University of Wrocław: Data Science

Theoretical Foundations of Large Data Sets, List 4

1. Consider a low dimensional setup: $n = 20$ and

   a) $\mu_1 = 1.2\sqrt{2\log n}, \mu_2 = \ldots = \mu_n = 0$;

   b) $\mu_1 = \ldots = \mu_5 = 1.02\sqrt{2\log\left(\frac{n}{10}\right)}, \mu_6 = \ldots = \mu_n = 0$;

   c) $\mu_i = \sqrt{2\log\left(\frac{20}{i}\right)}, i = 1, \ldots, 10, \mu_{11} = \ldots = \mu_n = 0$.

   Compare FWER, FDR and Power (proportion of identified alternative hypothesis among all alternative hypotheses) of the following procedures:

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

2. Consider a large dimensional setup: $n = 5000$ and

   a) $\mu_1 = 1.2\sqrt{2\log n}, \mu_2 = \ldots = \mu_n = 0$;

   b) $\mu_1 = \ldots = \mu_{100} = 1.02\sqrt{2\log\left(\frac{n}{200}\right)}, \mu_{101} = \ldots = \mu_n = 0$;

   c) $\mu_1 = \ldots = \mu_{100} = \sqrt{2\log\left(\frac{n}{200}\right)}, \mu_{101} = \ldots = \mu_n = 0$;

   d) $\mu_1 = \ldots = \mu_{1000} = 1.002\sqrt{2\log\left(\frac{n}{2000}\right)}, \mu_{1001} = \ldots = \mu_n = 0$.

   Compare FWER, FDR and Power (proportion of identified alternative hypothesis among all alternative hypotheses) of the following procedures:

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

3. Apply two-step Fisher procedure using

    - Bonferroni,
    - chi-square test

    for the first step in the following cases $n \in \{20, 5000\}$ and

    a) $\mu_1 = 1.2\sqrt{2\log n}, \mu_2 = \ldots = \mu_n = 0$;

    b) $\mu_1 = \ldots = \mu_5 = 1.02\sqrt{2\log\left(\frac{n}{10}\right)}, \mu_6 = \ldots = \mu_n = 0$;

    c) $\mu_i = \sqrt{2\log\left(\frac{20}{i}\right)}, i = 1, \ldots, 10, \mu_{11} = \ldots = \mu_n = 0$;

    d) $\mu_1 = \ldots = \mu_{1000} = 1.002\sqrt{2\log\left(\frac{n}{2000}\right)}, \mu_{1001} = \ldots = \mu_n = 0$.

    Compare FWER (in the strong sense), FWER (in the weak sense), FDR and Power (proportion of identified alternative hypothesis among all alternative hypotheses).

4. For $n = 5000$ simulate 1000 trajectories of the empirical process

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

    and 1000 trajectories of the Brownian bridge $B(t), t \in [0, 1]$ (see *BBridge {SDE}*). Plot 5 trajectories for each of these processes on the same graph. Based on these simulations 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 $\alpha = 0,8; 0,9; 0,95$. Discuss the results.