Theoretical Foundations of Large Data Sets

List 3

Paulina Podgórska

1. Estimation of the Type I Error Probability

In this analysis we will estimate the Type I error probability for the modified Higher Criticism Test, HC_{mod}, using the asymptotic critical value for 0.05 significance test, denoted as $c_{crit} = 4.14$. We consider sample sizes $n \in \{5000, 50000\}$. The HC_{mod} statistics is given by the following formula:

$$HC_{mod} = \underset{0<t<1}{max} \sqrt{n} \frac{F_n(t) - t}{\sqrt{t(1-t)q(t)}},$$

where

$$q(t) = loglog \frac{1}{t(1-t)},$$

$$F_n(t) = \frac{\sum_i V_i}{n},$$

$$V_i(t) = \mathbf{1}_{\{p_i \leq t\}}.$$

The estimated probabilities were calculated based on 1000 iterations and are presented in the table below:

n = 5000	n = 50000
0.04	0.052

We observe that the sample size n directly influences the probability of the type I error. For $n = 50000$ we see that the probability is very close to our significance level. This indicates that the modified Higher Criticism Test gains precision and accuracy with increased sample size.

2. Estimation of the critical values

Now, let us estimate critical values of both Higher-Criticism tests at the significance level $\alpha = 0.05$. The Higher Criticism Test statistic is given by:

$$HC^* = \underset{\frac{1}{n}<t<\frac{1}{2}}{max} \sqrt{n} \frac{F_n(t) - t}{\sqrt{t(1-t)}}.$$

The estimated critical values were calculated based on 1000 iterations and are presented in the table below:

HC^*	HC_{mod}
2.997	3.806

We can observe a higher critical value for modified version of HC - it's value is closer to the critical value from task 1. This means that the HC_{mod} test is more conservative in declaring significance than HC^*. This suggest that HC_{mod} might provide better control over the Type I error. This test might be beneficial in situations where it is important to avoid false positives.

3. Power of different tests

In this analysis, we will compare the power of the following tests: Higher Criticism, modified Higher Criticism, Bonferroni, chi-square, Fisher, Kolmogorov-Smirnov and Anderson-Darling in three scenarios:

1. $\mu_1 = 1.2\sqrt{2logn}, \mu_2 = ... = \mu_n = 0,$
2. $\mu_1 = ... = \mu_{100} = 1.02\sqrt{2log\left(\frac{n}{200}\right)}, \mu_{101} = ... = \mu_n = 0,$
3. $\mu_1 = ... = \mu_{1000} = 1.002\sqrt{2log\left(\frac{n}{2000}\right)}, \mu_{1001} = ... = \mu_n = 0,$

for $n = 5000$.

	HC^*	HC_{mod}	Bonferonni	chi-square	Fisher	K-S	A-D
1.	0.053	0.027	0.734	0.086	0.073	0.056	0.067
2.	1.000	0.999	0.978	1.000	1.000	0.684	0.997
3.	1.000	1.000	0.677	1.000	1.000	1.000	1.000

In the case of the needle in haystack problem (1), we can see that, as discussed in the lecture, Bonferroni's method is a clear winner. Its power is higher than 0.7, which is at least 7 times higher than for any other test. In the case of many small effects with the larger size of the needle (2), we observe that all tests did significantly better. But we have 4 top-runners with the maximum power - HC^*, HC_{mod}, chi-square and Fisher. Worst performer is the Kolmogorov-Smirnov test with power under 0.7. In the 3rd case, almost all test performed excellently, except the Bonferroni test, which we know performs worst in the case where the needle is small.

4. Sparse mixture model

Let us consider the sparse mixture model

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

with $\epsilon = n^{-\beta}$ and $\mu = \sqrt{2rlogn}$.

We will perform analysis for each of the settings $\beta = \{0.6, 0.8\}$, $r = \{0.1, 0.4\}$ and $n = \{5000, 50000\}$.

Critical values for the Neyman - Pearson test

We will simulate the critical values for the Neyman-Pearson test in the sparse mixture. The likelihood ratio for this problem is given by

$$L = \prod_{i=1}^{n}\left((1-\epsilon) + \epsilon\exp\{\mu X_i - \frac{\mu^2}{2}\}\right).$$

Table 4: n = 5000

	r_1, β_1	r_1, β_2	r_2, β_1	r_2, β_2
crit. val.	2.937129	1.301748	0.2061482	2.550398

Table 5: n = 50000

	r_1, β_1	r_1, β_2	r_2, β_1	r_2, β_2
crit. val.	2.95907	1.190324	0.0007656	2.614662

We observe that our test becomes more conservative (for both n) when r and β are smaller; in this cases, the critical value nearly reaches 3. Also, when both values are larger, the estimated critical value for test is second highest. The impact of n on the critical value is most noticeable when $r = 0.8$ and $\beta = 0.1$. The critical value for the larger n is over 200 times higher than for $n = 5000$. This indicates a strong influence of the combination of r and β on the critical value, however, the precise nature of this relationship is not immediately apparent.

Comparison of power of different tests

Now, let us compare the power of the Neyman-Pearson test to the power of both versions of HC, Bonferroni, Fisher and chi-square.

Table 6: n = 5000

	N-P	HC^*	HC_{mod}	Bonf.	Fisher	chi-sq
r_1, β_1	0.215	0.100	0.048	0.067	0.116	0.127
r_1, β_2	0.073	0.062	0.036	0.054	0.071	0.067
r_2, β_1	0.971	0.453	0.231	0.538	0.557	0.643
r_2, β_2	0.335	0.081	0.039	0.176	0.097	0.110

Table 7: n = 50000

	N-P	HC^*	HC_{mod}	Bonf.	Fisher	chi-sq
r_1, β_1	0.234	0.087	0.053	0.070	0.117	0.127
r_1, β_2	0.014	0.049	0.038	0.054	0.061	0.062
r_2, β_1	0.996	0.386	0.162	0.693	0.533	0.623
r_2, β_2	0.378	0.069	0.027	0.183	0.065	0.066

Firstly, let us analyze the influence of sample size on the power. It seems that the tests most affected by changes in n are the HC tests - for these, we can notice a significant increase in power for larger sample sizes. On the other hand, we cannot determine a strong influence of n on the performance of the Neyman-Pearson, Bonferroni, Fisher and chi-square tests. Overall, the test that almost consistently performs the best, regardless of the value of r and β is the Neyman-Pearson test. It is the only test that achieves the power very close to one for $r = 0.4$ and $\beta = 0.6$. We also observe that HC_{mod} achieves the worst results for all 8 scenarios.

From the theory learned in class, we know that the power of the test is depended on the threshold effect. There is a threshold curve for r of the form

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & \frac{1}{2} < \beta \leq \frac{3}{4}, \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} \leq \beta \leq 1 \end{cases},$$

such that if $r > \rho^*(\beta)$ we can the N-P test to achieve

$$\mathcal{P}_0(Type\ I\ Error) + \mathcal{P}_1(Type\ II\ Error) \longrightarrow 0$$

and if $r < \rho^*(\beta)$, then for any test

$$lim\ inf\ \mathcal{P}_0(Type\ I\ Error) + \mathcal{P}_1(Type\ II\ Error) \geqslant 1.$$

Let us calculate the threshold for our parameters:

1. $r = 0.1$

- $\beta = 0.6 \rightarrow \rho^*(\beta) = 0.1$
- $\beta = 0.8 \rightarrow \rho^*(\beta) = 0.31$

We can see that for $\beta = 0.6$, the threshold is equal to r. We can notice better performance in our tests in this case in contrast to β_2. As expected, for all tests in the second scenario, the power is very low - below 0.1.

2. $r = 0.4$

- $\beta = 0.6 \rightarrow \rho^*(\beta) = 0.1$
- $\beta = 0.8 \rightarrow \rho^*(\beta) = 0.31$

We can observe, that when r is a lot bigger than the threshold, our tests perform the best. The clear winner is the Neyman - Person test.