A test of hypotheses for random graph distributions

Andressa Cerqueira, Claudia Vargas, Daniel Fraiman, Florencia Leonardi

First Young Researchers Workshop

São Paulo, 05 de Maio de 2015

Main Goal

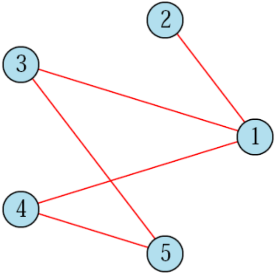
- The construction of a non parametric hypotheses test for samples of graphs;
- Application of this test to analyse brain functional networks constructed from electroencephalographic (EEG) data.

Graph

- A simple graph is a pair (V, E), where V is a finite set of vertices and $E \subseteq V \times V$ is a set of edges;

- The graph can be represented by its adjacency matrix, where

$$g_{ij} = \begin{cases} 1, & \text{if there is an edge between } i \text{ and } j \\ 0, & \text{otherwise} \end{cases}$$



$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Hypotheses test. Given two samples of graphs $\mathbf{g} = (g_1, \ldots, g_n)$ and $\mathbf{g}' = (g'_1, \ldots, g'_m)$, we want to test if they were originated from the same probability distribution, that is

$$\begin{cases} \mathrm{H}_0 : \pi = \pi' \\ \mathrm{H}_\mathrm{A} : \pi \neq \pi' \end{cases}$$

where π is the distribution which originated $\mathbf{g}$ and π' is the distribution which originated $\mathbf{g}'$.

Definition

Given two samples of graphs $\mathbf{g} = (g_1, \ldots, g_n)$ *and* $\mathbf{g}' = (g'_1, \ldots, g'_m)$ *we define the two-samples test statistic by*

$$T(\mathbf{g}, \mathbf{g}') = \max_{g \in \mathbb{G}(v)} \left| \overline{D}_{\mathbf{g}}(g) - \overline{D}_{\mathbf{g}'}(g) \right| \,,$$

where $\overline{D}_{\mathbf{g}}(g) = \frac{1}{n} \sum_{k=1}^{n} D(g, g_k)$ *and* $D(g, g_k) = \sum_{i<j} (g_{ij} - g_{ij}^k)^2$.

The critical region of the test is

$$\mathrm{R} = \{t : t(\mathbf{g}, \mathbf{g}') > q_{(1-\alpha)}\} \ ,$$

where $q_{(1-\alpha)}$ is the $(1-\alpha)-$quantile of the distribution of T under the null hypothesis (H_0).

Remark: $t \in R \Rightarrow$ we reject H_0

It is important to remark that

- We need to know the set $\mathbb{G}(V)$ to compute T;
- The set $\mathbb{G}(V)$ has $2^{\binom{|V|}{2}}$ graphs;
- If $|V| = 20$, then $|\mathbb{G}(V)| = 2^{190}$ - this is extremely LARGE !!!

How do we compute T ?

Proposition

Given two samples of graphs $\mathbf{g} = (g_1, \ldots, g_n)$ *and* $\mathbf{g}' = (g'_1, \ldots, g'_m)$ *we have that*

$$T(\mathbf{g}, \mathbf{g}') = \sum_{i<j} \left| \overline{\mathbf{g}}_{ij} - \overline{\mathbf{g}}'_{ij} \right| ,$$

where $\qquad \overline{\mathbf{g}}_{ij} = \frac{1}{n} \sum_{k=1}^{n} g^k_{ij}.$

- To compute the critical region of the statistical test we need to know the distribution of T.

What is the distribution of T?

Proposition

Let two samples of graphs $\mathbf{g} = (g_1, \dots, g_n)$ *and* $\mathbf{g}' = (g'_1, \dots, g'_m)$. *Under the null hypothesis* H_0 *we have*

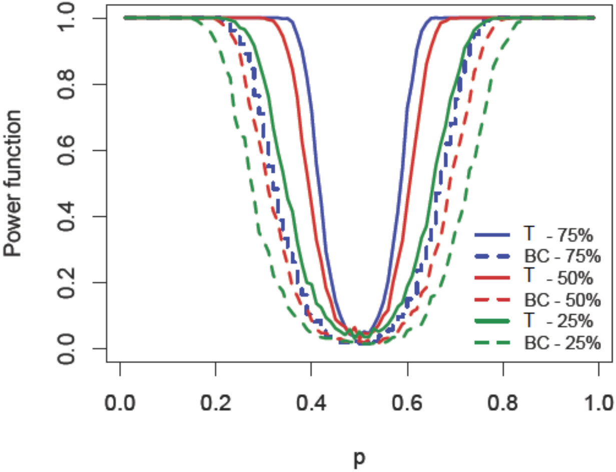
$$T(\mathbf{g}, \mathbf{g}') = \sum_{i<j} |T_{ij}|$$

$$\sqrt{\left(\frac{nm}{n+m}\right)}(T_{ij})_{ij} \xrightarrow[\substack{n\to\infty \\ m\to\infty}]{D} N(0, \Sigma)$$

where $\Sigma_{ij,kl} = \pi G_{ij,kl} - (\pi G_{ij})(\pi G_{kl})$ *and* $\pi G_{ij,kl} = \sum_{g \in \mathbb{G}(v)} g_{ij} g_{kl} \pi(g)$.

Simulation

We compared the power function of our test with the power function of the simultaneous testing procedure with Bonferroni correction (BC). The null model is the Erdös-Rényi model with parameter $p_0 = 0.5$ and the alternative hypothesis is (modified) Erdös-Rényi model with $v = 10$ nodes and $q\%$ of edges with parameter p and the remaining edges with parameter $p_0 = 0.5$.
The sample size was n=20.

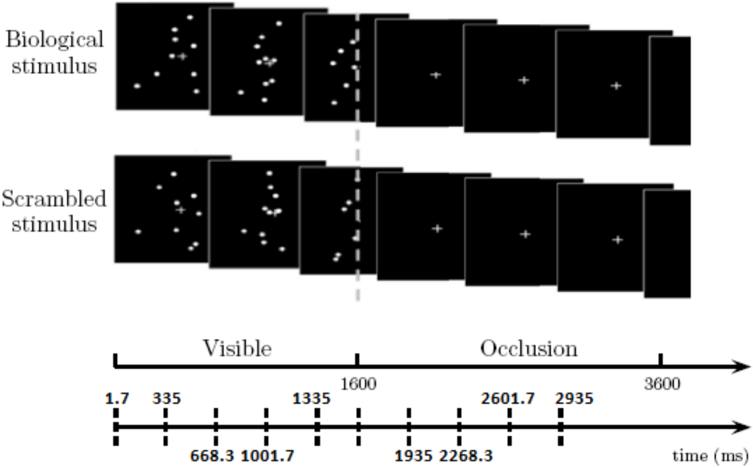


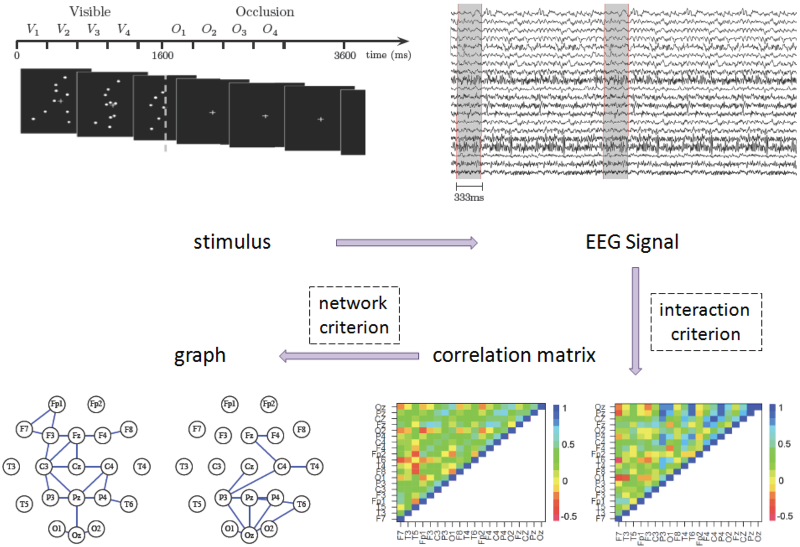
Power function
0.0
0.2
0.4
0.6
0.8
1.0
p
T - 75%
BC - 75%
T - 50%
BC - 50%
T - 25%
BC - 25%

Discrimination of EEG brain networks.

We want to compare graphs built from EEG data collected during the observation of videos depicting human locomotion.

Stimulus.


Biological stimulus
Scrambled stimulus
Visible
Occlusion
1600
3600
1.7
335
1335
2601.7
2935
668.3 1001.7
1935 2268.3
time (ms)



Visible
Occlusion
V_1 V_2 V_3 V_4
O_1 O_2 O_3 O_4
0
1600
3600
time (ms)
333ms
stimulus
EEG Signal
interaction criterion
correlation matrix
network criterion
graph
Fp1 Fp2 F7 F3 Fz F4 F8 T3 C3 Cz C4 T4 T5 P3 Pz P4 T6 O1 Oz O2
1
0.5
0
-0.5

Visible x Occlusion

- Biological movement
  - visible: 132 graphs for each window
  - occlusion: 132 graphs for each window
- Non-Biological Movement
  - visible: 132 graphs for each window
  - occlusion: 137 graphs for each window
- p-value of the test

Visible vs Occlusion	Windows			
	V_1 vs O_1	V_2 vs O_2	V_3 vs O_3	V_4 vs O_4
Biological	**0.0019**	0.4294	0.1984	0.0278
Non-biological	**0.0016**	0.8278	0.1249	0.6673

Our paper: A test of hypotheses for random graph distributions built from EEG data.

http://arxiv.org/abs/1504.06478

Acknowledgments

CNPq

FAPESP

modelagem estocástica e complexidade

NeuroMat

References

E. Bullmore e O. Sporns.
Complex brain networks: graph theoretical analysis of structural and functional systems.
Nat Rev Neurosci, 10:186–198.

J. R. Busch, P. A. Ferrari, A. G. Flesia, R. Fraiman, S. P. Grynberg e F. G. Leonardi.
Testing statistical hypothesis on random trees and applications to the protein classification problem.
The Annals of Applied Statistics, 3(2):542–563.

D. Fraiman, G. Saunier, E. F. Martins, e C. D. Vargas.
Biological motion coding in the brain: analysis of visually-driven eeg funcional networks.
Plos One. No prelo 2014.

References

- M. E. J. Newman.
  Networks: An Introduction.
  Oxford University Press.

- G. Saunier, E. F. Martins, E. C. Dias, J. M. de Oliveira, Thierry Pozzo e Claudia D. Vargas.
  Electrophysiological correlates of biological motion permanence in humans.
  Behavioural Brain Research, 236:166–174.