University of Wrocław: Data Science

Theoretical Foundations of Large Data Sets, List 3

Definitions

$$V_i(t) = \mathbf{I}_{\{p_i \le t\}}, \qquad F_n(t) = \frac{\sum V_i}{n}.$$

- Higher Criticism Test (Tukey 1976):

$$HC^\star = \max_{1/n < t < 1/2} \sqrt{n} \frac{F_n(t) - t}{\sqrt{t(1-t)}}$$

- Modification by Stepanova and Pavlenko (2014)

$$HC_{mod} = \max_{0<t<1} \sqrt{n} \frac{F_n(t) - t}{\sqrt{t(1-t)q(t)}}, \qquad q(t) = \log\log\frac{1}{t(1-t)}.$$

1. For $n \in \{5000; 50000\}$ estimate the probability of the type I error for HC_{mod} using the asymptotic critical value for 0.05 significance test $C_{crit} = 4.14$.

2. For $n = 5000$ estimate critical values of both Higher-Criticism tests at the significance level $\alpha = 0.05$.

3. Let $n = 5000$ and

   a) $\mu_1 = 1.2\sqrt{2\log n}, \mu_2 = \ldots = \mu_n = 0;$

   b) $\mu_1 = \ldots = \mu_{100} = 1.02\sqrt{2\log\left(\frac{n}{200}\right)}, \mu_{101} = \ldots = \mu_n = 0;$

   c) $\mu_1 = \ldots = \mu_{1000} = 1.002\sqrt{2\log\left(\frac{n}{2000}\right)}, \mu_{1001} = \ldots = \mu_n = 0;$

   Use the above settings to compare the power of the following tests and summarize the results:

   - Higher-Criticism,
   - modified Higher-Criticism,
   - Bonferroni,
   - chi-square,

- Fisher,
- Kolmogorov-Smirnov (ks.test),
- Anderson-Darling (ad.test {goftest}).

4. Consider the sparse mixture model

$$f(\mu) = (1 - \varepsilon)\delta_0 + \varepsilon\delta_\mu$$

with $\varepsilon = n^{-\beta}$ and $\mu = \sqrt{2r \log n}$.

For each of the settings $\beta = \{0.6; 0.8\}$, $r = \{0.1; 0.4\}$ and $n = \{5000; 50000\}$:

a) simulate the critical values for the Neyman-Pearson test in the sparse mixture;

b) Compare the power of the Neyman-Pearson test to the power of both versions of HC, Bonferroni, Fisher and chi-square.

Summarize the results referring to the theory learned in class.