Three testing perspective on connectome data

Start-up research - Follow up

A. Cabassi[1], A.Casa[2], M.Fontana[3], M. Russo[2] & A. Farcomeni[4]

University of Cambridge[1]
Università degli Studi di Padova[2]
Politecnico di Milano[3]
Sapienza Università di Roma[4]

Palermo, 19th June 2018

Data and Motivation

Data

- Multimodal and mixed-domain data:
  - Structural networks: anatomical interconnections among brain regions;
  - Dynamic functional activity: dynamical activity of each brain region during fMRI;
  - Functional networks: synchronization in brain activity for each pair of brain regions
- $V = 70$ brain regions with corresponding location and lobe information;
- $n = 24$ subjects with $k = 2$ scans each and additional information on age, handedness and psychological traits.

Aim and motivation

- Provide some insights about some of the statistical issues arising when dealing with analysis of MRI scans;
- Different perspectives and goals:
  1. Test correspondence among anatomical and functional connectivity;
  2. Check quality of available data estimating the effective number of white matter fibers connecting brain regions;
  3. Define a metric for functional networks in order to test for differences in functional connectivity of different groups of people.

Functional correlations in connectomic maps

Background

- **Neurological hypothesis**: functional connectome is strongly related to the underlying structural networks;
- Nature of this relation is not completely clear yet:
  - Is it possible to infer anatomical connections from functional ones?
  - How does the relation vary with time?
- **Aim**: Is the absence of white matter fiber connecting brain regions reflected in their functional correlation?

Literature review

- **Graphical models**: probabilistic models where a graph is used to express the conditional dependence between sets of random variables;
- Let X be an $n \times p$ matrix with $\{X_{i1}, \ldots, X_{ip}\} \sim N(0, \Sigma)$ and denote the precision matrix as $\Theta = \Sigma^{-1}$;
- Associating a node to each variable, the absence of an edge connecting nodes i and j indicates conditional independence among X_i and X_j;
- Maximizing $\mathcal{L}_p(\Theta) = log|\Theta| - tr(S\Theta)$, where $S = X^T X/n$, we get $\hat{\Theta} = S^{-1} \rightarrow$ what if $p > n$?

Literature review

- Friedman et al. (2008) proposed the graphical lasso;
- **Idea**: minimize the penalized profile likelihood

$$\mathcal{L}_{pen,p} = log|\Theta| - tr(S\Theta) - \lambda\|\Theta\|_1$$

  where $\|\Theta\|_1 = \sum_{i \neq j} \Theta_{ij}$;
- It provides $\hat{\Theta}$ even when S is singular and induces a sparse representation of the dependence among observed variables.
- Several inferential tools proposed to test if conditional dependences are statistically significant.

Proposed methodology

- **Proposal**: parametric bootstrap based test to check if absence of white matter fiber between regions is reflected in absence of a functional correlation among them;
- More sintetically:

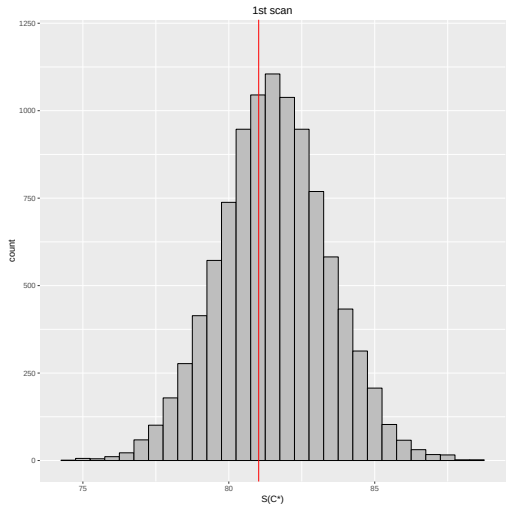
$$H_0 : \Omega = \Omega_0 \qquad \text{vs} \qquad H_1 : \Omega \neq \Omega_0$$

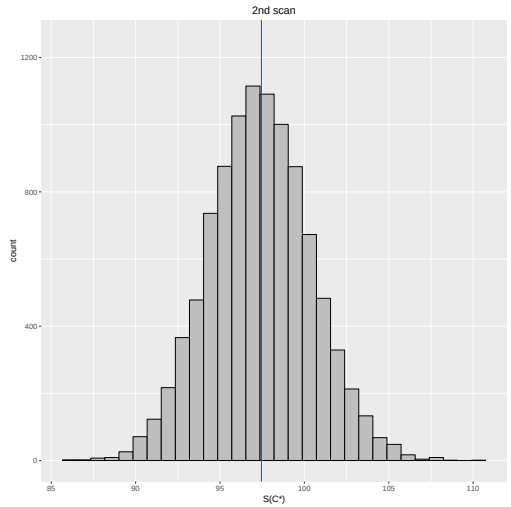
where Ω and Ω_0 are correlations matrices with the second one constrained by external information.

Proposed methodology

- X, $n \times p$ matrix of functional activities of p brain regions measured on n subjects, while D is a $p \times p$ structural network matrix;
- Estimate, via *glasso*, C^* s.t. $\left(C^*\right)^{-1}_{ij} = 0$ iff $D_{ij} = 0$ for all n subjects and obtain $C^*_1, \ldots, C^*_B$ matrices sampling from Wishart distribution with scale matrix C^*;
- Let $S(c)$ be the sum of squared correlations among unconnected regions
  $\rightarrow$ compare it with the bootstrap distribution of $S\left(c^*_i\right)$ with $i, \ldots, B$;
- Compute bootstrap p-value.

Results





- Temporal dimension is stacked allowing to consider Wishart distribution as the sampling one;
- Results are consisten with usual assumption in neuroscientific community;
- Similar results have been obtained considering LRT, even if tests have different rationale.

Conclusions

- We propose a simple and fast test to study the relation between functional and anatomical connectivity among brain regions;

New directions

- Study in greater details the properties of the test (e.g. the power) and compare with other solutions;
- Handle carefully time information;
- Incorporate spatial information (distances between regions) and characteristics of the specific subjects.

Bayesian method for fiber count validation

DTI white matter fiber count validation

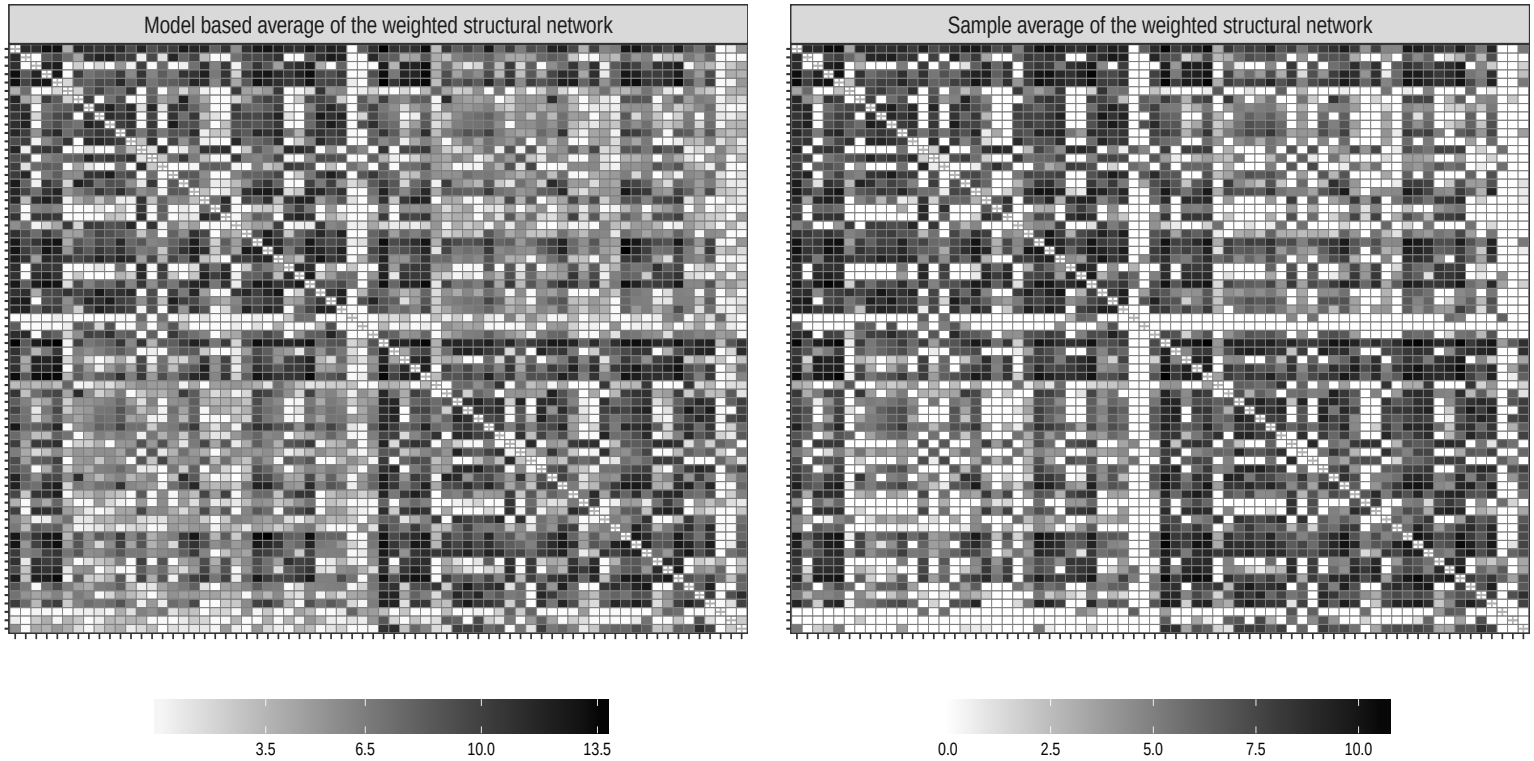
- DTI scan is a rather approximate techniques.
- It includes multiple source of variability ⇒ Scanner, lab, pre-processing & **Individual**.
- This uncertainty might lead to misleading results.

- To achieve more robust results we aim to estimate the unknown number of white matter fiber for each pair of brain region.
- We propose a **hierarchical Bayesian model**.

Our proposed approach

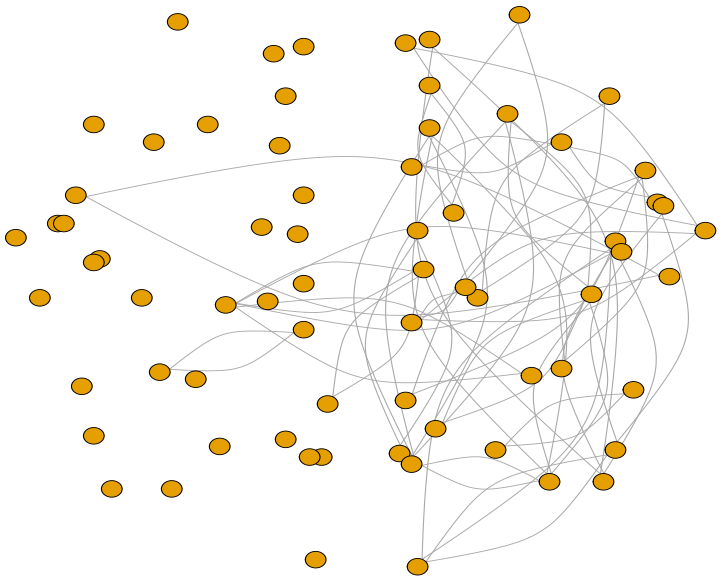
$$
\begin{aligned}
\left(\{n_{kij} : k = 1, \ldots, K\}\right) &\sim && \text{Bin}\left(M_{ij}, \pi_j\right), \\
\text{logit}\left(\pi_j\right) &= && \alpha_j + \alpha \texttt{MatchHemisphere}_j, \\
M_{ij} &\sim && \text{Pois}\left(\lambda_{ij}\right), \\
\log\left(\lambda_{ij}\right) &= && \beta_i + \beta_j + \beta \text{age}_i.
\end{aligned}
$$

Application to DTI data I



- Identified active areas agree with observed ones.
- As expected we find an higher number of white matter fibers.

Application to DTI data II



- the distribution of π_js gives info on regions in which is easier to observe connections.
- Connection with high probabilities to be observed share the same right hemisphere.

Conclusion

- DTI are still a valid source of information even if they should be used with care.
- Pre-processing and external source of information should be always be included in the model.
- In our opinion our proposed approach can mitigate undercount effect and be integrated in more refined analysis

Object-Oriented analysis of network data

Goal

Object Oriented Data Analysis: statistics for complex objects
E.g. Directed acyclic graphs, tensors, shapes, images, networks.
Main idea:

- Consider complex objects as the statistical units of our analysis;
- Analyse the data in the mathematical space in which they live.

Our goal: To define an object-oriented framework for structural and functional networks. In particular, we wish to define:

- A distance between networks;
- A test for the equality of the average networks of multiple groups.

Literature review

- Reducing each observed network to a vector of summary statistics.
- Univariate testing approaches considering each edge separately adjusted to control FDR or FWER taking into account the network structure.
- Use of auxiliary data (e.g. spatial proximity) to inform the posterior probability that some pairs of nodes interact differently.

Literature review

- Reducing each observed network to a vector of summary statistics.
- Univariate testing approaches considering each edge separately adjusted to control FDR or FWER taking into account the network structure.
- Use of auxiliary data (e.g. spatial proximity) to inform the posterior probability that some pairs of nodes interact differently.
- **Durante and Dunson (2017)** develop a Bayesian procedure for inference and testing of group differences in the network structure.

Literature review

- Reducing each observed network to a vector of summary statistics.
- Univariate testing approaches considering each edge separately adjusted to control FDR or FWER taking into account the network structure.
- Use of auxiliary data (e.g. spatial proximity) to inform the posterior probability that some pairs of nodes interact differently.
- **Durante and Dunson (2017)** develop a Bayesian procedure for inference and testing of group differences in the network structure.
- **Ginestet et al. (2017)** test the equality of two groups of networks using the concept of Fréchet mean of networks and deriving a CLT for sequences of network averages, using the Euclidean distance.

Distances

Procrustes size-and-shape distance

$$d_P(G_1, G_2) = \inf_{R \in O(D)} \|L_1 - L_2 R\| \tag{1}$$

where

L_i decomposition of G_i s.t. $G_i = L_i L_i'$, $i = 1, 2$;

$\|\cdot\|$ Frobenius norm;

D set of unitary operators.

Gromov-Wasserstein distance

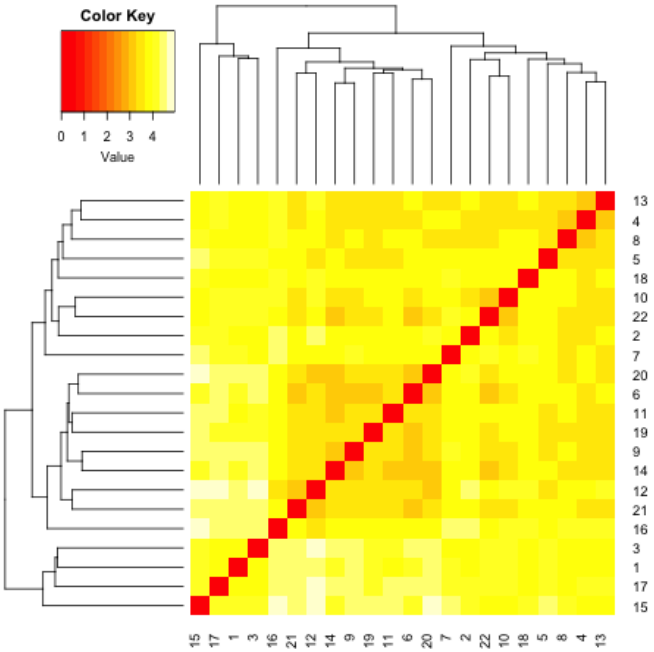
$$d_{GH} = \frac{1}{2} \inf_R \|d_X(x, x') - d_Y(y, y')\|_{L^p_{R \times R}} \tag{2}$$

where

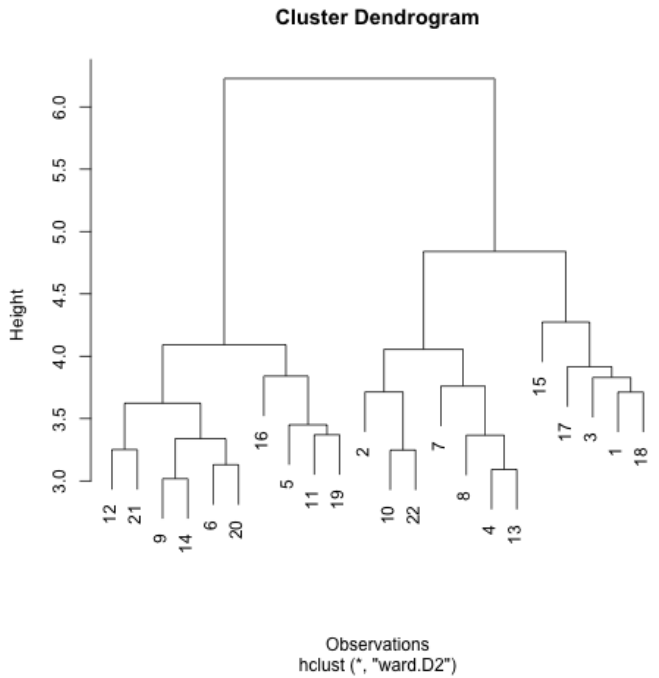
$R \in \mathcal{R}(X; Y)$, set of all correspondences between X and Y;

(X, d_X) and (Y, d_Y) compact metric spaces.

Exploratory analysis



Heatmap of the Procrustes distances.



Hierarchical clustering.

Test for the equality

$G_{11}, \ldots, G_{N_1 1}$ and $G_{12}, \ldots, G_{N_2 2}$ two groups of adjacency matrices, iid samples from 2 random processes with mean Γ_1 and Γ_2.

$$H_0 : \Gamma_1 = \Gamma_2 \quad \text{against} \quad H_1 : \Gamma_1 \neq \Gamma_2. \tag{3}$$

Test for the equality

$G_{11}, \ldots, G_{N_1 1}$ and $G_{12}, \ldots, G_{N_2 2}$ two groups of adjacency matrices, iid samples from 2 random processes with mean Γ_1 and Γ_2.

$$H_0 : \Gamma_1 = \Gamma_2 \quad \text{against} \quad H_1 : \Gamma_1 \neq \Gamma_2. \tag{3}$$

Similar strategy to the one used by Pigoli et al. (2014) for testing the equality of covariance operators of functional data, i.e. reformulate test as

$$H_0 : d(\Gamma_1, \Gamma_2) = 0 \quad \text{against} \quad H_1 : d(\Gamma_1, \Gamma_2) > 0 \tag{4}$$

Also possible to extend to the case of multiple groups (Cabassi et al. 2017).

Two-sample permutation test

Given a sample $G_1, \ldots, G_N$ of independent and identically distributed observations, the **sample Fréchet mean** is

$$\hat{\Gamma} = \arg\inf_{\Gamma} \sum_{n=1}^{N} d(G_n, \Gamma)^2. \tag{5}$$

Two-sample permutation test

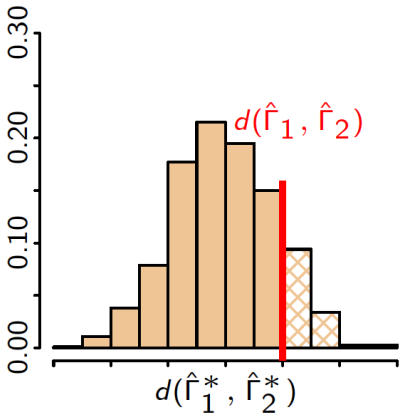
Given a sample $G_1, \ldots, G_N$ of independent and identically distributed observations, the **sample Fréchet mean** is

$$\hat{\Gamma} = \arg\inf_{\Gamma} \sum_{n=1}^{N} d(G_n, \Gamma)^2. \tag{5}$$

Algorithm

1. Compute $d(\hat{\Gamma}_1, \hat{\Gamma}_2)$, with $\hat{\Gamma}_i$ sample Fréchet mean of group i;
2. Apply B random permutations to the labels of the sample graphs;
3. For each of them compute $d(\hat{\Gamma}_1^*, \hat{\Gamma}_2^*)$;
4. The p-value of the test is

$$\lambda = \frac{\sum \mathbb{1}\left[d(\hat{\Gamma}_1^*, \hat{\Gamma}_2^*) \geq d(\hat{\Gamma}_1, \hat{\Gamma}_2)\right]}{B}.$$



Tests for the real data

Test	p-value	Adjusted p-value
1. Mental disorder diagnosis	0.914	1
2. Under vs. over 30	0.634	1
3. Under vs. over 50	0.091	0.273

Table$p$-values of the tests.

Tests for the real data

Test	p-value	Adjusted p-value
1. Mental disorder diagnosis	0.914	1
2. Under vs. over 30	0.634	1
3. Under vs. over 50	0.091	0.273

Table$p$-values of the tests.

Issues

- Probably need more observations;
- Not clear how to choose threshold for the age limit.

Concluding remarks

Summary

- No assumptions on the data generating process;
- Computationally intensive.

Future work

- Implement test using Gromov-Wasserstein distance;
- Compare to state-of-the-art methods where possible.