Theoretical Foundations of Large Data Sets

List 4

Paulina Podgórska

1. Low dimentional setting

In this task we will analyze a low dimensional setup $n = 20$ in three cases:

1. $\mu_1 = 1.2\sqrt{2 \log n}, \mu_2 = ... = \mu_n = 0$,
2. $\mu_1 = ... = \mu_5 = 1.02\sqrt{2 \log \frac{n}{10}}, \mu_6 = .. = \mu_n = 0$,
3. $\mu_i = \sqrt{2 \log \frac{20}{i}}, i = 1, ...10, \mu_{11} = ... = \mu_n = 0$.

We will compare FWER, FDR and Power of the following procedures:

- Bonferroni,
- Sidak's procedure with $a_n = 1 - (1 - \alpha)^{\frac{1}{n}}$,
- Holm,
- Hochberg,
- Benjamini-Hochberg.

Table 1: Power

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.472	0.475	0.473	0.473	0.478
2.	0.032	0.033	0.032	0.032	0.038
3.	0.108	0.108	0.110	0.110	0.156

Table 2: FWER

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.047	0.051	0.050	0.050	0.084
2.	0.048	0.050	0.051	0.051	0.064
3.	0.034	0.035	0.036	0.036	0.073

Table 3: FDR

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.037	0.040	0.038	0.038	0.057
2.	0.045	0.047	0.047	0.047	0.052
3.	0.018	0.018	0.019	0.019	0.026

From our analysis, it is clear that Bonferroni, Sidak, Holm and Hochberg methods control FWER at the nominal level $\alpha = 0.05$, just as we saw in our lectures. These methods achieve nearly identical values

for all considered statistics in all three cases, although it is notable that Sidak seems to be slightly less conservative. We observe that the Benjamini-Hochberg procedure controls the False Discovery Rate. Moreover, it consistently achieves the highest power in all three scenarios, although with the cost of higher FWER and FDR. Generally, all tests perform the best in the 'needle in the haystack' problem. We also notice that again, as discussed in lecture, in every case FDR is lower than FWER, which balances the rate of false discoveries and the ability to detect true effects.

2. Large dimensional setup

Now, let us experiment with the number of hypothesis. We will see how the number of hypothesis influences the FWER, FDR and power of the procedures. Let us consider a large dimensional setup $n = 5000$ in four cases:

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

Table 4: Power

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.727	0.730	0.727	0.727	0.728
2.	0.033	0.033	0.033	0.033	0.098
3.	0.030	0.030	0.030	0.030	0.083
4.	0.001	0.001	0.001	0.001	0.003

Table 5: FWER

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.039	0.040	0.039	0.039	0.072
2.	0.041	0.044	0.041	0.041	0.406
3.	0.046	0.047	0.046	0.046	0.347
4.	0.047	0.047	0.047	0.047	0.168

Table 6: FDR

	Bonferroni	Sidak	Holm	Hochberg	Benjamini-Hochberg
1.	0.025	0.026	0.025	0.025	0.043
2.	0.012	0.013	0.012	0.012	0.048
3.	0.014	0.014	0.014	0.014	0.045
4.	0.028	0.027	0.028	0.028	0.040

We can see that with increased number of hypothesis, all tests achieve significantly greater power in the needle in the haystack scenario. We also observe that all test fail in the forth scenario, which represents the many small effects case. Beyond this, similar conclusions to the first task can be drawn.

In low-dimensional settings, we might be more interested in controlling FWER.Doing so might be important, because with a smaller number of hypothesis, there's a higher chance that even one false positive might significantly impact the overall results. Also, with smaller number of hypothesis, it is very likely that each test is of high importance.

In high-dimensional settings, controlling FWER might be too conservative, potentially leading to the loss of important information and power, making it difficult to detect any significant results. Controlling FDR might be more interesting, as it allows a small proportion of false positives while increasing the ability to detect true discoveries.

3. Two-step Fisher procedure

Now, let us apply the two-step Fisher procedure using Bonferroni and chi-square test for the first step in the following cases $n \in \{20, 5000\}$ and

1. $\mu_1 = 1.2\sqrt{2 \log n}, \mu_2 = ... = \mu_n = 0$,
2. $\mu_1 = ... = \mu_5 = 1.02\sqrt{2 \log \frac{n}{10}}, \mu_6 = .. = \mu_n = 0$,
3. $\mu_i = \sqrt{2 \log \frac{20}{i}}, i = 1, ...10, \mu_{11} = ... = \mu_n = 0$,
4. $\mu_1 = ... = \mu_{1000} = 1.002\sqrt{2 \log \frac{n}{2000}}, \mu_{1001} = .. = \mu_n = 0$.

Then, in order to analyze the differences we will compare FWER (in the strong sense), FWER (in the weak sense), FDR and Power. Results are shown below.

Table 7: FWER in both senses for n = 20

	$FWER_{strong}$ (Bonf.)	$FWER_{strong}$ (chi-sq.)	$FWER_{weak}$ (Bonf.)	$FWER_{weak}$ (chi-sq.)
1.	0.316	0.296	0.051	0.053
2.	0.131	0.216	0.041	0.034
3.	0.303	0.428	0.047	0.048

Table 8: FWER in both senses for n = 5000

	$FWER_{strong}$ (Bonf.)	$FWER_{strong}$ (chi-sq.)	$FWER_{weak}$ (Bonf.)	$FWER_{weak}$ (chi-sq.)
1.	0.735	0.068	0.053	0.037
2.	0.681	0.165	0.053	0.053
3.	0.087	0.089	0.051	0.051
4.	0.648	1.000	0.045	0.064

Table 9: FDR and Power for n = 20

	FDR (Bonf.)	FDR (chi-sq.)	Power (Bonf.)	Power (chi-sq.)
1.	0.186	0.186	0.492	0.345
2.	0.069	0.109	0.064	0.098
3.	0.074	0.110	0.283	0.378

Table 10: FDR and Power for n = 5000

	FDR (Bonf.)	FDR (chi-sq.)	Power (Bonf.)	Power (chi-sq.)
1.	0.732	0.068	0.735	0.068
2.	0.668	0.162	0.648	0.158
3.	0.086	0.088	0.035	0.038
4.	0.273	0.422	0.177	0.273

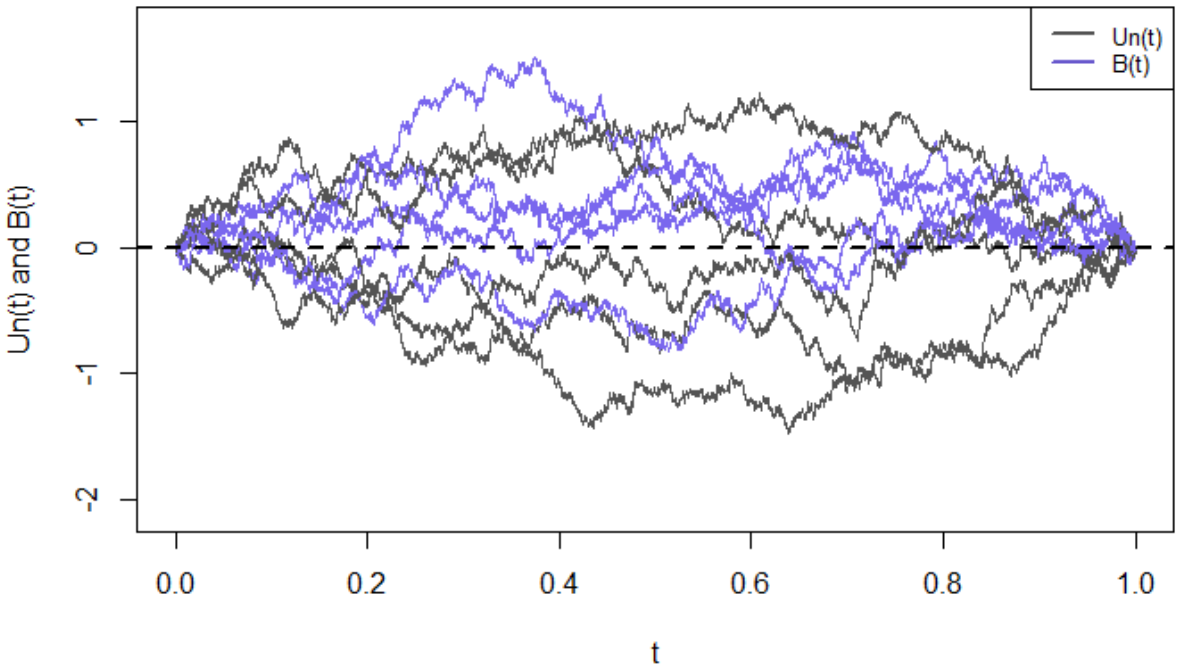
For both values of n, the Bonferroni method has the biggest power in the first scenario - when facing a needle in the haystack problem, but at the cost of the highest $FWER_{strong}$ and FDR. The power in this case, as expected, gets higher with the increase of n. The chi-square test performs the best in the 'many small effects' scenario, achieving the highest power and lowest FDR, yet with the highest $FWER_{strong}$. Both tests show suboptimal performance in the third scenario for the smaller n, and in the second scenario for the bigger n. We can see that the two-step procedure does not control FWER in the strong sense, as the values are significantly higher than the significance level α. However, we observe that in every case, FWER is controlled in a weak sense. We also notice that, as we discussed in the lecture, FDR is smaller than FWER across all scenarios.

4. Simulation of trajectories

In our final section, we will simulate 1000 trajectories of the empirical process

$$U_n(t) = \sqrt{n}(F_n(t) - t), \qquad t \in [0, 1]$$

and 1000 trajectories of the Brownian bridge $B(t), t \in [0, 1]$. Firstly, we will plot 5 trajectories for each of these processes on the same graph.



Next, based on our simulations we will estimate the α quantile of the K-S statistics under the null hypothesis as well as α quantile of $T = sup_{t\in[0,1]}|B(t)|$ for $\alpha = 0.8, 0.9, 0.95$.

Table 11: Estimated quantiles of our statistics

	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 0.95$
K-S	1.655	1.796	1.867
T	1.641	1.717	1.756

The estimated quantiles for both statistics increase with the significance level, aligning with the principle that the higher confidence levels correspond to larger critical values. We know that

$$KS = sup_{t\in[0,1]}\sqrt{n}(\hat{F}_n(t) - t) \to sup_{t\in[0,1]}B(t),$$

which is reflected in the similarity of the quantiles for both K-S and T statistics observed in the table.