Statistics of Big Data, Fall 21/22

Homework 3

Due date: 6 January 2022

1. **Detecting signal in noise**

   In this exercise we seek to identify some signal hidden in high dimensional noise. The file *covtrain.csv* contains a matrix X of $n = 1000$ observations of dimension $p = 500$. Data were generated from the model Boaz Nadler used in his talk:

   $$x \sim \mathcal{N}(0_p, \sum_{j=1}^{K} \lambda_j v_j v_j^T + I)$$

   with $K << p$ dimension of signal. Note that this also assumes that $v_j \perp v_l$ for $j \neq l$, and that we used $\sigma^2 = 1$ for simplicity.

   Our task is to investigate the eigen decomposition of $X^T X/n$ (or PCA of X) to try and find K, the directions, and relate it to the theory and results presented by Boaz.

   (a) Plot the empirical distribution of the eigenvalues of $X^T X/n$ and compare it to the null distribution under the Marchenko-Pasteur law. What do you conclude about the likely number of identifiable non-null signals in this data?

   (b) Compare the top eigenvalues to the magnitude $(1+\sqrt{p/n})^2$ expected if signal is below the "phase transition" threshold. Are your conclusions similar?

   (c) Now project the matrix X on the 10 top eigenvectors/PCs $\hat{v}_j$ (by multiplying each row by $\hat{v}_j$), and calculate the norms of these vectors. How are they related to the corresponding eigenvalues? Explain it algebraically.

   (d) Next read another independent matrix drawn from the same distribution in *covtest.csv*. Perform the same 10 projections for this matrix and calculate the norms. Explain the results in light of your findings in the previous items.

   (e) (* Extra credit) Next, can we infer on the nature of the vectors v_j?
   Hint: The structure is relatively simple.
   You can use any graphical, intuitive or other method to try and figure it out, but to get credit you then need to find a way to justify your guess in a relevant measurable way.

   Some code hints for this problem are in the file *pca.r*.

2. **Selective inference and multiple testing**

Our general setup: we are either testing m hypotheses or building confidence intervals for m parameters. We may select a subset $S \in \{1, ..., m\}$ of them as "interesting".

(a) State whether each of these claims is true or false and explain **briefly and clearly**:

i. Building confidence intervals at the Bonferroni level $1 - \alpha/m$ guarantees FCR control at level α for any subset S.

ii. If we choose a set of rejected hypotheses by the BH procedure at level α, obtaining R rejections, and then build confidence intervals at level $1 - \alpha \cdot R/m$, then the FCR is also controlled at level α.

iii. If we decide to select all m hypotheses as "interesting", then selective inference (i.e., controlling FCR) is reduced to inference "on average", meaning we are controlling the expected percentage of errors of our m hypotheses.

(b) Consider the Science paper by Zeggini et al. referenced in slide 28 of Yoav's deck ("Selection by table") (link to the original paper can be found the class homepage).

i. Assume we calculate FCR-corrected confidence intervals for the second to last column of the table. Considering the results from the first part of this problem, and the p values in the table, explain why these FCR-corrected intervals are not expected to cross below 1.

ii. Assume now that we were to take a different approach, collect all the SNPs that were significant in *any* of the participating studies, and declare all of them "selected" (their number can be much bigger than the ten on slide 11), and then build FCR-corrected CI's for them at FCR level α, based on the entire meta analysis (like the last two columns of the table). Do you expect that some of these intervals will cross 1? Explain. What percentage of non-coverage do you expect over these selected? Specifically, do you expect this percentage to be about α, smaller than that, or larger than that? Explain.

(c) Assume now that instead of selecting interesting results, we can order our hypotheses a-priori from "the most important" to the least important. We care most about the first hypothesis and least about the last. The following procedure is known as hierarchical testing:

- Test the most important null hypothesis at level α. If not rejected, stop and don't consider the other hypotheses.
- If rejected, continue to the second null hypothesis and test it at level α. If this second hypothesis not rejected, stop.
- Continue until a non-rejected null. Then stop.

i. Does this procedure control FWER at level 0.05? Prove your answer (a full formal proof is not required, but a clear and correct argument is required).

ii. Given a large number m of hypotheses, consider Bonferroni, α-spending, and this hierarchical approach, all at level α. Can you predict which one would make more rejections? Why yes or why not?