Stat4DS / Homework 03

Pierpaolo Brutti

Due before next semester (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: When Brutti and Galasso collide...

1. The Two-sample Testing Problem

If I was a bit smarter in allocating time-to-topic this semester, one of the first example of hypothesis testing you'd learn would be the **two-sample test** problem. Hence, before anything else, I strongly suggest to go back to Moodle, grab my slides on Hypothesis Testing, read the section on Wald test included the "Compare the algorithms'' example (slides 19-27), and possibly follow along with the provided Zoom recordings from last year.

Now, based on this intro, you might think that this problem is old news. But it has had a revival: there is a lot of recent research activity on this seemingly simple problem. What makes the problem still interesting and challenging is that, these days, we need effective **high dimensional** versions.

Incidentally, most of the following comments and remarks about two-sample tests also applies to the problem of testing whether two random variables/objects X and Y are **independent**. The reason is that testing for independence really amounts to testing whether two distribution are the same, namely, the joint distribution $F_{X,Y}$ for the bivariate random vector (X, Y) and the product distribution $F_X \cdot F_Y$.

The Problem

We observe two **independent** samples of size n and m respectively:

$$\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \overset{\text{iid}}{\sim} F_{\boldsymbol{X}} \qquad \text{and} \qquad \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\} \overset{\text{iid}}{\sim} F_{\boldsymbol{Y}}.$$

Please notice that, in general, these are random vectors.

The problem is to test the null hypothesis

$$\mathsf{H}_0 : F_{\boldsymbol{X}} = F_{\boldsymbol{Y}} \qquad \text{vs} \qquad \mathsf{H}_1 : F_{\boldsymbol{X}} \neq F_{\boldsymbol{Y}}.$$

Various tests that tackle this problem typically define a test statistic T which is a function of the overall dataset $\mathcal{D}_{n,m} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$. As usual, we reject H_0 if $(T > t)$ for some **critical value** t.
We choose t such that, for a pre-fixed level $\alpha \in (0, 1)$, **IF** H_0 is **true**, then

$$\mathbb{P}_{\mathsf{H}_0}(T > t) \leqslant \alpha,$$

where $\mathbb{P}_{\mathsf{H}_0}(\cdot)$ is the distribution of T when H_0 is true.

Some of the test statistics T can be pretty complicated. This raises the question of how we choose t or, in other words, how do we find the distribution of T under the null hypothesis.
There are four main ways:

1. We may be able to find $\mathbb{P}_{\mathsf{H}_0}(\cdot)$ exactly and explicitly.
2. We may be able to find the large sample limit of $\mathbb{P}_{\mathsf{H}_0}(\cdot)$ and use this approximation.
3. Bootstrapping.
4. Permutation testing.

In all likelihood the latter, that is the *permutation approach*, is the most general although somewhat underrated technique to retrieve $\mathbb{P}_{\mathsf{H}_0}(\cdot)$ and it works for essentially any test statistics T. More details can be found in my *other* slide-set on testing.

Regarding the choice of the test statistics T, instead, there are plenty of choices. Here's a few interesting options that, nevertheless, will **not** be the focus of this homework:

- **Energy Test**: do you remember distance covariance? Well, you should, because Szekely and Rizzo (2004, 2005) defined a test (which later they turned into a test for independence: see Szekely and Rizzo 2009) based on estimating the distance covariance we defined in class. In 2012, Sejdnovic, Gretton, Sriperumbudur and Fukumizu showed that, under certain conditions, the resulting test corresponds exactly to the following, *kernel based* family of testing methods.
- **Kernel Test**: kernel machines are everywhere in ML, and for a number of years, a group of researchers has used kernels and reproducing kernel Hilbert spaces (RKHS) to define very general tests. Two key references are: Gretton, Fukumizu, Harchaoui, Sriperumbudur (2012), Gretton, Borgwardt, Rasch, Scholkopf and Smola (2012). These authors derive many interesting properties including the limiting distribution of the test statistics. But, for practical purposes, we don't need to know the distribution. We simply apply the permutation method!
- **The Cross-Match Test**: there is a long tradition of defining two sample tests based on certain "local coincidences". The cross-match test due to Paul Rosenbaum (2005) is a relatively new tool with the added benefit of not requiring the (computationally demanding) permutation methods. This might not seem like a big deal but it is useful if you are doing many tests.

At this point you may ask, is there a *best* test? Of course the answer is no! More precisely, the **power of the test** depends on $F_{\boldsymbol{X}}$ and $F_{\boldsymbol{Y}}$. No test will uniformly dominate (in power) any another test while assuring a prefixed Type-I error control at level α. If you are undecided about which test to use, you can always combine them into one meta-test. Then you can use the permutation method to get the p-value!

But there is more...

A Classifier is all you need... almost...

Many scientific questions are naturally posed as two-sample tests. For example, say we are interested in determining whether a particular brain region responds differently for a person with a medical condition (patient) and a person without the condition (control). Often, one collects and analyzes brain data for different normal and ill patients under the same stimulus to study the effect of that medical condition.

Since the work of Golland and Fischl (2003) where the authors examined permutation tests for classification with application to neuroimaging analysis, it has been increasingly common to assess whether there is a significant difference between the two sets of data collected by **learning a classifier** to differentiate between them (because, for instance, they may be more familiar with classification than two-sample testing). Neuroscientists call this style of brain decoding as pattern discrimination and a positive answer can be seen as preliminary evidence that the mental process of interest might occur within the portion of the brain being studied.

This classification approach to two-sample testing has been considered in many other application areas including particle physics, genetics, speech analysis, credit scoring, churn prediction, and video analysis.

In practice, researcher familiar with machine learning but not the hypothesis testing literature often find it intuitive to perform testing in the following way: first learn a classifier, and then **see if its accuracy is significantly different from chance**

and if it is, then conclude that the distributions are different. This approach is surely very appealing to the practitioner but not trivial at all. In particular only recently there has been an effort to better understand the **power** achievable at a fixed false positive level α by specific families of classification-based testing techniques.

The Friedman's way

The idea highlithed in the previous section is that the **held-out accuracy** of any classifier in any dimension can be used as the test statistic, and Type-I error can always be controlled non-asymptotically at the desired level using permutations as suggested above. Hence, the main interest is in studying the **power** of such a test as a function of the classifier selected and of the peculiarities of the underlying distributions.

But we can also follow a different route. As a matter of fact, the idea of using **binary classifiers** for two-sample testing was conceptualized by Friedman (2004). However, his proposal was fundamentally different:

1. Train a classifier on **all** data-points.
2. Use that classifier to assign a score to each point.
3. Compare the scores in each class using a **univariate** two-sample test like Mann-Whitney or Kolmogorov-Smirnov.

In other words, Friedman proposed using classifiers to **reduce a multivariate two-sample test into a univariate one**. And this is what you'll do in this homework!

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

In this exercise we use (again) 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).

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`).

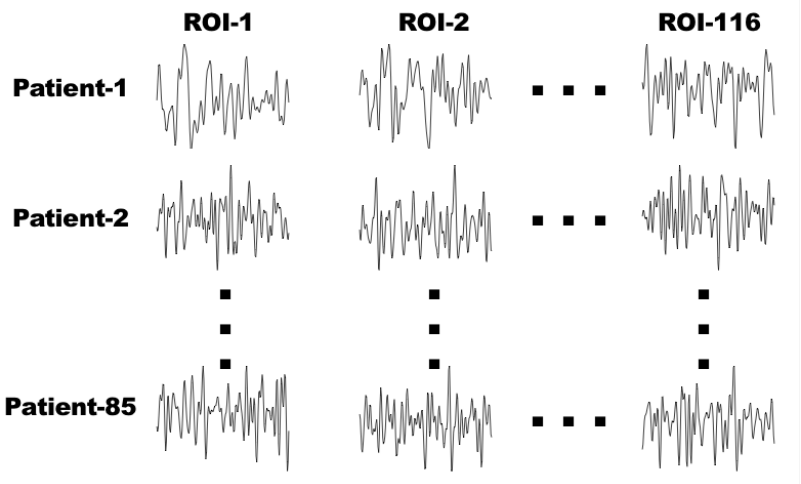
The final result are two `lists` called `asd_data` and `td_data` containing fMRI data for 85 and 93 patients respectively. For each patient we observed a set of 116 time series (one per `ROIs`) but this time with different lengths depending on the patient. To give you an idea, let's load the data and make some basic stats:

```
# Load the data
load("hw3_data.RData")
# Count how many patients in the ASD group have time series of a given length
sort( table( sapply(asd_data, function(x) nrow(x)) ), decreasing = T )
```

```
##
## 175 235 295 145 115 195 245 205  77
##  15  13  13  12  10  10   6   5   1
```

So, for 15 out of 93 patients the recorded time series are 175 time instants long, for 13 patients 235 time instants, and so on.

Important Remark: for the purpose of this homework it is quite important to start thinking about the data in each group (`ASD` and `TD`) as a very structured rectangular array having the `ROIs` as columns/features as displayed below:



⇝ Your job ⇜

1. Review/study the concepts of size, power and p-value. As mentioned already, go back to Moodle, grab my slides on Hypothesis Testing, and follow along with the provided Zoom recordings from last year.

2. Read and understand Friedman's paper. Write a short essay ($<$ 1 A4 page) to summarize it. Also explain (without copy-paste from Wikipedia or ChatGPT!) the essence of the auxiliary tests adopted (i.e. Mann-Whitney or Kolmogorov-Smirnov) and why we use them and not others.

3. Now go back to professor Galasso's material and pick any binary classifier you feel confortable with (e.g. *logistic regression* ⇝ in `R` see `?glm`). Feel free to ask me (on Moodle!) info on packages in `R` that implement a specific technique. Based on this classifier, implement Friedman's procedure and **setup a simulation study** to gather information on the size and power of the associated test under different scenarios.

   In practice, once you decide the simulation size M, you should pick: a sample size, say n_0 and n_1, for each of the two classes, say `0` and `1`; the dimension of the feature vector, say k; and the conditional k-variate distributions $F_0(\cdot)$ and $F_1(\cdot)$ from which you sample the actual data.

   Remember that here we are testing

   $$\mathsf{H}_0 : F_0 = F_1 \qquad \text{vs} \qquad \mathsf{H}_1 : F_0 \neq F_1.$$

   Hence, as also summarized in my additional scribbles, to approximate the power of our test, you must work under H_1 by picking two different (k-variate) distributions to sample from. Clearly, for fixed sample sizes n_0 and n_1, the closer (in some distance) F_0 is to F_1, the harder it is to distinguish them based on the finite sampling information provided ⇝ the lower the power achievable. Consequently here the main question of genuine mathematical interest is how the power varies as we tweak the distance between F_0 and F_1. You may play around with k, n_0 and n_1 and possibly any other relevant quantity.

   As usual, properly comment all the results and complement the numbers with suitable plots.

4. Time to apply Friedman's procedure to our fMRI data. To start with, identify the `TD` group with class `0` and the `ASD` group with class `1`. Easy.
   Now, the next big step is the crucial one: although functional data analysis is a thing, here you will proceed in a different way by *manually* embedding our original 116 functional features (one per `ROI`) in a vector space of your choosing.

   So, pick any number and type of relevat summary/statistics and extract them from each time series of the available patients. You can also reuse ideas and code from the previous homework if so you wish.

   Critically **comment** and **explain** your choices also in light of the classification performance achieved.

   In the end, based on the available data and the choices you made (the classifier, the embedding, etc), is $F_{\texttt{ASD}}$ significantly different from $F_{\texttt{TD}}$?
   No surprise here: properly comment all the results and complement the numbers with suitable (and possibly nice) plots.