Random matrices, perturbations and their applications in statistics

Sebestyén Kovács

Supervisor: Ágnes Backhausz

Eötvös Loránd University

Monday 8th January, 2024

Content

1. Introduction, motivations
2. Singular vectors and singular values of matrices
3. Most important results, theorems
4. Future plans, references

Introduction, motivations

- Importance of examining data and covariance.
- How can we describe matrices of data effectively?
- Perturbation problem: how can we model noisy observation.

- First singular value and vector of A.
- $\sigma_1 := \max_{|v|=1} |Av| \quad \text{and} \quad v_1 := \operatorname{argmax}_{|v|=1} |Av|.$
- 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

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

- Features of singular vectors and values.

Quadratic form

Interesting theorem (O'Rourke, Vu, Wang, 2016): If $E = [E]_{i,j}$ is a squared real symmetric matrix with independent entries and zero mean in and above the main diagonal and there exists a $K \geq 1$ with

$$P(E_{i,j} < K) = 1 \qquad \text{(for every } i,j)$$

then for every normalized vectors u, v ($|u| = |v| = 1$) and every $t > 0$ we have

$$P((Eu)^T v \geq t) \leq 2 \exp\left(-\frac{t^2}{K^2}\right).$$

The proof of this statement was not elaborated in the paper, I worked out the details and studied some generalizations.

- The matrix E is called Bernoulli matrix if

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

with independent coordinates.

- We have seen the main theorem of O'Rourke, Vu, Wan with Bernoulli matrix: if A is data matrix with (low) rank r and E is a random Bernoulli matrix, then 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

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

fulfils.

Future plans

- Making simulations about perturbed random matrices.
- Understanding the perturbed random matrices and their statistical applications.
- Seeing the connections between these and Principal Component Analysis (PCA).

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.

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)

Thank you for your attention!