Theoretical Foundations of the Analysis of Large Data Sets

Laboratory 2

Needle in Haystack

1. Let $L(X) = \frac{1}{p}\sum_{i=1}^{p}\exp(X_i\mu - \mu^2/2)$ be the statistic of the Neyman-Pearson test for the "needle in haystack" problem, and $\tilde{L}(X) = \frac{1}{p}\sum_{i=1}^{p}\left(\exp(X_i\mu - \mu^2/2)1_{\{X_i<\sqrt{2\log p}\}}\right)$ be its truncated version. For each of the settings $\mu = (1+\epsilon)\sqrt{2\log p}$ with $\epsilon \in \{-0.3, -0.2, -0.1\}$ and $p \in \{5000, 50000, 500000\}$

   a) Estimate $P_{H_0}(L(X) \neq \tilde{L}(X))$.

   b) Calculate the sample mean and the sample variance of $L(X)$ and $\tilde{L}(X)$ (use at least 500 replicates).

   c) Based on at least 500 replicates calculate the maximum of $L(X)$ and $\tilde{L}(X)$.

   d) Report 0.95 quantile of $L(X)$ and $\tilde{L}(X)$.

   How do these quantities change with p ? - comment referring to the theory learned in class. How does $L(X)$ compare to $\tilde{L}(X)$?

2. For $p = 5000$ and $p = 50000$ estimate the critical values of the optimal Neyman-Pearson test for the "needle in haystack" problem against alternatives:

   a) $\mu^{(p)} = 1.2\sqrt{2\log p}$

   b) $\mu^{(p)} = 0.8\sqrt{2\log p}$

   Use the significance level $\alpha = 0.05$. Comment on the results referring to the theory given in class.

3. For $p = 5000$ and $p = 50000$ and $\alpha = 0.05$ compare the power of the above Neyman-Pearson test with the power of the Bonferroni test when

   a) $\mu_1 = 1.2\sqrt{2\log p},\quad \mu_2 = \ldots = \mu_p = 0$

   b) $\mu_1 = 0.8\sqrt{2\log p},\quad \mu_2 = \ldots = \mu_p = 0.$

   Comment on the results referring to the theory given in class.

   Next two problems are for additional points.

4. For $p = 5000$ and $p = 50000$ implement the optimal Neyman-Pearson test against the alternative such that

   a) $||\mu||^2 = (2*p)^{2/5}$

   b) $||\mu||^2 = (2*p)^{3/5}$

   and estimate its critical values. Use the significance level $\alpha = 0.05$. Comment on the results referring to the theory given in class.

5. For $p = 5000$ and $p = 50000$ and $\alpha = 0.05$ compare the power of the above Neyman-Pearson test with the power of the chi-square test when μ is uniformly distributed on the sphere such that

   a) $||\mu||^2 = (2*p)^{2/5}$

   b) $||\mu||^2 = (2*p)^{3/5}$

   Use the significance level $\alpha = 0.05$.

   Comment on the results referring to the theory given in class.

Malgorzata Bogdan