Statistics

List 5

Let $X_1, \ldots, X_n$ be the independent identically distributed random variables coming from the population with the continuous cumulative distribution function F. Once more, we test the hypothesis

$$H_0 : F = F_0 \quad \text{against the alternative} \quad H_1 : F \neq F_0, \tag{1}$$

where F_0 is a known cumulative distribution function.
We define the new variables $U_1 = F_0(X_1), \ldots, U_n = F_0(X_n)$. Then, the testing problem (H_0, H_1) is equivalent to verifying

$$H_0 : U_1 \sim U(0,1) \quad \text{against} \quad H_1 : U_1 \nsim U(0,1), \tag{2}$$

where $U(0,1)$ denotes the uniform distribution on $(0,1)$.

This time, we will analyze three data-driven tests, which are based on a data-driven selection of the number of summands in the Neyman's smooth statistic. Recall, that the Neyman's smooth statistic with the k components has the form

$$N_k = \sum_{j=1}^{k} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j(U_i) \right\}^2, \tag{3}$$

where $\{b_j\}_{j \in \mathbb{N}}$ is the orthonormal system of the Legendre's polynomials in $L^2((0,1), du)$.
We will select the number of summands in the statistic N_k using the rule S, T, and A. Specifically,

(i) the simplified Schawrz (BIC) selection rule has the form

$$S = \min\{k : 1 \leq k \leq K, \;\; N_k - k \log n \geq N_j - j \log n, \;\; j \in \{1, \ldots, K\}\}, \tag{4}$$

(ii) the simplified Akaike (AIC) rule has the form

$$A = \min\{k : 1 \leq k \leq K, \;\; N_k - 2k \geq N_j - 2j, \;\; j \in \{1, \ldots, K\}\}, \tag{5}$$

(iii) the rule T has the form

$$T = \min\{k : 1 \leq k \leq K, \;\; N_k - \Pi(k,n) \geq N_j - \Pi(j,n), \;\; j \in \{1, \ldots, K\}\}, \tag{6}$$

where

$$\Pi(k,n) = k \log n \, \mathbf{1}\Big(\max_{1 \leq j \leq K} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j(U_i) \Big| \leq \sqrt{c \log n} \Big) + 2k \, \mathbf{1}\Big(\max_{1 \leq j \leq K} \Big| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j(U_i) \Big| > \sqrt{c \log n} \Big), \tag{7}$$

for some $c > 0$, while $\mathbf{1}(\cdot)$ is the indicator of the set $\cdot$. Set $c = 2.4$ and $K = 12$.

Under the null model, the statistics N_S and N_T have an asymptotic chi-square distribution with 1 degree of freedom. We reject the hypothesis H_0 for large values of the statistic N_S, N_T, N_A.

Exercise 1.

Repeat the numerical experiment from List 3 for the tests based on the statistics N_S, N_T, N_A. Discuss the results and compare them with the outcomes from List 3.

References

Inglot, T., Ledwina, T. (2006). Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Applications*, 417, 124–133.