Stat4ACSAI / Homework 03

Pierpaolo Brutti

Due Monday, May 16 (on Moodle…with tolerance)

General Instructions

I expect you to upload your solutions (one per team) on Moodle as a **single running** `Notebook` (`.ipynb`) **named with your surnames** + pictures converted and collected in a single `pdf`-file of your handwritten exercises. Alternatively, a `zip`-file with all the material inside will be fine too.

Your responses must be supported by textual explanations, the code **you** wrote to produce the results and **extensive comments** on both the code and the results.

Please Notice

- 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 1: Major League Baseball

Polynomial regression is linear regression with increasing powers of a single covariate as explanatory variables:

$$\mathbb{E}(Y_i|\boldsymbol{x}_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p.$$

Quadratic regression is simply polynomial regression up to degree $p = 2$.

The dataset `MLB` contains the horizontal distance (in feet) traveled by a baseball hit at various angles (the initial speed of the ball at the bat is constant).

1. Download the data and perform quadratic regression with `Distance` as response variable and `Angle` as explanatory variable, with both `Python` routines and matrix computation. Comment on the goodness of fit of the model.
2. What is the average distance we would observe with an angle of 27 degrees? What is the interpretation of $-\beta_1/2\beta_2$?
3. Compute the correlation between `Distance` and `Angle`. Would simple linear regression be adequate for this problem? Motivate your answer.

Exercise 2: 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.

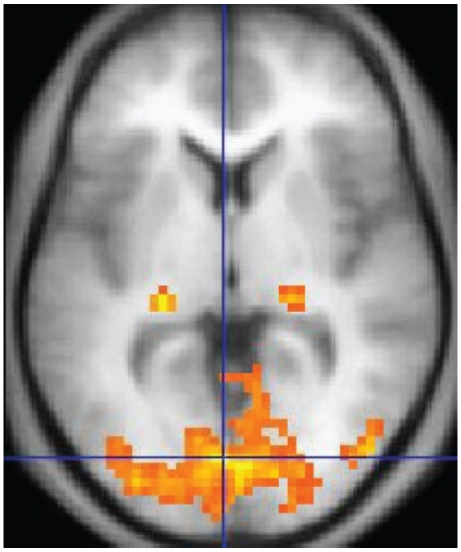


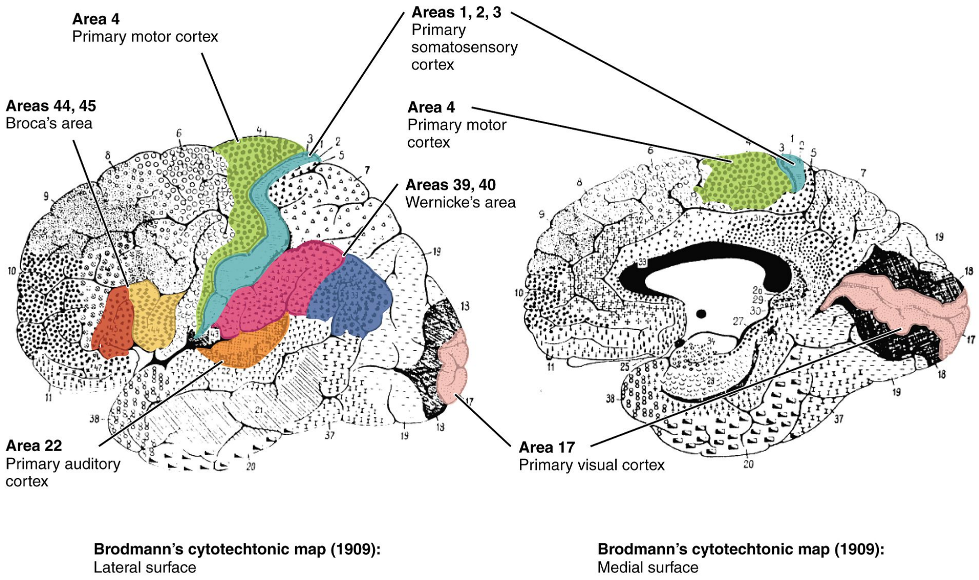
Figure 1: An fMRI image with yellow areas showing increased activity compared with a control condition

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.



Brodmann's cytotechtonic map (1909): Lateral surface

Brodmann's cytotechtonic map (1909): Medial surface

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, in particular those for a common association measure like Pearson correlation, are 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: introducing **association graphs**! Here's a quick summary…

Let $\mathcal{D}_n = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\}$ be a random sample (IID) from some joint D-dimensional distribution $f_{\boldsymbol{X}}(\cdot)$ where the random vector $\boldsymbol{X}_i = \left[X_i(1), \dots, X_i(\mathrm{D})\right]^{\mathsf{T}} \in \mathbb{R}^{\mathrm{D}}$. 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 $\mathcal{V} = \{V_1, \dots, V_{\mathrm{D}}\}$ 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* measure of *association* like *Pearson's correlation*, 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 **true** (at the population level!) 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 $\rho(j,k)$.

Now, 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.

⇝ IN PRACTICE ⇜

We necessarily need to control for multiplicity in order to meaningfully talk about the overall graph topology by avoiding a ridiculous overflow of false discoveries/edges (here $H_0^{j,k} : \big|\rho_{j,k}\big| < t$ or $H_0^{j,k} : \rho_{j,k} = 0$ for all $j < k$).

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.

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

In this exercise we use a (*very small part of 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].

⇝ Your job ⇜

1. Load the pre–processed data for a single TD subject and a single ASD subject contained in the two files `td_data.csv` and `asd_data.csv` respectively. Each one will correspond to a data-matrix of size (145×116): the 116 columns are related to different ROIs, whereas the 145 rows are the observation times.

2. Let ρ be the Pearson correlation. Use **Nonparametric Bootstrap** + **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% bootstrapped confidence intervals for $\{\rho_{j,k}^{\texttt{ASD}}\}_{j,k}$ and $\{\rho_{j,k}^{\texttt{TD}}\}_{j,k}$.

3. With the help of the HW03 – Addendum, graphically represent the estimated graphs playing around with the threshold t to filter out "weakly" associated ROIs.

   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?

[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.