Stat4DS / Homework 02

Pierpaolo Brutti

Due in the... epiphany stocking?!? (on Moodle)

General Instructions

I expect you to upload your solutions on Moodle as a **single running** `R Markdown` file (`.rmd`) + its `html` output, **named with your surnames**. Alternatively, a `zip`-file with all the material inside will be fine too.

R Markdown Test

To be sure that everything is working fine, start `RStudio` and create an empty project called `HW1`. Now open a new `R Markdown` file (`File > New File > R Markdown...`); set the output to `HTML mode`, press `OK` and then click on `Knit HTML`. This should produce a web page with the knitting procedure executing the default code blocks. You can now start editing this file to produce your homework submission.

Please Notice

- For more info on `R Markdown`, check the support webpage that explains the main steps and ingredients: R Markdown from RStudio or, equivalently, read about Quarto. For more info on how to write math formulas in LaTex: Wikibooks.
- Remember our **policy on collaboration**: *collaboration on homework assignments with fellow students is* ***encouraged****. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had* ***discussions*** *(no more) concerning your solution. You may* ***not****, however, share written work or code after discussing a problem with others. The solutions should be written by* ***you and your group*** *only.*

Exercise: Connect your brain

1. Background: MRI and fMRI

Since its invention in the early 70s by Lauterbur and Mansfield (2003 Nobel prize in Physiology and Medicine), **magnetic resonance imaging** (MRI) has evolved into a versatile tool for the in vivo examination of tissue. Unlike X-ray computed tomography (CT) and positron emission tomography (PET), it does not rely on high energetic radiation but on the nuclear magnetic resonance phenomenon. Consequently, it does in principle not harm the examined tissue and can be applied also in healthy subjects. Thus, MRI is a perfect tool for the examination of the **living brain** in neuroimaging.

Functional magnetic resonance imaging (fMRI) is a technique to examine the human (or animal) brain "at work". fMRI is used to analyze (neuro-)scientific questions, e.g., on the localization of neural capabilities, on the consequences of neuronal diseases or on brain function. For this, in fMRI, a **time series** of MRI volumes is acquired, while the subject in the scanner is typically performing some cognitive task.

What fMRI images visualize is the so called blood oxygenation level–dependent (BOLD) contrast: as active neurons rely on increased oxygen supply, the neural activity is related to a local change in support of blood oxygenation. Thus, fMRI can be used as a natural, yet indirect, contrast for detecting neural activity. In order to achieve a sufficient temporal resolution the spatial resolution of fMRI is typically limited. An fMRI dataset then consists of more than 100 image volumes with a spatial voxel dimension of about 2-4 mm.

⇝ IMPORTANT DISCLAIMER ⇜

Data from fMRI experiments suffer from several artifacts that require special preprocessing ahead of the statistical analysis, like *slice time correction*, *motion correction*, *registration*, *normalization*, *brain masking* and *brain tissue segmentation*.

For the sake of this exercise, I'll provide you with a clean, pre-processed dataset extracted from the *Autism Brain Imagine Data Exchange* (ABIDE) project, but be aware that these early data analytic stages are crucial and not at all trivial.

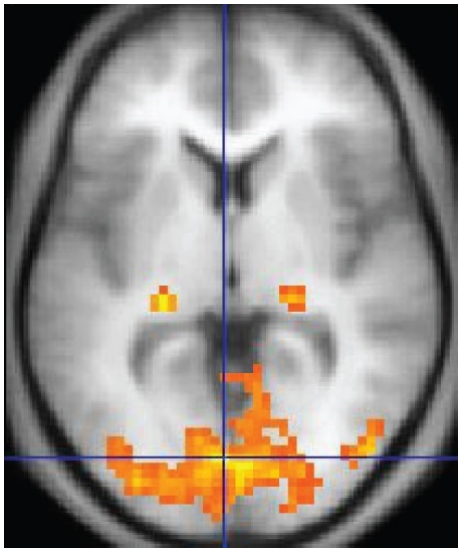


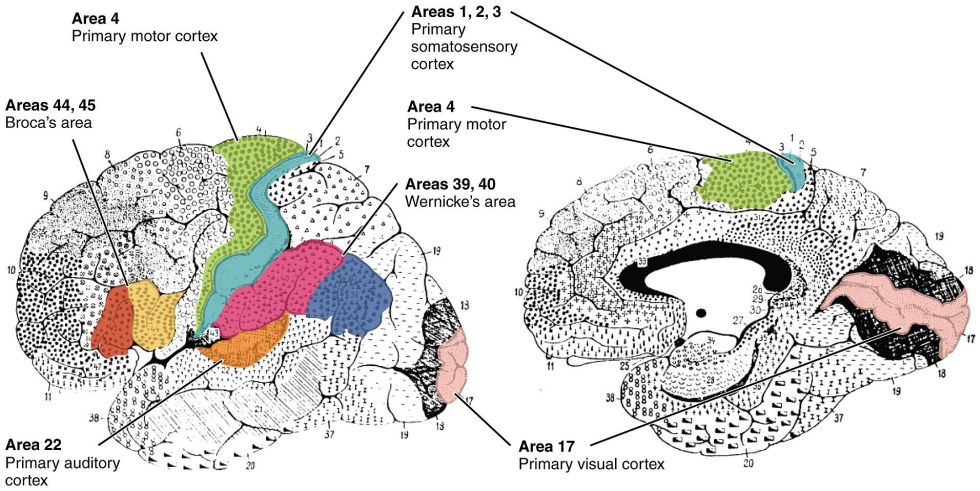
Figure 1: An fMRI image with yellow areas showing increased activity compared with a control condition

2. The Task: Functional Connectivity

The development of MRI and fMRI has paved the way to **connectomics**, i.e., modeling the brain as a network in order to tackle fundamental neuroscience research questions as a *graph-analysis* problem.

Generally speaking, a *connectome* is a map describing neural connection between brain **regions of interest** (`ROIs`), either by observing anatomic fiber density (*structural connectome*), or by computing a suitable statistical association measure (e.g. Pearson correlation) between time series of activity associated to `ROIs` (*functional connectome*). Of the two, the latter is the case of interest to us and from now on we will focus on it.

Nevertheless, before going any further, we need to clarify what these *regions of interest* actually are. Typically, `ROIs` are defined in terms of a suitable **functional brain atlas** which provides information about the spatial location of functional brain regions aggregating knowledge on brain functionality and anatomy accumulated over more that 100 year of brain research. In other words, we essentially use *these* atlases – yes, *these*, because there's more than one – to tag fMRI voxels with specific cortical brain regions. The oldest atlas system dates back to the German anatomist Brodmann who defined 52 cortical areas based on the cytoarchitectural organization of neurons.



This is all nice and good, but to attach an observed fMRI voxels to a specific area of your functional atlas of choice we first need to *normalize* each individual brain or, in other words, we need to map it onto a "standard brain" in order to then be able to identify the corresponding brain regions. As an example, Talairach coordinates, also known as *Talairach space*, is one famous 3-dimensional coordinate system (atlas) that uses Brodmann areas as the labels for brain regions.

3. The Toolkit: Association Graphs

We said that **functional connectivity** addresses the interaction between cortical brain regions, and it is usually quantified by measuring the level of dependency between the observed fMRI time series in these regions. A brain network can then be defined by creating a link between two `ROIs` that exhibit a *co-activation* (e.g. a **strong correlation** in their **time series**).

⇝ IMPORTANT DISCLAIMER ⇜

At this point, if you REALLY paid attention, you may argue that all the (plug-in) estimators we introduced for common association measures like Pearson/Spearman/Kendall correlation, or Distance Correlation, or any other for what matters, is based on an random (i.e. **independent**) sample, not a stream of **temporally dependent** data. Well...good catch!

As a matter of fact we *could* be more "respectful" of the time-dependency shown by the data, but, as done in a large portion of the current literature on this topic, in the following and in your implementation we will simply **ignore it**.

Okay, so, the problem is one of evaluating dependency between cortical regions: time to put the **association graph** machinery introduced in class to good use. Here's a quick recap...

Let $\mathcal{D}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ be a random sample (IID) from some joint D-dimensional distribution $f_{\boldsymbol{X}}(\cdot)$ where the random vector $\boldsymbol{X}_i = \big[X_i(1), \ldots, X_i(\mathrm{D})\big]^{\mathsf{T}} \in \mathbb{R}^{\mathrm{D}}$. The **vertices** (nodes) of the graph refer to the D features/variables, whereas the **edges** represent *relationships* between them.

The graph is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{V_1, \ldots, V_{\mathrm{D}}\}$ is the vertex–set and $\mathcal{E}$ the edge–set. We can regard the edge–set $\mathcal{E}$ as a $(\mathrm{D} \times \mathrm{D})$ adjacency matrix $\mathbf{E}$ where $\mathbf{E}(j,k) = 1$ if there is an edge between feature j and feature k and 0 otherwise. Alternatively, you can regard $\mathcal{E}$ as a list of *unordered* pairs where $\{j,k\} \in \mathcal{E}$ if there is an edge between j and k.

⇝ IN PRACTICE ⇜

For what concerns our application:

1. $n = 145$ represents the length of the time series recorded on a single subject/patient: again, yes, for the sake of this exercise we will drop the time dependency and treat them as IID data.
2. $\mathrm{D} = 116$ will be the number of cortical regions of interest (`ROIs`) from the functional atlas of choice (the AAL atlas, to be precise)
3. Each features and, consequently, each node/vertex of the association graph, correspond to one of the $\mathrm{D} = 116$ `ROIs`.

Denoting by $\rho(j,k)$ any association measure, in an **association graph** $\mathcal{G}(t) = \big(\mathcal{V}, \mathcal{E}(t)\big)$ we put an edge between V_j and V_k if

$$\big|\, \rho(j,k) \,\big| \geqslant t \quad \text{for any} \quad j,k \in \{1, \ldots, \mathrm{D}\},$$

that is, if the association between feature j and feature k is "strong" enough. The choice of ρ and the threshold t are typically application specific, but often we set $t = 0$, in which case there is an edge if and only if $\rho(j,k) \neq 0$. We also write $\rho_{j,k}$ or $\rho(X_j, X_k)$ to mean the same as $\rho(j,k)$.

Remember that the **population parameter** ρ is required to have the following property:

$$X \text{ independent from } Y \quad \overset{\text{implies}}{\Rightarrow} \quad \rho(X,Y) = 0.$$

In general, the reverse may **not** be true. We will say that ρ is *strong* if

$$X \text{ independent from } Y \quad \overset{\text{if and only if}}{\Leftrightarrow} \quad \rho(X,Y) = 0.$$

Let's mention, for example, the recent work by Bergsma and Dassios (2014) – TauStar package – who extended the Kendall's τ into a *strong* correlation, and also the now well–known distance covariance – energy package – defined in Szekely et al. (2007).

⇝ GOAL ⇜

Starting from the data $\mathcal{D}_n$ and for a specific choice of the threshold t, we need to get a statistically sound estimate $\widehat{G}_n(t) = \big(\mathcal{V}, \widehat{\mathcal{E}}_n(t)\big)$ of the **true** population graph $G_n(t) = \big(\mathcal{V}, \mathcal{E}_n(t)\big)$.

⇝ IN PRACTICE ⇜

Once we get a $(1-\alpha)$ confidence interval $\mathrm{C}_n^{j,k}(\alpha)$ for $\rho(j,k)$, we can then place an edge between feature j and feature k whenever $[-t, +t] \cap \mathrm{C}_n^{j,k}(\alpha) = \emptyset$ (the empty-set).

For the **Pearson correlation**, for example, we know we can build these intervals via nonparametric bootstrap or going for the asymptotic Fisher Z-transform: of course the bootstrap is a viable alternative for any other association measures!

This idea is perfectly fine in case we have only a small number of edges/cortical areas to check. In general though, the graphs are characterized by a large number of nodes ($\mathrm{D} = 116$), and consequently, a huge amount of edges: $m = \binom{\mathrm{D}}{2}$ to be precise.

⇝ IMPORTANT DISCLAIMER ⇜

We necessarily need to control for multiplicity in order to meaningfully talk about the overall graph topology by avoiding a ridiculous overflow of false edges/discoveries.

⇝ IN PRACTICE ⇜

The easiest procedure to implement – although quite conservative – is the so called Bonferroni correction that simply asks for adjusting the nominal level of the intervals from α to α/m where $m = \binom{\mathrm{D}}{2}$ is the number of intervals we are building, but of course there are viable, less conservative, alternatives.

As an alternative, partial (linear) correlations can be considered. In our setup, they effectively measure the association between two `ROIs` **with the (linear) effect of series in all other regions removed**.
As highlighted in class, partial correlations can be effectively computed from inverse covariance or precision matrices. More specifically, if $\mathbf{R}_{(p)} = \left[\rho^{(p)}_{j,k}\right]_{j,k}$ denotes the $(\mathrm{D} \times \mathrm{D})$ matrix of partial correlations, then

$$\mathbf{R}_{(p)}[j,k] = -\frac{\Lambda_{j,k}}{\sqrt{\Lambda_{j,j} \cdot \Lambda_{k,k}}},$$

where $\Lambda = \Sigma^{-1}$ is the precision matrix associated to the covariance matrix Σ.

In the **low–dimensional** setting ($\mathrm{D} << n$), we can estimate $\mathbf{R}_{(p)}$ as follows. Let $\widehat{\Sigma}_n$ be the sample covariance and $\widehat{\Lambda} = \widehat{\Sigma}_n^{-1}$ its inverse. Then define

$$\widehat{\mathbf{R}}_{(p)}[j,k] = \widehat{\rho}^{(p)}_{j,k} = -\frac{\widehat{\Lambda}_{j,k}}{\sqrt{\widehat{\Lambda}_{j,j} \cdot \widehat{\Lambda}_{k,k}}}.$$

Similarly to Pearson correlation, there is also a large sample Z-transform based *Normal approximation* for the sampling distribution of $\widehat{\rho}^{(p)}_{j,k}$, namely

$$Z_{j,k} = h\big(\widehat{\rho}^{(p)}_{j,k}\big) = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}^{(p)}_{j,k}}{1-\widehat{\rho}^{(p)}_{j,k}}\right) \quad \text{then} \quad Z_{j,k} \dot{\sim} \mathrm{N}_1\left(\theta_{j,k}, \frac{1}{n-g-3}\right), \tag{1}$$

where $\theta_{j,k} = h\big(\rho^{(p)}_{j,k}\big)$ and $g = (\mathrm{D}-2)$.

An implementation of this method is provided by the function `sinUG()` in the package `R` package `SIN`. The only difference is that, instead of Bonferroni, it adopts a less conservative correction for multiplicity.

⇝ IMPORTANT DISCLAIMER ⇜

In **high–dimensional** settings – i.e. $\mathrm{D} \approx n$ or even $\mathrm{D} >> n$ – this protocol will **not** work since $\widehat{\Sigma}_n$ is **not** invertible. In fact, as can be seen from Equation 1,

$$\mathbb{V}\mathrm{ar}\big(\widehat{\rho}^{(p)}_{j,k}\big) \approx \frac{1}{n-\mathrm{D}},$$

and this blows up when D is close to, or even larger than, n.

In such a bad situation, we have at least two exit strategies:

1. Apply some **shrinkage** (statistical lingo) or **regularization** (more math inclined) scheme to "fix" the ill–conditioned matrix $\widehat{\Sigma}_n$ and make it non–singular + bootstrap on the entries of the resulting matrix. See, for example, Schager and Strimmer (2005) and Ledoit and Wolf (2004).
2. Use the *graphical lasso* – `glasso` package.
   Warning! The reliability of the graphical lasso depends on lots of non-trivial, uncheckable assumptions.

4. The Data: The `Autism Brain Image Data Exchange` Project

Up to now we have essentially sketched how to generically learn the brain connectivity from data coming from a single person.

An important related problem is to identify connection patterns that might be associated to specific cognitive phenotypes or mental dysfuntions. Given fMRI scans of (many) patients affected by a mental disorder and scans of (many) healthy individuals, the goal is to discover patterns in the corresponding connectomes that explain differences in the brain mechanism of the two groups. This task can be formalized in many ways. Here we will approach it as a *multiple comparison problem*.

To be more precise, in this exercise we use a publicly available dataset released by the `Autism Brain Imagine Data Exhange` (ABIDE) project. The dataset contains neuroimaging data of patients suffering from *Autism Spectrum Disorder* (`ASD`) and *Typically Developed* (`TD`) subjects. Since fMRI data are strongly influenced by a variety of **confounding factors**, in an effort to mitigate this intrinsic variability we will consider only male patients with an age between 15 and 20 years (adolescents)[1].

⇝ GOAL ⇜

Denoting by $\mathcal{G}^{\texttt{ASD}}(t)$ and $\mathcal{G}^{\texttt{TD}}(t)$ the two **true** (but *unknown*) association graphs characterizing the `ASD` and the `TD` group respectively, our statistical goal is to learn (from data) if and how $\mathcal{G}^{\texttt{ASD}}(t)$ differs from $\mathcal{G}^{\texttt{TD}}(t)$ at a prefixed threshold t.

We can reformulate this problem in a slightly different way. More specifically, for any association measure ρ and threshold $t \geqslant 0$ we pick, we can define the **difference graph** $\mathcal{G}^{\Delta}(t) = \big(\mathcal{V}, \mathcal{E}^{\Delta}(t)\big)$ where we put an edge between V_j and V_k if

$$\big|\Delta_{j,k}\big| = \big|\,\rho^{\texttt{ASD}}_{j,k} - \rho^{\texttt{TD}}_{j,k}\,\big| \geqslant t \quad \text{for any} \quad j,k \in \{1,\dots,\mathrm{D}\}.$$

Here $\{\rho^{\texttt{ASD}}_{j,k}\}_{j,k}$ and $\{\rho^{\texttt{TD}}_{j,k}\}_{j,k}$ denote the **true** association coefficients between `ROI` j and `ROI` k in the `ASD` and `TD` group.

⇝ IN PRACTICE ⇜

As before, to get an estimator $\widehat{\mathcal{G}}^{\Delta}(t)$ of $\mathcal{G}^{\Delta}(t)$, we can start by building a $(1-\alpha)$ confidence interval $\mathrm{C}^{j,k}_n(\alpha)$ for $\Delta_{j,k}$, and then place an edge between `ROI` j and `ROI` k whenever $[-t,+t] \cap \mathrm{C}^{j,k}_n(\alpha) = \emptyset$ (α **must** adjusted for multiplicity).

⇝ Your job ⇜

1. Load the pre–processed data contained in the file `hw2_data.RData`. You will get two `lists` of length 12, `asd_sel` and `td_sel` (one per group). In other words we have data from 12 `ASD` subjects and 12 `TD` subjects. Each slot of these lists contains a `data.frame` of size (145×116): the 116 columns are related to different `ROIs`, whereas the 145 rows are the observation times. Data coming from different subjects can/must also be considered independent from each other.

   REMARK: you've 12 subjects per group ⇝ decide how to **pool** together their data (whatever the choice, please explain).

2. Let ρ be the Pearson correlation. Set the threshold t equal to a meaningfully high (trial-and-error, explain your choice) percentile of the correlation values observed in the two groups (HINT: you can simply use `cor()` inside an `lapply()` to make `asd_sel` and `td_sel` into lists of 116×116 correlation matrices).
   Use **Fisher Z-transform + Bonferroni correction** to get two separate estimates, $\widehat{\mathcal{G}}^{\texttt{ASD}}(t)$ and $\widehat{\mathcal{G}}^{\texttt{TD}}(t)$, of the *true* association graphs based on 95% asymptotic confidence intervals for $\{\rho^{\texttt{ASD}}_{j,k}\}_{j,k}$ and $\{\rho^{\texttt{TD}}_{j,k}\}_{j,k}$.

3. If you haven't already, take a look at basic tools to deal with graphs in `R` such as the `igraph`, `ggraph` packages. Graphically represent all the estimated graphs and try to draw some conclusion: are there clear co-activation differences between the two groups? What happens if you skip the Bonferroni correction and work with unadjusted intervals? For a fixed α, if you vary the threshold t from small to large values, what connections are the more "resilient"; that is, the *last ones to die*? Are there differences in the two groups?

4. Repeat the analysis using this time the **partial correlation coefficient** and compare the results.

[1] To extract the data, we have followed a preprocessing strategy called DPARSF, followed by a band-pass filtering + global signal regression. To parcellate the brain we adopt the AAL atlas (116 `ROIs`). The final result for a single patient is a set of 116 time series of length 145 each.