Theoretical Foundations of the Analysis of Large Data Sets

Laboratory 3, 06.04.2017, Due 20.04.2017

Higher Criticism test and detection of signals is sparse mixtures

1. For $p \in \{5000, 50000, 500000\}$ estimate critical values of the Higher-Criticism test at the significance level $\alpha = 0.05$.

2. Using the settings from Problem 4 in Lab 1 and additionally the setup:

$$\mu_1 = \ldots = \mu_{100} = 2, \quad \mu_{101} = \ldots = \mu_{5000} = 0$$

   compare the power of the following tests: Higher-Criticism, Bonferroni, chi-square, Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D). Summarize the results.

3. For each of the settings $\beta = 0.6, \beta = 0.8$, $r \in \{0.1, 0.2, 0.3, 0.4\}$ and $p \in \{5000, 50000, 500000\}$

   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 the Higher-Criticism, Bonferroni, K-S, A-D and chi-square tests.

   Summarize the results referring to the theory learned in class.

4. Simulate 1000 trajectories of the empirical process $U_p(t)$ with $p = 5000$ and 1000 trajectories of the Brownian bridge $B(t)$. Plot 5 trajectories for each of these processes on the same graph. Based on these simulations estimate the 0.8 quantile of the K-S statistics under the null hypothesis as well as 0.8 quantile of $T = sup_{t\in(0,1)}|B(t)|$. Discuss the results.

Malgorzata Bogdan