_k = 1) - (1-\eta)^2\sigma^2\right\}}_{(B2)}$$
$$\underbrace{- \left\{\frac{n}{n_1} - \eta^{-1}\right\}n\eta^2\sigma^2 - \left\{\frac{n}{n_2} - (1-\eta)^{-1}\right\}n(1-\eta)^2\sigma^2}_{(B3)} + o_P(1)$$

And we recognize each of (B1)-(B3) as products of mean-zero asymptotically linear estimators. What to do with them?

A very useful lemma.

Lemma

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two asymptotically linear estimators of θ_1 and θ_2, with influence functions Ψ_1 and Ψ_2 respectively. Then, $n(\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2)$ is asymptotically equivalent to $E(\Psi_1\Psi_2) + nU$ where U is a U-statistic with first-order degenerate kernel:

$$h(X_i, X_j) = \{\Psi_1(X_i)\Psi_2(X_j) + \Psi_1(X_j)\Psi_2(X_i)\}/2.$$

Furthermore, for $f(x_1, x_2)$ with $Ef^2(X_1, X_2) < \infty$ we have $Ef(X_1, X_2)h(X_1, X_3) = 0$.

Thus, we can re-express our estimator as being asymptotically equivalent to a U-statistic!

But what's the null distribution of T?

In summary,

$$SST = \sigma^2 + n\binom{n}{2}^{-1} \sum_{j<k} h^*(X_j, X_k) + o_P(1),$$

where

$$\begin{aligned} h^*(X_j, X_k) = d(M_j, M_k)\{1 - (1 - D_j)(1 - D_k)/\eta - D_j D_k/(1 - \eta)\} \\ - \left(1 - \frac{1 - D_j}{\eta}\right)\{(1 - D_k)E_M d(M_k, M) - \sigma^2\eta\} \\ - \left(1 - \frac{1 - D_k}{\eta}\right)\{(1 - D_j)E_M d(M_j, M) - \sigma^2\eta\} \\ - \left(1 - \frac{D_j}{1 - \eta}\right)\{D_k E_M d(M_k, M) - \sigma^2(1 - \eta)\} \\ - \left(1 - \frac{D_k}{1 - \eta}\right)\{D_j E_M d(M_j, M) - \sigma^2(1 - \eta)\} \\ - (1 - D_k - \eta)(1 - D_j - \eta)\sigma^2\eta^{-1}(1 - \eta)^{-1}. \end{aligned}$$

Which can be seen easily to also be first order degenerate.

But what's the null distribution of T?

Now[3], we write the spectral representation $h^*(x_1, x_2) = \sum_{l=1}^{\infty} \lambda_l \phi_l(x_1)\phi_l(x_2)$ (from Hilbert-Schmidt Theory), where λ_l are the eigenvalues and ϕ_l are the eigenfunctions of the transformation $\Psi : g(x) \longmapsto Eh^*(x, X_j)g(X_j)$. Then, under regularity conditions,

$$
\begin{aligned}
nU &= \frac{2}{n-1} \sum_{j<k} h^*(X_j, X_k) = \frac{2}{n-1} \sum_{j<k} \sum_{l} \lambda_l \phi_l(X_j)\phi_l(X_k) \\
&= \sum_{l} \lambda_l \frac{1}{n-1} \left[\left\{ \sum_{j} \phi_l(X_j) \right\}^2 - \sum_{j} \phi_l(X_j)^2 \right] \\
&= \sum_{l} \lambda_l \left[\left\{ \frac{1}{\sqrt{n-1}} \sum_{j} \phi_l(X_j) \right\}^2 - \frac{1}{n-1} \sum_{j} \phi_l(X_j)^2 \right] \Rightarrow \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)
\end{aligned}
$$

Next, where $Z_j \sim N(0, 1)$ by the central limit theorem which are independent by the orthonormality of ϕ_l. Now, the eigenvalues of Ψ satisfy $Eh^*(x, X_j)\phi_l(X_j) = \lambda_l \phi_l(X_j)$, so to estimate λ_l we can use the empirical version of this expression, $H_n \phi_{l,n} = \lambda_{l,n} \phi_{l,n}$.

Thus, $SST \Rightarrow \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$ as $n \to \infty$.

[3] Lee, A.J. (1990). U-statistics: Theory and Practice

But what's the null distribution of T?

Theorem

Under the null hypothesis that $p(M_i \mid D_i) = p(M_i)$, SST is asymptotically equivalent to $\sigma^2 + nU^*$ where U^* is the U-statistic with a first-order degenerate kernel.
Thus,

$$SST \Rightarrow \sum_{j=1}^{\infty} \lambda_j (Z_j^{*2} - 1) + \sigma^2$$

where $\left(\sum D_i\right)/n \to \eta \in (0, 1)$, $\sigma^2 = E[d(M_j, M_k)]$, $Z_1^*, Z_2^*, \ldots$ are normally distributed random variables with $Z_j^* \perp Z_{j'}^*$ for $j \neq j'$ and $\lambda_1, \lambda_2, \ldots$ are eigenvalues of an integral equation that can be approximated using a singular value decomposition.

End result.

Corollary

Under the null hypothesis that $p(M_i \mid D_i) = p(M_i)$*, we have*

$$T \Rightarrow \sigma^{-2} \left\{ \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1) + 1 \right\}$$

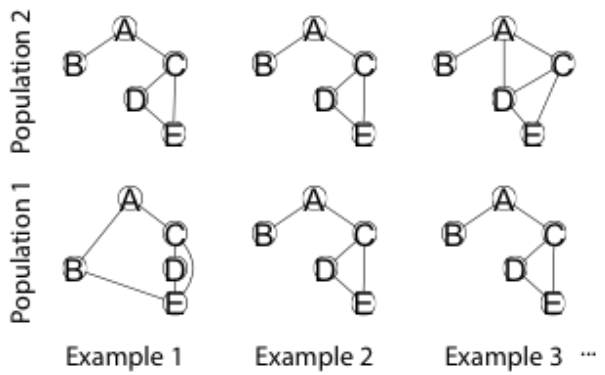
Simulations

To assess the performance of our test, for $i = 1, \ldots, n$, we let $D_i \sim Bern(0.5)$ and consider two outcome distributions:

Scalar Case $M_i \sim N(D_i, \sigma^2)$

Graph Case M_i is sampled using the adjacency matrix:

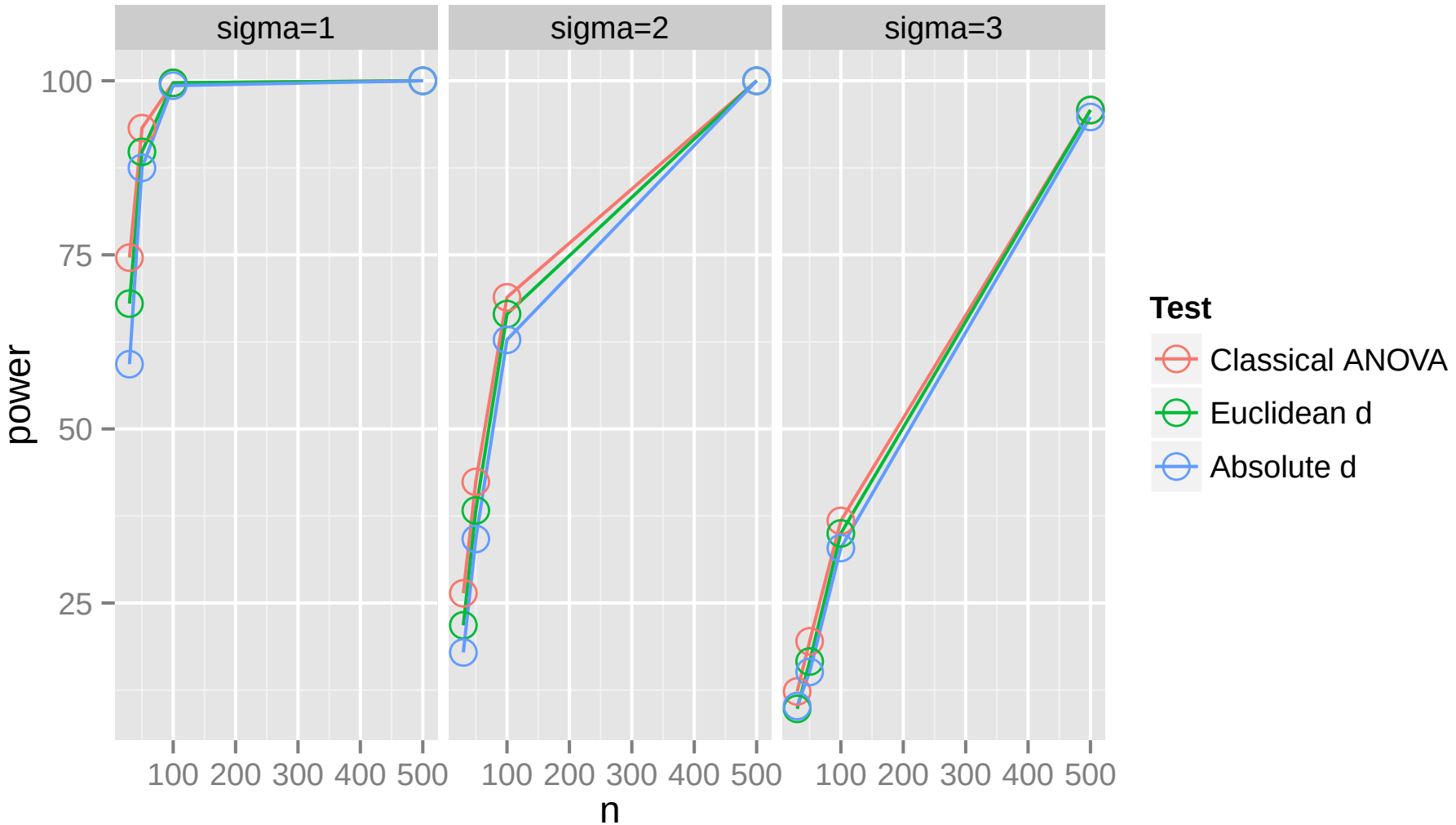
$$\begin{bmatrix} 1 & 1 & 1-\tau_{D_i} & \tau_{D_i} & \tau_{D_i} \\ 1 & 1 & \tau_{D_i} & \tau_{D_i} & \tau_{D_i} \\ 1-\tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \\ \tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \\ \tau_{D_i} & \tau_{D_i} & 1 & 1 & 1 \end{bmatrix}$$



We simulated $B = 1000$ datasets from the above distributions for various parameter values. All simulations showed type I error below the 5% rate.

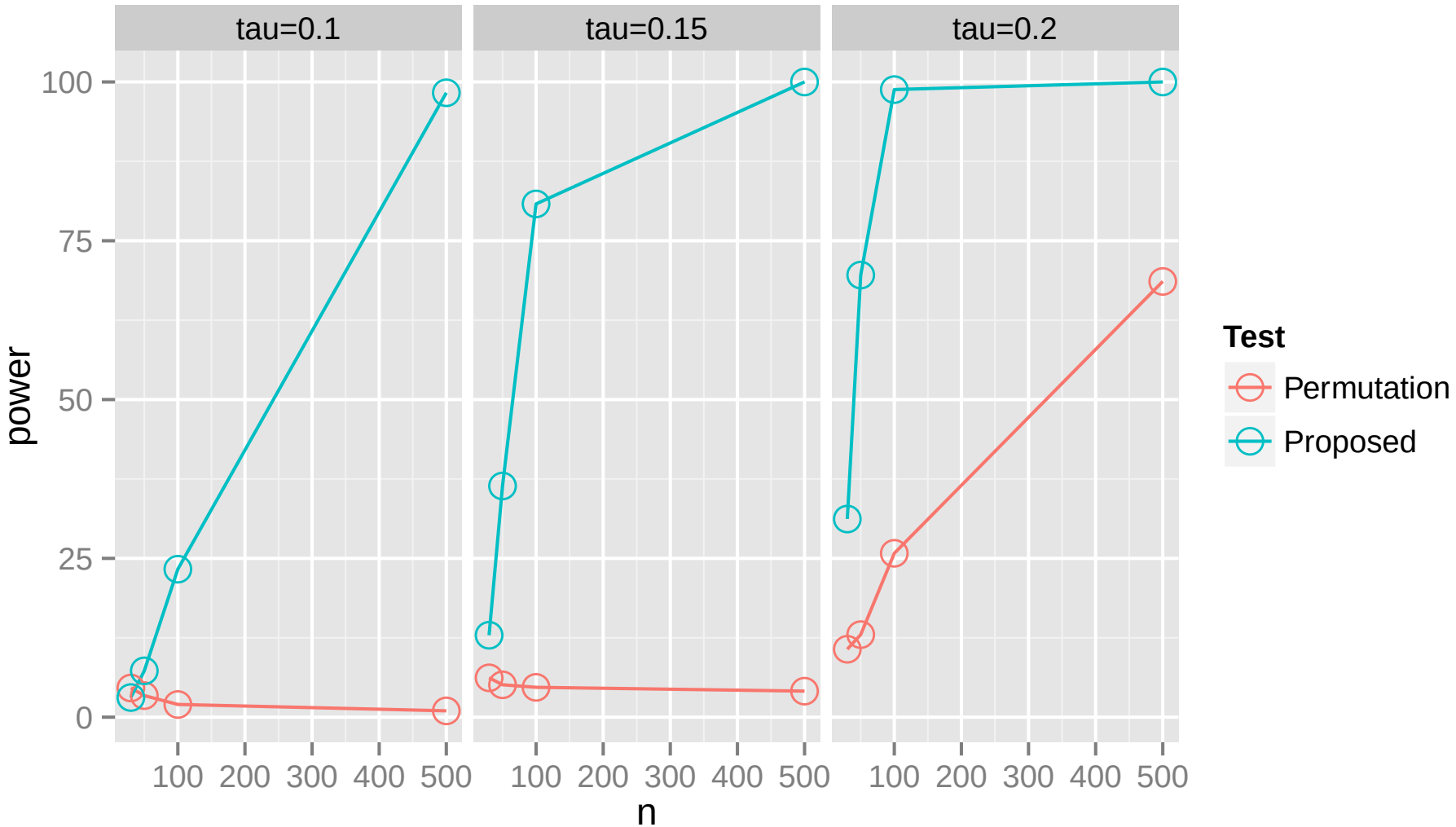
Simulations - Scalar Case

For the scalar simulations, we used two distance functions: the squared Euclidean distance $d_1(M_j, M_k) = (M_j - M_k)^2/2$ and the absolute distance $d_2(M_j, M_k) = |M_j - M_k|/2$.



Simulations - Graph Case

For the graph simulations, we used the number of edge disagreements between subjects as the discrepancy measure and fixed $\tau_1 = 5\%$.



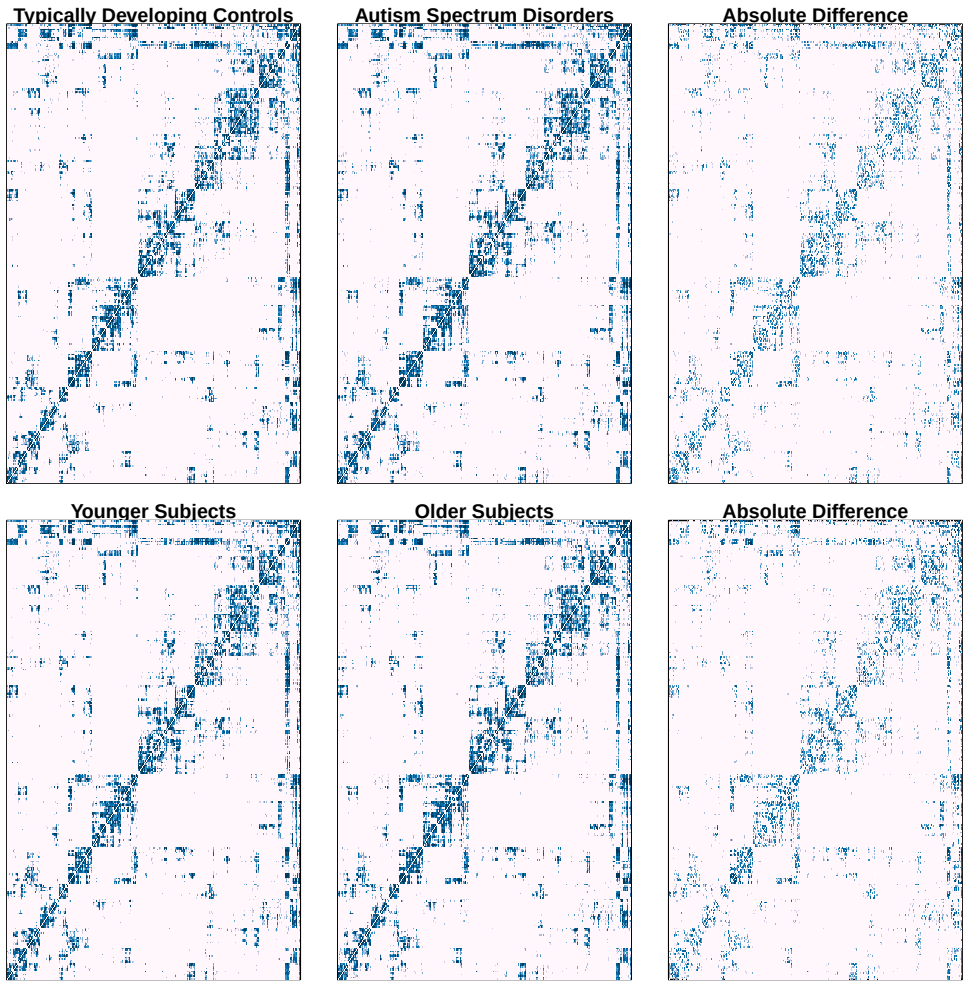
DTI Connectivity in the CURE Study

Our motivating study of autism consisted of age-matched 144 subjects with autism and 120 who were typically developing.

Diagnoses were confirmed using expert consensus by two independent psychologists following the guidelines set by Collaborative Programs of Excellence in Autism (CPEA). Briefly, 30-direction DTI were 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 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.

DTI Connectivity in the CURE Study



DTI Connectivity in the CURE Study

For discrepancies in DTI connectivity we used the squared Frobenius norm of the differences in the streamline count matrices and applied our test.

No diagnosis-related differences were observed in our small sample for connectivity in structure ($p=0.45$).

We did, however see **differences between age groups** using our testing framework ($p<0.01$).

We are currently examining how more sensitive distance measures and targeted subnetwork analyses can refine our results, and help us to understand changes in the connectivity of the brain in autism.

Thanks!

Thank you very much for your attention!

Please check out our website at http://www.med.upenn.edu/pennsive/ if you are interested in learning more about our work.