Modern Inference (Statistical Learning) - University of Milano-Bicocca

Homework 4

To submit via e-mail by 21/04/2020 h 14:00.

You will have to submit your solution by the deadline. Feel free to choose the format for your solution (.txt, .tex, .pdf etc.), the nicer the better, and the number of files in attachment (one or more, but not too many). Answer with clarity and precision. All R code must be reproducible. For theoretical questions, try to provide a well-reasoned mathematical argument. Simulations can help form your intuition; but a purely empirical answer will only receive partial points. It is encouraged to discuss the problem sets with others, but every group needs to turn in a unique write-up. Use of sources (people, books, internet and so on) without citing them in homework sets results in failing grade.

1 Two-step Benjamini-Hochberg

Suppose we wish to test $m > 1$ hypotheses $H_1, \dots, H_m$. In this problem, we are interested in a procedure which operates in two-step:

Step 1: Select a set $\mathcal{S} \subseteq \mathcal{H} = \{H_1, \dots, H_m\}$ of "interesting" hypotheses; consider the selection rule:

$$\mathcal{S} = \{H_i \in \mathcal{H} : p_i \leq \alpha\}$$

for some pre-specified $\alpha \in (0, 1)$

Step 2: Apply the Benjamini-Hochberg method at level α to the selected set of hypotheses $\{H_i, i \in S\}$

Calculate the FDR of this Benjamini-Hochberg two-step procedure by assuming that all m hypotheses are true and $p_1, \dots, p_m$ are i.i.d. Uniform(0,1). Would you expect FDR control at level α? Explain why and why not.

2 Storey method under positive dependence

Perform a simulation study to compare the FDR control and power for the following procedures:

- BH: original Benjamini-Hochberg procedure at level $\alpha = 0.05$
- STO: Storey adaptive procedure with $\lambda = 1/2$ at level $\alpha = 0.05$
- OBH: Oracle Benjamini-Hochberg: the Oracle knows the true value of π_0, and this corresponds to perform Benjamini-Hochberg procedure at level $\alpha = 0.05/\pi_0$

In the simulation study, consider

- The number of tests m set at $m = 32, 128, 512$
- The fraction of the true null hypotheses $\pi_0 = m_0/m = 1, 0.75, 0.5$
- The p-values generated in the following way. First, let $Z_0, Z_1, \dots, Z_m$ be i.i.d. $N(0, 1)$, Next, let $Y_i = \sqrt{\rho} Z_0 + \sqrt{1 - \rho} Z_i + \mu_i$ for $i = 1, \dots, m$ and let $p_i = 1 - \Phi(Y_i)$
- The correlation $\rho = 0, 0.5, 0.9$
- The values of μ_i are 0 for $i = 1, \dots, m_0$, and $\mu_{m_0+1} = 1$, $\mu_{m_0+2} = 2$, $\mu_{m_0+3} = 3$, $\mu_{m_0+4} = 4$, $\mu_{m_0+5} = 1$, $\mu_{m_0+6} = 2$ etc. (this cycle 1, 2, 3, 4, is repeated to produce the desired m_1 values)

For each configuration of m, π_0, ρ, replicate many times to estimate

- FDR = Average (number of type I errors / number of rejections). If you get 0 rejections, the ratio evaluates 0.
- Average power = Average (correct rejections / number of false hypotheses) where correct rejections is the number of false hypotheses rejected.

for BH, STO and OBH. Comment your results.

3 Letters

Suppose we have a moderate-scale spatial problem, structured in a $m = 100 \times 100$ square matrix, where for each unit (pixel) we measure the local magnitude of the effect.

Moreover, imagine that in the substantive theory of our data, it is the shape of the area of effects that is of primary interest. The "Alphabet theory" conjectures that effects manifest themselves only with the shape of the letters of the alphabet A, B, C, ..., Z. You have

- `pvals.RData` is a 100×100 matrix of $m = 10000$ p-values
- `letters.RData` is a list of 26 matrices 100×100 giving the shape of the 26 letters

Find the the most likely letter and justify your finding.

For example, you can provide an estimate/lower bound for the amount of signal present within a selected region. For this, you can use the R package `hommel`: given 10 p-values in locations 1 to 10, the 95% lower bound for number of effects in the first 5 locations can be evaluated by

```
set.seed(102)
m <- 10
pvalues <- c(runif(0.5*m,0,0.02), runif(0.5*m,0,1))
res <- hommel(pvalues)
locations <- 1:5
discoveries(res, ix=locations, alpha=0.05)
```