Név: **Kovács Sebestyén**
NEPTUN-kód: **FP959I**
Témavezető: **Backhausz Ágnes Mariann**

Math project 1
Random matrices, perturbations and their applications in statistics
(Véletlen mátrixok, perturbációk és statisztikai alkalmazásaik)

Covariance matrices play an important role in multivariate statistical analysis, so it can be useful to examine random matrices that arise as covariance matrices of data. This semester we considered a perturbation problem in random matrix theory. More precisely, we worked with such problems that are motivated by analysing big data and matrices. These matrices can be examined with their singular vectors and singular values. We are often curious about how the singular vectors and singular values of our matrix will change in case we add a random noise to our matrix. These vectors and values can characterize this new data set well. Generally we assume that we have a low rank matrix. To model the randomness of the observations, we add random noise to this low-rank matrix. The main aim of this project is to review the scientific literature of this area, to get acquinted with statistic applications and to make some computer simulations related to this topic, which helps us to understand cases where analitic results are not available. Firstly, we have to define the singular values and singular vectors of matrices.

Definition. Let σ_1 be the first singular value of matrix A and and let denote the first singular vector of matrix A by v_1, if

$$\sigma_1 = \max_{|v|=1} |Av| \qquad \text{and} \qquad v_1 = \operatorname{argmax}_{|v|=1} |Av|.$$

By induction, let σ_i be the i-th singular value of matrix A (for $i = 2 \dots r$) and let denote the i-th singular vector of matrix A by v_i, if

$$\sigma_i = \max_{v:|v|=1, v\perp v_1, \dots v_{i-1}} |Av| \qquad \text{and} \qquad v_i = \operatorname{argmax}_{|v|=1, v\perp v_1, v_2 \dots v_{i-1}} |Av|.$$

In this situation v_r is the last singular vector of matrix A, if $\sigma_r > 0$ and

$$\max_{v\perp v_1, v_2, \dots v_r} |Av| = 0.$$

This semester I came to know the most important features of singular vectors and values of matrices, by reading the scientific literature of [2]. For example, we could see that the subspace spanned by the first k singular vectors of matrix A $(1 \leq k \leq r)$ is the best-fit k-dimensional subspace for the row vectors of matrix A.

The main aim of Wang's article (cf. [1]) is to give better estimations to the Weil bounds (these are in the 5th and 7th theorem of his article). I read this article and came to know the most important results and methods, too. Here we assume that the random noise E is a Bernoulli matrix, so

$$E = [E]_{i,j}, \qquad P(E_{i,j} = 1) := P(E_{i,j} = -1) := 0.5$$

with independent coordinates. We have proven the Corollary 8 from the Weil bounds. The main result of Wang's article is his ninth theorem, which is contained in the next

Theorem. (cf. [1]) Assume that E is a Bernoulli matrix and $A, E \in \mathbb{R}^{n \times n}$, furthermore let the rank of A be denoted by r. For every $\varepsilon > 0$ there exist constants $C, \delta_0 > 0$ such that if

$$\delta \geq \delta_0 \quad \text{and} \quad \sigma_1 \geq \max\{n, \sqrt{n} \cdot \delta\}$$

then with a probability at least $1 - \varepsilon$ the inequality

$$\sin(< (v_1, v_1')) \leq C \cdot \frac{\sqrt{r}}{\delta}$$

fulfils. Here v_1 is the first singular vector of matrix A and v_1' is the first singular vector of $A + E$ (the new matrix).

In order to prove these results, we have to examine the concentration of a Bernoulli matrix. We worked a lot with Lemma 35 and 36. Lemma 35 was proven with the help of Hoeffding's inequality but Lemma 36 was written down without any proof. After understanding the proof of Lemma 35, I could give a proof to lemma 36, too. These lemmata assert that the quadratic form of a Bernoulli matrix can be large only with low probability. By these lemmata we had to assume that the absolute values of the components of Bernoulli matrices are upper-bounded. We could see that this assumption is really important. By constructing appropriate sequences of matrices I could show that in some cases we can only get a trivial upper bound on the the quadratic form of E.

In the future we will examine these results more deeply. We will make some simulations about random matrices and their singular vectors and values, in the cases where no analytic results are available. On the other hand, we would like to understand the perturbed random matrices and their statistic applications more profoundly. We want to see the connections between this topic and Principal Component Analysis, too.

Last but not least I would like to generalize the results that were proven until now, for example to examine more complicated random noise matrices than Bernoulli matrices, partly by using scientific literature and partly on my own.

References

[1] O'Rourke, S.; Vu, V.; Wang, K.: *Random perturbation of low rank matrices: Improving classical bounds*, Linear Algebra and its Applications **540** (2016), 26–59.

[2] Blum, A; Hopcroft, J.; Kannan, R.: *Foundations of Data Science*, Cambridge University Press, 2020.
(`https://doi.org/10.1017/9781108755528`, cf. `https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/book-chapter-4.pdf`)