Exact Topological Inference of the Resting-State Brain Network in Twins

Moo K. Chung
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
www.stat.wisc.edu/~mchung
Abstract

A cycle in a brain network is a subset of a connected component with redundant additional connections. If there are many cycles in a connected component, the connected component is more densely connected. While the number of connected components represents the integration of the brain network, the number of cycles represents how strong the integration is. However, it is unclear how to perform statistical inference on the number of cycles in the brain network. In this lecture, we present a new Exact Topological Inference framework for determining the statistical significance of the number of cycles through the Kolmogorov-Smirnov (KS) distance, which was recently introduced to measure the similarity between networks across different filtration values using the zeroth Betti number. We show how to extend the method to the first Betti number. Using a twin imaging study, which provides biological ground truth, the methods are applied in determining if cycles are heritable network features in the resting-state functional brain networks of 217 twins. This talk is based on a paper of the same title: doi.org/10.1162/netn_a_00091. The MATLAB codes as well as the connectivity matrices used in the paper are freely available at www.stat.wisc.edu/~mchung/TDA.
Codes, data & lecture slides given in

www.stat.wisc.edu/~mchung/TDA

More codes & published brain imaging data given in

https://www.stat.wisc.edu/~mchung/software.html
Acknowledgement

Yixian Wang, Shih-Gu Huang, Andrey Grisenko, Ross Luo, Nagesh Adluru, Andrew Alexander, Richard Davidson, Hill Goldsmith

University of Wisconsin-Madison, USA

Yuan Wang \ University of South Carolina
Hyekyung Lee \ Seoul National University
Hernando Ombao \ KAUST

NIH grants: R01 EB022856, R01 MH101504, P30 HD003352, U54 HD09025
Full day course
Topological and Object Oriented Data Analysis

International Biometric Conference (IBC2020)
COEX Seoul, Korea

Sunday July 5, 2020

Steve Marron (UNC)
Yuan Wang (USC)
Moo K. Chung (UW-Madison)

http://www.tda-brain.com/teaching/ibc2020
Motivation of this talk

There is a still huge gap between TDA theory to applications.

Theory  Must integrate multiple images: statistical problem  Neuroimaging application
Previous works & Preliminary
3T MRI research scanner in Madison

Structural MRI

Functional MRI

Error using double
Requested
1083154800x1
(8.1GB) array exceeds
maximum array size.
Persistence Diagrams of Cortical Surface Data

Moo K. Chung\textsuperscript{1,2}, Peter Bubenik\textsuperscript{3}, and Peter T. Kim\textsuperscript{4}

Chung et al., 2009 Information Processing in Medical Imaging (IPMI)
Kernel density Estimation (uniform kernel)
Permutation test

\[ x = (x_1, x_2, \cdots, x_m) \]
\[ y = (y_1, y_2, \cdots, y_n) \]
\[ (x, y) = (x_1, \cdots, x_m, y_1, \cdots, y_n) \]
\[ \pi(x, y) \in S_{m+n} \]

Permutation group of order \( m+n \)

\[ p\text{-value} = \frac{1}{(m + n)!} \sum_{\tau \in S_{m+n}} I(f(\tau(x), \tau(y)) > f(x, y)) \]
Permutation test

Observation: \( x=(x_1,x_2)=(1,3), \; y=(y_1,y_2)=(2,4) \)

Hypothesis: \( H_0: x = y \) vs. \( H_1: x > y \)

Test stat: \( f(x,y) = x_1 + x_2 - y_1 + y_2 \)

\[ f \text{ large } \rightarrow H_1 \text{ is more likely} \]

Permutations

\[
(1,3)(2,4) \quad (2,4)(1,3) \quad (1,2)(3,4) \quad (3,4)(1,2) \quad (1,4)(3,2) \quad (3,2)(1,4)
\]

\[
\begin{align*}
f & \quad -2 & \quad 2 & \quad -4 & \quad 4 & \quad 0 & \quad 0 \\
\end{align*}
\]

\( p\text{-value}= \frac{4}{6} \)

We do not reject \( H_0 \)
Permutation test on persistent diagrams

Max $t = 3.9507$
Min $t = -3.0961$
95 percentile $= 3.6432$
5 percentile $= -4.0237$

More pairings for the control subjects
= More cortical folding
History of permutation test

Fisher 1935, The Design of Experiment

\[
\binom{8}{4} = 70
\]

Thompson et al. 2001, Nature Neuroscience

\[
\binom{40}{20} = 1.34 \cdot 10^{11}
\]

Nichols et al. 2002, Human Brain Mapping

4279 citations

\[
\binom{6}{3} = 20
\]
Graph filtration based network analysis

Attention deficit hyperactivity disorder (ADHD)

Autism spectrum disorder (ASD)

Pediatric controls (PedCon)

I-correlation

Lee. et al. 2012, IEEE Transactions on Medical Imaging
Graph theory based network analysis in year 2010.

Threshold at 0.5

Threshold at 0.7
24 attention deficit hyperactivity disorder (ADHD) children
26 autism spectrum disorder (ASD) children
11 pediatric control subjects

The normal brain networks merges to a single component faster than other clinical populations.

Lee et al. 2010 ISBI
Resting-state functional magnetic resonance imaging (fMRI)
Time series with 1200 time points at 300000 voxels per subject measured over 14min 33 seconds inside MRI scanner

416 subjects
= 131 Monozygotic (MZ) twins
77 Dizygotic (DZ) twins

4GB x 416 = 1.6TB data
Permutation test impractical if sample size > 200

>> nchoosek(200,100)

**Warning:** Result may not be exact. Coefficient is greater than 9.007199e+15 and is only accurate to 15 digits

> In nchoosek (line 92)
ans =
   9.0549e+58
Dense brain network

Brain network where each voxel is a node.

AMD vega64 GPU
Time series averaged into 116 brain regions
116 time series at 1200 time points

\[ \zeta_i(t) = \sum_{l=0}^{k} d_{li} \psi_l(t), \quad t \in [0, 1] \]

\[ \psi_0(t) = 1, \quad \psi_l(t) = \sqrt{2} \cos(l\pi t) \]

120 features → \( \mathbf{d}_i = (d_{0i}, d_{1i}, \cdots, d_{ki}) \)
Subject level brain connectivity matrix

$c_{ij} = corr(d_i, d_j)$

Correlation of Fourier coefficients
ACE model for twins

\[ \rho_{MZ} = A + C \]

Twin correlation Additive genetics Common environment

\[ \rho_{DZ} = A/2 + C \]

Falconer's formula for heritability index (HI)

\[ HI = A = 2(\rho_{MZ} - \rho_{DZ}) \]

MZ-twins share 100% of genes
DZ-twins share 50% of genes
Correlation (group) of correlation (subject)

\[ c_{ij}^1 = (c_{ij}^{11}, \cdots, c_{ij}^{1m}) \]

\[ c_{ij}^2 = (c_{ij}^{21}, \cdots, c_{ij}^{2m}) \]

\[ c_{ij}^{MZ} = \text{corr}(c_{ij}^1, c_{ij}^2) \]
MZ- and DZ-twin correlation difference

$C_{ij}^{MZ}$

$C_{ij}^{DZ}$

$h_{ij} = 2(c_{ij}^{MZ} - c_{ij}^{CZ})$

Heritability index = amount of genetic contribution
Heritable brain regions

\[ h_{ij} \geq 1 \]

Statistical significance?
Betti-plots

Monotonicity:

*Chung et al. 2019 Network Neuroscience*
\( \beta_0 \) and \( \beta_1 \) are monotone over graph filtration.

**Monotonicity of \( \beta_0 \):** Deletion of edge increases the number of connected components by at most 1. \( \beta_0 \) increases by 0 or 1.
$\beta_0$ and $\beta_1$ are monotone over graph filtration.

Monotonicity of $\beta_1$:

Euler characteristic:

$$\chi = \beta_0 - \beta_1 = p - q$$

$\beta_1 = \beta_0 - p + q$

Nodes, edges, 0, +1, fixed, -1
Betti-plots on graph filtration

MZ

DZ

\( \beta_0 \): Connected components

\( \beta_1 \): Cycles

Correlation
Exact Topological Inference (ETI)
Kolmogorov Smirnov (KS) distance

\[ G^1 = \{ G^1_\lambda : 0 \leq \lambda \leq 1 \} \]
\[ G^2 = \{ G^2_\lambda : 0 \leq \lambda \leq 1 \} \]
\[ D(G^1, G^2) = \sup_{\lambda \in [0,1]} | \beta_i(G^1_\lambda) - \beta_i(G^2_\lambda) | \]

\( D \) satisfies all the axioms of metric except identity:

\[ D(G^1, G^2) = 0 \quad \text{\( \neq \)} \quad G^1 = G^2 \]
\[ P(D(G^1, G^2) = 0) = 0 \]
Inference on Betti-plots using KS-distance

Null hypothesis:

\[ H_0 : \beta_1 (G_1^\lambda) = \beta_1 (G_2^\lambda) \quad \text{for all} \quad \lambda \]

Need to determine the probability of observed event under the null hypothesis.

Under the null, generate every possible events (sample space) by permutations.
Permutation test on monotone features

1) Combine features
   1, 3, 2, 4

2) Permutation
   3, 2, 4, 1

Observed data: (1,3) (2,4)

(0,0) (2,2)
Exact Topological Inference

Theorem

\[ D_q = \sup_{1 \leq j \leq q} |\beta_i(G_{\lambda j}^1) - \beta_i(G_{\lambda j}^2)| \]

\[ P(D_q \geq d) = 1 - \frac{A_{q,q}}{\binom{2q}{q}} \]

Permutations can be mapped one-to-one to walks on the square grid.

Chung et al. 2017 IPMI
Run time

$q = \text{Number of edges}$
Permutation test impractical if sample size > 200

>> nchoosek(200,100)

Warning: Result may not be exact. Coefficient is greater than 9.007199e+15 and is only accurate to 15 digits

> In nchoosek (line 92)
ans =
  9.0549e+58
Asymptotic

$$\lim_{q \to \infty} P\left( \frac{D_q}{\sqrt{2q}} \geq d \right) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$
Validation via simulation
The purpose of (statistical) simulation is to generate synthetic data with the ground truth, where the performance of a method can be compared against existing methods.
Network simulation

$n \times 1$ data vector $x_i$ at node $i$.

$$x_i \sim N(0, I_n) \rightarrow C = (c_{ij}) = (\text{corr}(x_i, x_j))$$

$$E C = I_n$$

Network with $k$ modules

$$y_1, \ldots, y_c = x_1 + N(0, \sigma^2 I_n)$$

$$y_{c+1}, \ldots, y_{2c} = x_{c+1} + N(0, \sigma^2 I_n)$$

$$\vdots$$

$$y_{c(k-1)+1}, \ldots, y_{ck} = x_{c(k-1)+1} + N(0, \sigma^2 I_n)$$
Matrix norm based distance

\[ \chi^1 = (V, w^1) \quad \chi^2 = (V, w^2) \]

\[ D_l(\chi^1, \chi^2) = \left( \sum_{i,j} |w^1_{ij} - w^2_{ij}|^l \right)^{1/l} \]

\[ D_\infty(\chi^1, \chi^2) = \max_{\forall i,j} |w^1_{ij} - w^2_{ij}| \]
Performance based on 100 simulations

<table>
<thead>
<tr>
<th>p=20</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_\infty$</th>
<th>GH</th>
<th>KS ($\beta_0$)</th>
<th>KS ($\beta_1$)</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 vs. 4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>5 vs. 5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>10 vs. 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>4 vs. 5</td>
<td>0.63</td>
<td>0.40</td>
<td>0.33</td>
<td>0.15</td>
<td>0.27</td>
<td>0.06</td>
<td>0.9</td>
</tr>
<tr>
<td>2 vs. 4</td>
<td>0.71</td>
<td>0.48</td>
<td>0.42</td>
<td>0.53</td>
<td>0.18</td>
<td>0.00</td>
<td>0.95</td>
</tr>
<tr>
<td>5 vs. 10</td>
<td>0.94</td>
<td>0.80</td>
<td>0.78</td>
<td>0.72</td>
<td>0.44</td>
<td>0.24</td>
<td>0.96</td>
</tr>
</tbody>
</table>

False positives

False negatives

Modularity
Performance based on 100 simulations

<table>
<thead>
<tr>
<th>p=100</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_\infty$</th>
<th>GH</th>
<th>KS ($\beta_0$)</th>
<th>KS ($\beta_1$)</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 vs. 4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.26</td>
<td>0.54</td>
<td>0.03</td>
</tr>
<tr>
<td>5 vs. 5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
<td>0.43</td>
<td>0.05</td>
</tr>
<tr>
<td>10 vs. 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>4 vs. 5</td>
<td>0.51</td>
<td>0.37</td>
<td>0.35</td>
<td>0.16</td>
<td>0.11</td>
<td>0.00</td>
<td>0.93</td>
</tr>
<tr>
<td>2 vs. 4</td>
<td>0.66</td>
<td>0.45</td>
<td>0.57</td>
<td>0.61</td>
<td>0.03</td>
<td>0.00</td>
<td>0.91</td>
</tr>
<tr>
<td>5 vs. 10</td>
<td>0.94</td>
<td>0.86</td>
<td>0.79</td>
<td>0.72</td>
<td>0.11</td>
<td>0.00</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Performance based on 100 simulations

<table>
<thead>
<tr>
<th>p=500</th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_{\infty}$</th>
<th>GH</th>
<th>KS ($\beta_0$)</th>
<th>KS ($\beta_1$)</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 vs. 4</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.08</td>
<td>0.20</td>
<td>0.26</td>
<td>0.02</td>
</tr>
<tr>
<td>5 vs. 5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>10 vs. 10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.06</td>
<td>0.18</td>
<td>0.05</td>
</tr>
<tr>
<td>4 vs. 5</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.11</td>
<td>0.00</td>
<td>0.20</td>
</tr>
<tr>
<td>2 vs. 4</td>
<td>0.14</td>
<td>0.11</td>
<td>0.14</td>
<td>0.12</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>5 vs. 10</td>
<td>0.20</td>
<td>0.18</td>
<td>0.19</td>
<td>0.16</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
</tr>
</tbody>
</table>

We need to come up with better **topology-aware** network distances!
What next?
Coidentification of cycles over multiple networks

Lee et al. 2019 MICCAI
Thank you!
Question? mkchung@wisc.edu