University of Wrocław: Data Science

Theoretical Foundations of Large Data Sets, List 5

1. Let $X_{ij}, j = 1, \ldots, m$, be iid random variables from the exponential distribution with the expected value $\mu_i = E(X_{ij})$. Construct a test for testing the hypothesis
$$H_{0i} : \mu_i = 3 \quad vs \quad H_{Ai} : \mu_i > 3.$$

2. For all combinations of parameters

   - $m \in \{20, 100\}$,
   - $n \in \{200, 1000\}$,
   - $\varepsilon \in \{0,01; 0,05; 0,1; 0,2\}$

   consider the following mixture model:

   μ_i are iid random variables from the two-point distribution
   $$\mathrm{P}(\mu_i = 3) = 1 - \varepsilon = 1 - \mathrm{P}(\mu_i = 5,5), \quad i = 1, \ldots, n$$
   and for $j = 1 \ldots m$, X_{ij} are iid $\mathrm{Exp}(\mu_i)$.

   For each $i = 1, \ldots, n$ and $q \in \left\{0,1; 0,1\sqrt{\frac{200}{m}}\right\}$ use test from the Problem 1 with the following multiple testing corrections

   i) Bonferroni procedure at the FWER level q,
   ii) procedure controlling Bayesian FDR at the level q (find the respective critical values by using the command *uniroot()*),
   iii) classical BH procedure at the FDR level q.

   Use at least 1000 replicates to estimate FDR, Power (see, the List 4) and the expected value of the total experiment cost under the assumptions

   i) $c_0 = c_A = 1$ (note that in this case the expected cost is equal to the expected number of misclassifications),
   ii) $c_0 = 2$, $c_A = 1$,
   iii) $c_0 = 1$, $c_A = 2$,

   here c_0 is the cost for the type I error and c_A is the cost for the type II error.

   For the Bonferroni procedure and the BFDR controlling procedure **calculate the exact values** of the Power and of the expected cost.

3. For each of the above values of ε and combinations of c_0 and c_A derive the optimal Bayesian classifier and theoretically calculate its BFDR, Power and the expected value of the corresponding cost function.

4. Assume that you do not know ε and μ. For each replicate of the whole experiment estimate ε and μ using Expectation Maximization algorithm. Evaluate the accuracy of your estimators: draw histograms, calculate the bias, the variance and the mean squared error.

5. Given the above estimates

   i) for each of the above values of ε and combinations of c_0 and c_A construct a plug-in Bayesian classifer *(based on estimated values of ε and μ)* and estimate its Power, FDR and the expected value of the corresponding cost function;

   ii) For each of the above values of ε and q construct the plug-in version of the rule controlling BFDR *(based on estimated values of ε and μ)* and the modifed version of BH *(define $i_0 = \max\left\{i : p_{(i)} \leq \frac{iq}{(1-\varepsilon)n}\right\}$ and reject all hypothesis with p-values smaller or equal then $p_{(i_0)}$).* Estimate Power, FDR and the expected values of the above cost functions for these modifications of the BFDR controlling rule and of BH.

6. Present values of each of the above characteristics (FDR, Power, expected cost) as a function of ε. Compare results of direct and plug-in procedures on the same graph.