**Statistical Inference II - PhD ECOSTAT - University of Milano-Bicocca**

# Homework 5

*To submit via e-mail by June 29, 2020.*

## 1

*fMRI data analysis*
Multi-subject event-related fMRI. These data sets comprise contrast images from single-subject fMRI analyses or 'first-level' analyses from the repetition priming experiment described in

Henson, R.N.A., Shallice, T., Gorno-Tempini, M.-L. and Dolan, R.J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex, 12, 178-186.

The dataset contains 12 contrast images (one per subject) which are then used in a 'second-level' analysis allowing you to make inferences about the population from which the subjects were drawn.

Analyse the data by using tools for post-selection inference on the interesting regions. Interesting regions may be data-driven or knowledge-based (or some mix of the two). A form of domain-knowledge may be an atlas of the brain provides locations and names of interesting structures and subdivisions of the brain, and this side information can be used to drive interpretations of the results.

## 2

*Analysis of Politics and U.S. Economy*

Consider the data from the @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance. We are interested in the association between Politics and U.S. Economy. Perform either

1. A more rigorous analysis on the connection between politics and the economy. See for example the work of Larry Bartels (Unequal Democracy: The Political Economy of the New Gilded Age) and Alan Blinder and Mark Watson (Presidents and the Economy: A Forensic Investigation).
2. $p$-hackaton data challenge: using the data provided, find the smallest (i.e.) most significant $p$-value. Propose a correction for the resulting inferences.

## 3

*Variable importance in random forests*
Breiman's popular permutation technique (2001) for Random Forests (RF) Variable Importance (VI) has been criticized for failing to properly account for correlations between features; for example see

- Nicodemus et al., (2010) The Behaviour of Random Forest Permutation-Based Variable Importance Measures under Predictor Correlation
- Gregorutti et al., (2015) Grouped variable importance with random forests and application to multiple functional data analysis

One solution is presented in

- Watson, Wright (2019+) 'Testing Conditional Independence in Supervised Learning Algorithms'

Summarise RF VI pitfalls and compare the standard approach to your solution to some dataset (either simulated or real)

## 4

*Testing the precision matrix in Gaussian graphical models*
Consider a random sample of $n$ realisations of an $m$-dimensional normal random vector $X \sim \mathsf{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^m$ and $\Sigma$ an $m \times m$ positive definite. It is well known that there is a simple relationship between the zero pattern of the precision matrix $\Omega = (\omega_{ij}) = \Sigma^{-1}$ and the relation of conditional independence between components of $X$:

$$w_{ij} = 0 \equiv X_i \text{ is independent of } X_j \text{ given } X_{-(i,j)},$$

where $X_{-(i,j)}$ denotes the components of $X$ after removing $X_i$ and $X_j$.

On the other hand, conditional independence is closely related to the relation of graph separation in undirected graphs. Namely, if we consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, m\}$ is the set of nodes and $E = \{(i, j) : X_i \text{ and } X_j \text{ are not conditionally independent given } X_{-(i,j)}\}$ is the set of edges, than all relations of conditional independence in $X$ can be read of from $G$. For instance, whenever two sets of nodes $A, B \in V$ are separated by $S \in V$, then $X_A$ is conditionally independent from $X_B$ given $X_S$. This property has been greatly utilised in the area of Gaussian graphical models.

Given the connection between $\Omega$ and $G$, think about possible ways of estimating $G$ by means of hypothesis testing.

## 5

*Universal Inference Using the Split Likelihood Ratio Test*
Wasserman, Ramdas and Balakrishnan (2020+) propose a general method for constructing hypothesis tests that have finite sample guarantees without regularity conditions. The method is very simple and is based on a modified version of the usual likelihood ratio statistic, called the *split likelihood ratio test.*
The method is especially appealing for irregular statistical models. Find an application (either one proposed in the paper or one of your own) and try it (best if with real data).

## 6

*SARS-CoV-2 data*
Apply one or more of the inferential methods discussed in the course (or something related to the course, like selection, multiplicity, post-selection inference etc.) to SARS-CoV-2 data.

## 7

*Adaptive Benjamini-Hochberg*
Given $m$ ordered $p$-values $p_1 \leq \ldots \leq p_m$ for $m$ hypotheses:

1. If $p_i > \dfrac{i\alpha}{m}$ for all $i$, set $R = 0$ or if $p_m \leq \alpha$, set $R = m$;
2. Otherwise, compute the upper bound $\hat{m}_\alpha$ for $m_0$, where

$$\hat{m}_\alpha = \max\Big\{i \in \{1, \ldots, m-1\} : p_{m-i+j} > \frac{j\alpha}{i} \text{ for } j = 1, \ldots, i\Big\} \tag{1}$$

and set

$$R_\alpha = \max\Big\{i \in \{1, \ldots, m-1\} : p_i \leq \frac{i\alpha}{\hat{m}_\alpha}\Big\}.$$

This adaptive Benjamini-Hochberg procedure rejects the $R_\alpha$ hypotheses with smallest $p$-values. The FDR of the procedure can be bounded by

$$\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\Big(\frac{|R_\alpha \cap M_0|}{|R_\alpha| \vee 1}\Big) \\
&= \mathbb{E}\Big(\frac{|R_\alpha \cap M_0|}{|R_\alpha| \vee 1}|\hat{m}_\alpha \geq m_0\Big)\mathrm{P}(\hat{m}_\alpha \geq m_0) + \mathbb{E}\Big(\frac{|R_\alpha \cap M_0|}{|R_\alpha| \vee 1}|\hat{m}_\alpha < m_0\Big)\mathrm{P}(\hat{m}_\alpha < m_0) \\
&\leq 2\alpha
\end{aligned}$$

Can you improve the lower bound of $2\alpha$ by assuming a Dirac-uniform configuration?
*Assumption*
Assume the Dirac-uniform configuration $Du(m_0)$, where $m_0$ $p$-values corresponding to true hypotheses are independent uniform $U(0, 1)$, while $m - m_0$ $p$-values corresponding to false hypotheses follow a Dirac distribution with point mass 1 in zero.
This is an open problem. Useful references will be provided.

## 8

*Propose your project and write a description. It should contain the application of one or more inferential methods discussed in the course (or something related to the course).*