Statistical Learning

Due on Moodle 48 hours before the discussion day

Final Homework

Exercise 1: 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.

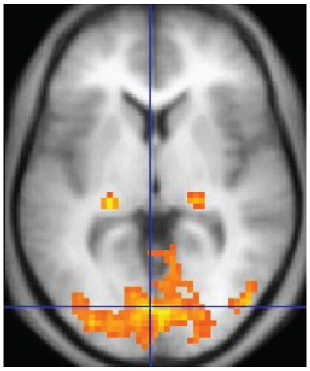


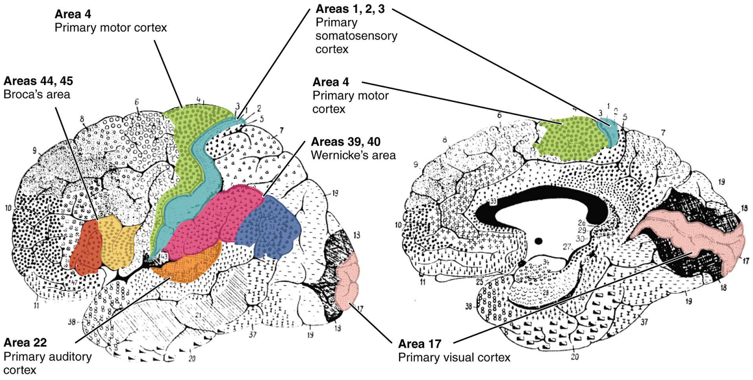
Figure 1: An fMRI image with yellow areas showing increased activity compared with a control condition

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.

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

In this exercise we use a (*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 also consider age and sex as additional covariate[1].

I will provide you with time series of activity associated to 116 distinct ROIs observed at 115 time instants in order to predict mental dysfunctions (...possibly building and then using a connectome structure for each subject in the study...). So, in other words, given data extracted from fMRI scans of patients affected by a mental disorder (autism) and scans of healthy individuals (control), the goal is to discover patterns that explain differences in the brain mechanism of the two groups.

More specifically, the **training** dataset has columns/features that can be broken down in the following way:

- id: simply the row index (ignore)
- age: the age of the subject
- sex: the sex of the subject
- y: the target variable to predict (only available in training, of course...)
- From column 5 on, you will find 116 vectorized **time series** (one for each ROI) of length 115 each

⇝ Your job (A) ⇜

Compete in the HW03 Hackathon to show me how good you are!

As always, I expect you to upload on Moodle a well commented working code that covers the entire pipeline with all the due explanations behind your choices: from data loading/pre-processing and feature engineering/dim-reduction to model fit and prediction on test.

[1] To extract the data, I have followed a preprocessing strategy called DPARSF, plus 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.

Exercise 2: Variable Importance (Skip if you're presenting in the first session)

1. Introduction

Being able to quantify the importance of a covariate in predicting a response of interest is crucial in most real applications... even more so nowadays, when the use of very complex, nonlinear, overparametrized models is the rule rather than the exception.

Despite this, the very idea of "importance" is slippery and need to be precisely framed, defined and handled (...yes, even when talking about linear models!). Here we'll focus on a very simple and general technique.

2. The LOCO

At page 32 of their paper, Lei and coauthors proposed a simple, general and, essentially assumptions–free idea for measuring variable importance, called **leave-one-covariate-out** (LOCO) inference. The algorithm is extremely simple.

Let $\ell(y, \widehat{y})$ be a suitable error measure/metric/loss for the learning task at hand. Then,

1. Randomly split the training data into two, non overlapping, parts: $\mathsf{D}_n = \mathsf{D}_{n_1}^{(1)} \cup \mathsf{D}_{n_2}^{(2)}$ with $n_1 + n_2 = n$.
2. Run any algorithm you like to compute an estimate $\widehat{f}_{n_1}(\cdot)$ on first part $\mathsf{D}_{n_1}^{(1)}$.
3. Select some variable $\boldsymbol{X}[j]$ of interest to you, and recompute $\widehat{f}_{n_1}^{-j}(\cdot)$ on $\mathsf{D}_{n_1}^{(1)}$ again (rerun algorithm without access to variable $\boldsymbol{X}[j]$).
4. Use $\mathsf{D}_{n_2}^{(2)}$ to construct finite-sample, distribution-free confidence interval (e.g., use non-parametric bootstrap or sign-test or Wilcoxon-test) for the following new (population) parameter[2]:

$$\theta_j\left(\mathsf{D}_{n_1}^{(1)}\right) = \text{median}_{(Y,\boldsymbol{X})}\left(\ell\big(Y, \widehat{f}_{n_1}^{-j}(\boldsymbol{X})\big) - \ell\big(Y, \widehat{f}_{n_1}(\boldsymbol{X})\big) \,\Big|\, \mathsf{D}_{n_1}^{(1)} \right), \quad j \in \{1, \ldots, p\}. \tag{1}$$

   Since you're using the same dataset to build more than one confidence-interval, apply any reasonable correction to adjust for multiplicity (e.g. Bonferroni or Benjamini-Hochberg FDR).

θ_j has a very clear interpretation without resorting to linearity or any other *uncheckable* model assumption: it's just how much extra prediction error you would pay by not having access to variable $\boldsymbol{X}[j]$.

In addition, from a more technical point of view, this parameter is "smooth" enough (Hadamard differentiable) to guarantee the success of resampling techniques like the nonparametric bootstrap. In other words, confidence intervals, tests, etc for θ_j are easy to obtain no matter what is the underlying predictive model you picked (LASSO, LASSO + CV, Random Forest, Boosting, Deep nets, etc).

Of course there are also problems with the LOCO. More specifically:

1. It is **not** on an intuitive scale but we could fix this by rescaling.
2. Results are tied to our choice of the algorithm. In theory we could use a "meta–cross–validation" scheme that cycles over different candidate predictor classes (computational expensive but trivially parallelizable).
3. Results are also sensitive to: the ratio of training to test set sizes; the metric $\ell(\cdot,\cdot)$ selected; the correlation between features.
4. It is conditional on $\mathsf{D}_{n_1}^{(1)}$, meaning that it measures "*how important is variable X_j, to our algorithm's estimates on $\mathsf{D}_{n_1}^{(1)}$?*"...not obvious at all how to fix this...

⇝ Your job (B) ⇜

1. Starting from the model used in *Exercise 1*, evaluate variable importance via LOCO for a subset of 10 features. More in particular:
   - If the technique you picked comes with a *default* method to evaluate variable importance (e.g. Random Forest), compare the top-10 in this ranking with their LOCO scores.
   - On the other hand, if the technique you picked does not come with a default method to evaluate variable importance, pick 10 features providing some reasoning behind your choice.
2. Provide a suitable visualization of the results and comment.

[2] Alternatively, we could also consider the median of the ration between the two losses.

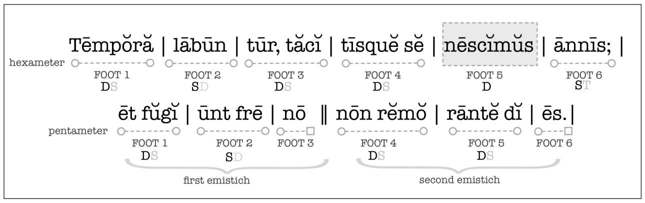
Exercise 3: It’s a question of `style`...

1. Background

One of the main open problems is Latin literature is the authorship and the chronological dating of the so called Double Heroides, six books of a poem traditionally attributed to Ovid with a certain degree of uncertainty.
In this exercise we consider the problem of establishing the period of composition of these books based on a single metric feature extracted from Ovid’s poetic production (**27 books** each considered as a single observation).
The peculiarity of this analysis is in the nature of this variable: its observed values are frequency distributions ⇝ we will handle it as compositional data via a suitable Mercer Kernel.

Basics of metric: the elegiac couplet

Meter is the basic rhythmic structure of a verse. We here focus on the so called *elegiac couplet* or *distich*, a poetic form initially introduced in Greek lyric, later on adopted by Roman poets, and in particular by Ovid. To illustrate the “anatomy” of the elegiac couplet, let us consider one of the 406 distichs from Ovid’s Fasti VI:



The main steps in the metrical analysis are summarized as follows:

- Each syllable of the words of a verse is categorized as *long* (−) or *short* (∪) according to its *weight*. (e.g. in `Tempora`, `Tem-` is long, `-po-` and `-ra` are short).
- Specific sequences of syllables define a *foot* (delimited by |...|), for instance

Dactyl (D)	formed by 3 syllables	− ∪ ∪
Spondee (S)	formed by 2 syllables	− −
Trochee (T)	formed by 2 syllables	− ∪

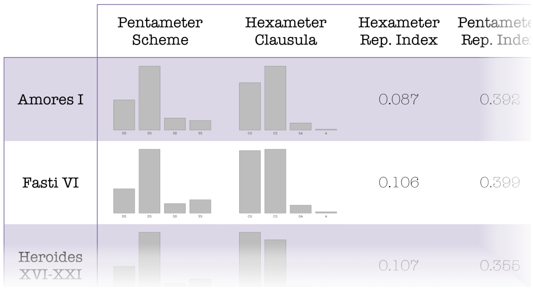
- Each verse is formed by a specific number of *feet*. The 1$^{\text{st}}$ verse of the distich is called *hexameter*; the 2$^{\text{nd}}$ one *pentameter*.
- The metrical pattern of feet in a verse is then summarized by a sequence of letters. For instance, in the couplet above the scheme is DSDDDS · DD − DD−

The choice of a particular scheme for the verses, together with many other metrical features, yield a great variability in the realization of the elegiac distich, and strongly characterizes the style of each single author. In this sense, the stylometric analysis can be helpful in the attempt of attributing a poem to a specific author.

Stylometric features

The goal of a metric study is to identify characterizing stylistic features of a poet with respect to the metric language of the tradition, as well as to detect deviations of metric features of some parts of his own poetic production with respect to his entire work.

All the poetic production in elegiac couplets attributed to Ovid has been examined from a metrical point of view in Ceccarelli (2012). In this study many quantitative information has been collected on each section of *every* poem. Most of these variables are frequency distributions summarizing the occurrence of a specific metrical phenomenon over a whole poem, and therefore they should be treated as compositional data. The figure below shows 3 variables observed on two of the poems: the 1$^{\text{st}}$ and the 2$^{\text{nd}}$ consist of distributions associated to the realizations of the pentameter scheme, and to the clausula in the exameter respectively, whereas the third scalar one is another particular stylometric quantity.



	Pentameter Scheme	Hexameter Clausula	Hexameter Rep. Index	Pentamete Rep. Inde
Amores I	DD DS SD SS	CG CS GA A	0.087	0.392
Fasti VI	DD DS SD SS	CG CS GA A	0.106	0.399
Heroides XVI-XXI			0.107	0.355

2. Data & Tools

For each of the 27 books we are interested in, the `esa_scheme_only.csv` dataset contains the observed distribution of the 16 allowed hexameter schemes, together with the book name (`opera`) and `period` of production (1^{st}, 2^{nd} and 3^{rd}, the latter coinciding with Ovid exile to Tomis). Notice that the *Double Heroides* (denoted as `Ovid. Her. XVI-XXI`) do **not** have an associated period because of their attribution problem.

Compositional Data

Compositional data have been commonly observed in many scientific fields, such as bio-chemistry, ecology, finance, economics, to name just a few. The most notable aspect of compositional data is the **restriction on their domain**, specifically that the sum of the variables is fixed. The compositional domain is **not** a classical vector space, but instead a (regular) **simplex**:

$$\Delta^d = \left\{(x_1, \ldots, x_{d+1}) \in \mathbb{R}^{d+1} \;\text{ such that }\; \textstyle\sum_{j=1}^{d+1} x_j = 1, x_j \geqslant 0\right\}$$

Arguably the most prominent approach to handle data on a simplex is to take a *log-ratio* transformation (for which one has to properly handle zeros in the dataset, of course!).

For our purposes we would like to handle compositional data via a suitable Mercer Kernel. However, our usual choices for a kernel (e.g. linear kernel, polynomial kernel and Gaussian kernel), do **not** directly apply to compositions since the data belongs to a simplex.

Fortunately, there is an easy "fix" to our beloved Gaussian kernel: simply substitute the euclidean norm in the exponent with the **Aitchison distance** between composition profiles. More specifically, let $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d+1})$ for $i \in \{1, \ldots n\}$ be the observed compositions, then the *compositional Gaussian kernel* between two generic observations r and s is given by:

$$k(\boldsymbol{x}_r, \boldsymbol{x}_s) = \mathrm{e}^{-\frac{1}{h} d(\boldsymbol{x}_r, \boldsymbol{x}_s)} \quad \text{where} \quad d(\boldsymbol{x}_r, \boldsymbol{x}_s) = \sqrt{\textstyle\sum_{j=1}^{d+1} \left\{\ln\left(\frac{x_{r,j}}{\operatorname{geo}(\boldsymbol{x}_r)}\right) - \ln\left(\frac{x_{s,j}}{\operatorname{geo}(\boldsymbol{x}_s)}\right)\right\}^2},$$

and $\operatorname{geo}(\boldsymbol{x})$ denotes the *geometric mean* of $\boldsymbol{x}$. Here h is the usual tuning parameter that can be set in different ways (e.g. equal to the observed *median* distance).

⇝ Your job (C) ⇜

1. Implement and then use the **kernel based two-sample test** described in class (see our notes and also the original paper) to check if there is a statistically significant stylometric difference in Ovid's work between the 3^{rd} and the other two periods combined. Comment properly and extensively the results.
2. Be creative and use the kernel machinery above to check if the *Double Heroides* are stylometrically compatible with Ovid's work (or at least with its work in one of the three periods).