Assignment 3 - Numerical Linear Algebra

Arthur Rabello Oliveira[1], Henrique Coelho Beltrão[2]

17/06/2025

Abstract

coming soon

Contents

[1]Escola de Matemática Aplicada, Fundação Getúlio Vargas (FGV/EMAp), email: arthur.oliveira.1@fgv.edu.br
[2]Escola de Matemática Aplicada, Fundação Getúlio Vargas (FGV/EMAp), email: henrique.beltrao@fgv.edu.br

1. Introduction

2. Norm Distribution (a)

2.1. The Chi-Square Distribution

Here we construct a theoretical basis for our analysis of the histograms shown in Section 2.2

When we generate a matrix $A \in \mathbb{R}^{m \times n}$, with $A_{ij} \sim N(0,1)$ independent, each column c_i is a gaussian vector in $\mathbb{R}^m$. if

$$x = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \in \mathbb{R}^m \tag{1}$$

Is a column, then:

$$\begin{aligned} V = \|x\|_2 = \sqrt{\sum_{i=1}^{m} X_i^2} \\ V^2 = \sum_{i=1}^{m} X_i^2 \end{aligned} \tag{2}$$

Is of our interest. The expected value and variance are:

$$\begin{aligned} \mathbb{E}[V^2] = \mathbb{E}\left[\sum_{i=1}^{m} X_i^2\right] = \sum_{i=1}^{m} \mathbb{E}[X_i^2] = m \\ \mathrm{Var}(V^2) = \mathrm{Var}\left(\sum_{i=1}^{m} X_i^2\right) = \sum_{i=1}^{m} \mathrm{Var}(X_i^2) = 2m \end{aligned} \tag{3}$$

But we know that if $X_i \sim N(0,1)$ are independent:

$$\sum_{i=1}^{m} X_i^2 \sim \chi_m^2 \tag{4}$$

where χ_m is the chi-squared distribution with m degrees of freedom, better discussed in Section 2.1.

Taking the square root on eq. (4), we have:

$$V = \|x\|_2 = \sqrt{\sum_{i=1}^{m} X_i^2} \sim \sqrt{\chi_m^2} \sim \chi_m \tag{5}$$

The 2-norm of a vector x is distributed as a chi distribution with m degrees of freedom, in order to understand the distribution for many values of m, we can calculate the expected value and variance of this distribution as a function of m. The PDF of the chi distribution (with m degrees of freedom) is:

$$f_V(\varphi) = \frac{1}{2^{\frac{m}{2}-1} \cdot \Gamma\left(\frac{m}{2}\right)} \varphi^{m-1} e^{-\frac{\varphi^2}{2}} \tag{6}$$

So from this, the expected value is:

$$\mathbb{E}(V) = \sqrt{2} \cdot \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \tag{7}$$

And from this, the variance:

$$\mathrm{Var}(V) = m - \left(\sqrt{2} \cdot \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\right)^2 \tag{8}$$

The Stirling Approximation provides a good approximation for the expected value and variance:

$$\mathbb{E}(V) \approx \sqrt{m} \cdot \left(1 - \frac{1}{4m} + O\left(\frac{1}{m^2}\right)\right) \tag{9}$$

$$\mathrm{Var}(V) \approx \frac{1}{2} + O\left(\frac{1}{m}\right) \tag{10}$$

2.2. Histograms

The first cell of this notebook has as expected output, with input being matrices with fixed $n = 1000$ and $m \in \{10, 20, 100, 200, 1000, 2000\}$, the following plots:

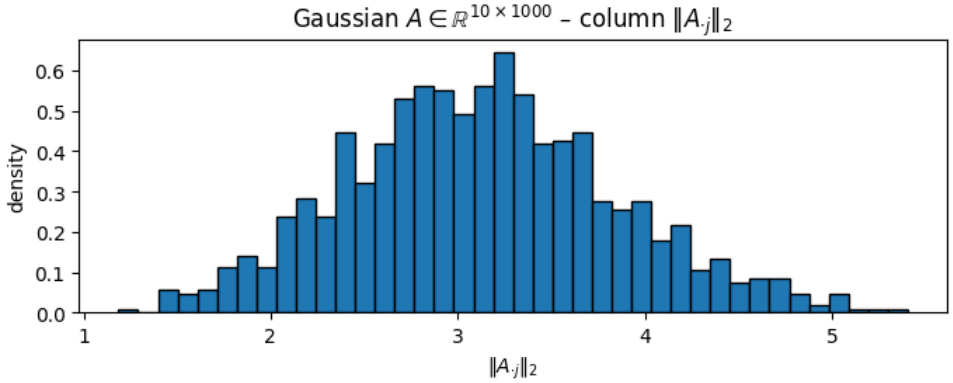


Figure 1: 10×1000 gaussian matrix

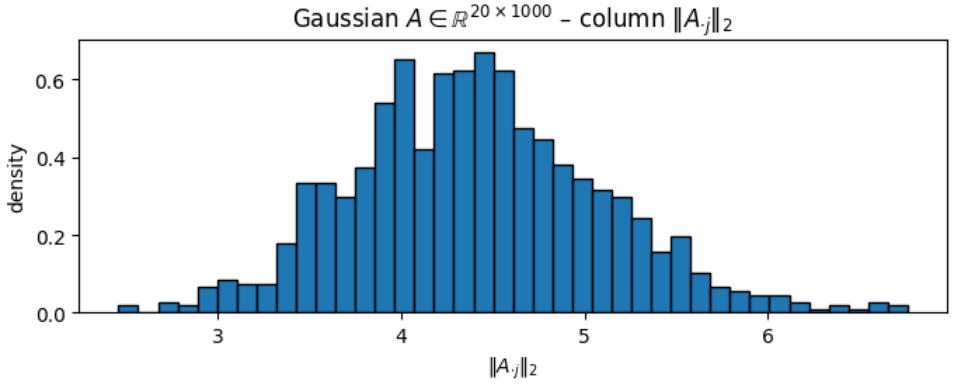


Figure 2: 20×1000 gaussian matrix

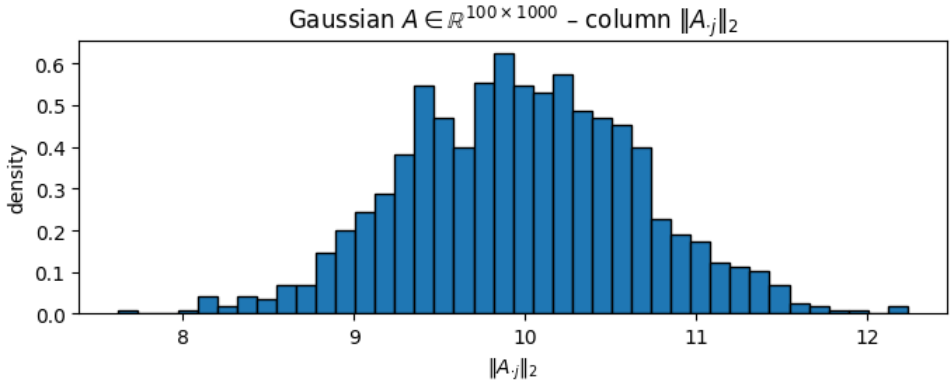


Figure 3: 100×1000 gaussian matrix

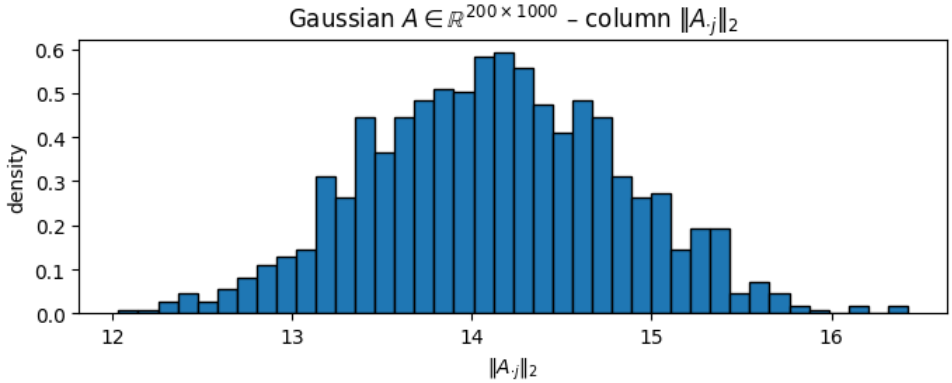


Figure 4: 200×1000 gaussian matrix

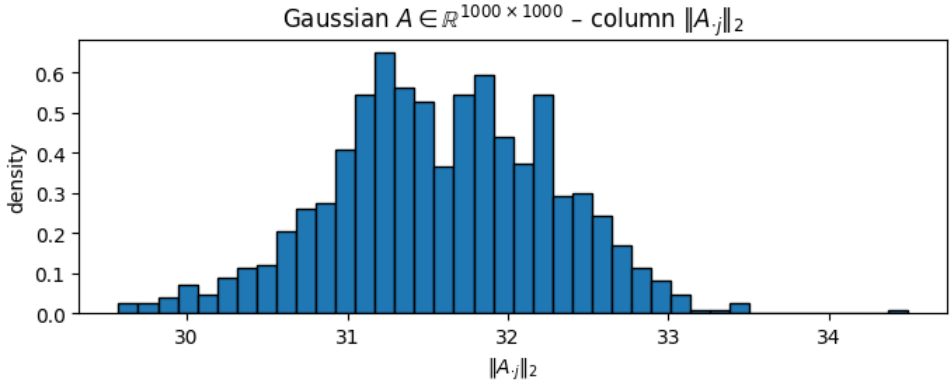


Figure 5: 1000×1000 gaussian matrix

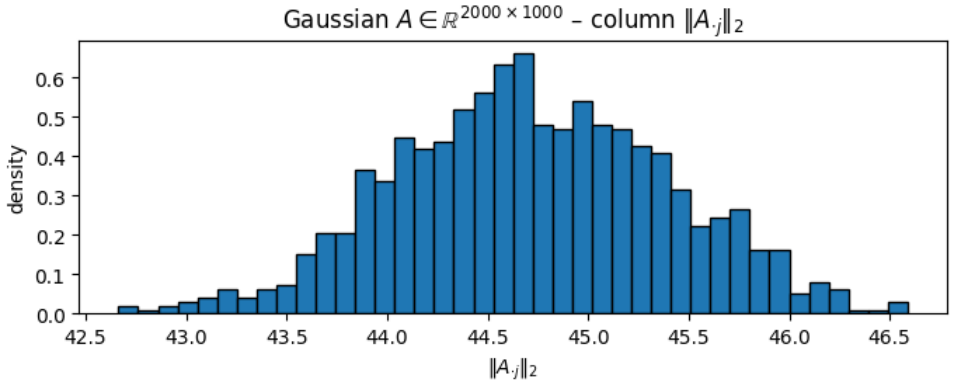


Figure 6: 2000×1000 gaussian matrix

m	**approximate μ_m (theory)**	**$[\mu \pm 3\sigma]$ (theory)**	**observed spike**	**visual range**
10	3.08	$1.0 - 5.18$	≈ 3.1	$1.2 - 5.1$
20	4.42	$2.3 - 6.52$	$\approx 4.3 - 4.4$	$2.9 - 6.4$
100	9.98	$7.9 - 12.1$	$\approx 9.9 - 10.0$	$7.8 - 12.0$
200	14.12	$12.0 - 16.2$	≈ 14.1	$12.2 - 16.2$
1000	31.61	$29.5 - 33.7$	$\approx 31.7 - 32.0$	$29.9 - 33.3$
2000	44.72	$42.6 - 46.8$	$\approx 44.5 - 45.0$	$42.8 - 46.6$

This table illustrates the expected value μ_m and the range $[\mu - 3\sigma, \mu + 3\sigma]$ for Figure 1 to Figure 6.

So apparently as m grows, the size of the gaussian vectors rapidly converge to $\sqrt{m}$, with small errors.

3. Inner Products (b)

Here we construct a theoretical basis for our analysis of the inner products shown in Section 3.1.

When we generate a matrix $A \in \mathbb{R}^{m \times n}$, with $A_{ij} \sim N(0,1)$ independent, each column c_i is a gaussian vector in $\mathbb{R}^m$. If The inner product of two gaussian vectors $x = (X_1, ..., X_n), y = (Y_1, ..., Y_n)$ is:

$$Z = \langle x, y \rangle = \sum_{i=1}^{m} X_i Y_i \tag{11}$$

With $X, Y \sim N(0,1)$. Since X_i, Y_j are independent, we have:

$$\begin{aligned} \mathbb{E}[Z] &= \sum_{i=1}^{m} \mathbb{E}[X_i Y_i] = \sum_{i=1}^{m} \mathbb{E}[X_i]\mathbb{E}[Y_i] = 0 \\ \mathrm{Var}(Z) &= \sum_{i=1}^{m} \mathrm{Var}(X_i Y_i) = \sum_{i=1}^{m} \mathbb{E}[X_i^2]\mathbb{E}[Y_i^2] = \sum_{i=1}^{m} 1 = m \end{aligned} \tag{12}$$

If $W = X_i Y_i$, we have:

$$M_W(\varphi) = \mathbb{E}[e^{\varphi W}] = \frac{1}{\sqrt{1 - \varphi^2}}, |\varphi| < 1 \tag{13}$$

Over all $W_i = X_i Y_i$:

$$M_Z(\varphi) = \mathbb{E}[e^{\varphi Z}] = (M_W(\varphi))^m = \left(\frac{1}{\sqrt{1-\varphi^2}}\right)^m = (1-\varphi^2)^{-\frac{m}{2}}, |\varphi| < 1 \tag{14}$$

And magically:

$$M_{\frac{Z}{\sqrt{m}}}(\varphi) = \left(1 - \frac{\varphi^2}{m}\right)^{-\frac{m}{2}} \Rightarrow \lim_{m\to\infty} M_{\frac{Z}{\sqrt{m}}}(\varphi) = e^{\frac{\varphi^2}{2}} \tag{15}$$

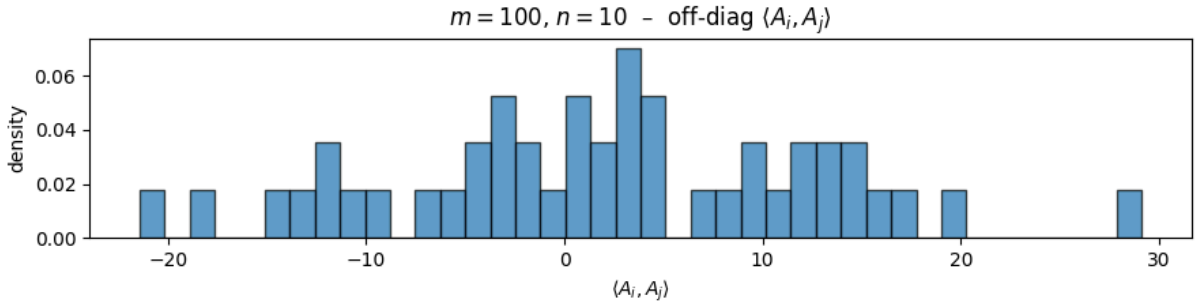
Precisely the moment generating function of a standard normal distribution, so as $m \to \infty$:

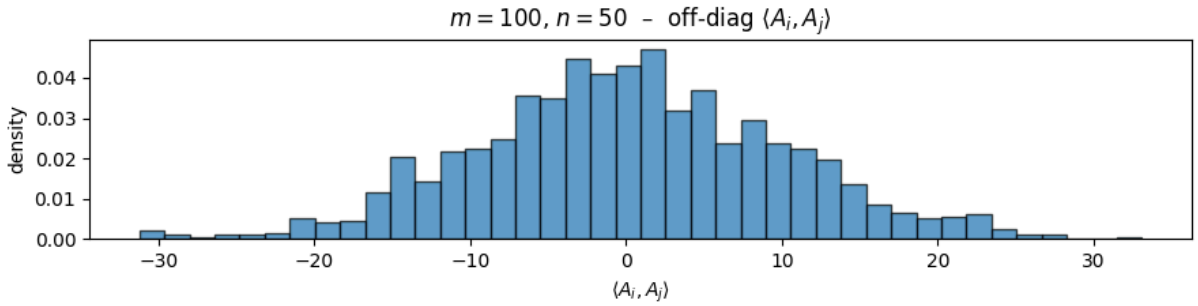
$$\frac{Z}{\sqrt{m}} \sim N(0,1) \tag{16}$$

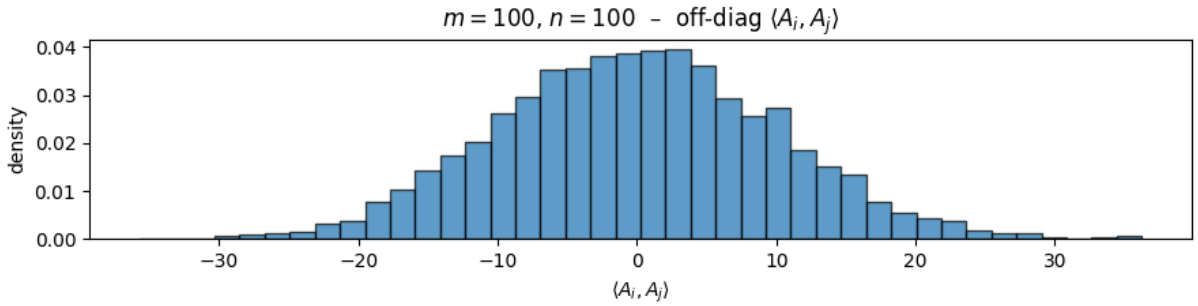
With a fixed $m = 100$, when $n \to \infty$ we can see the distribution approaching $N(0,1)$, as shown in Section 3.1

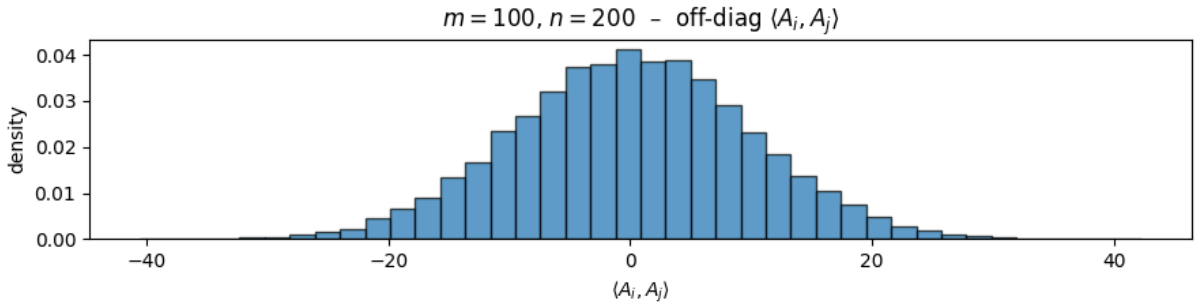
3.1. Histograms

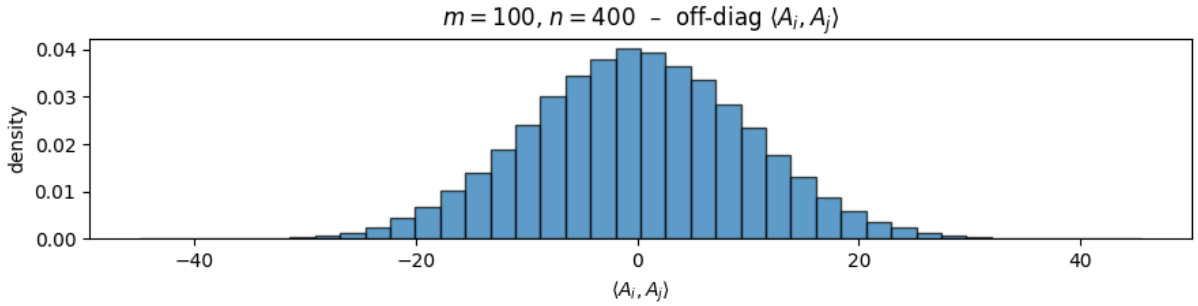
The following plots are an expected output for the second cell of this notebook, with input $m = 100, n \in \{10, 20, 30, 40, 50, 60, ..., 1000\}$:

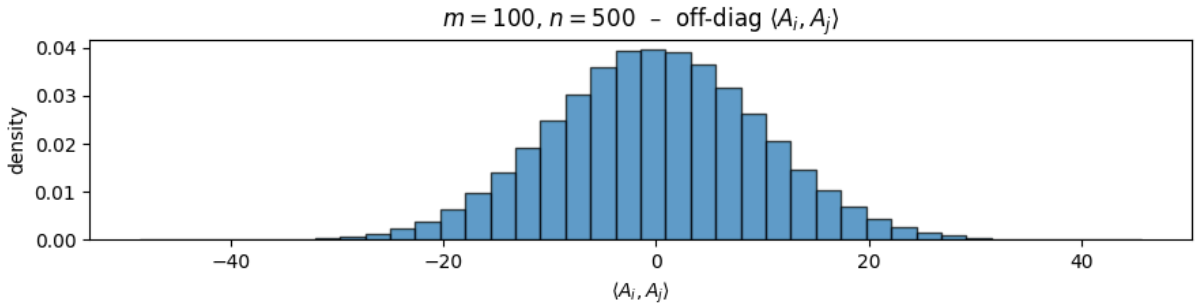


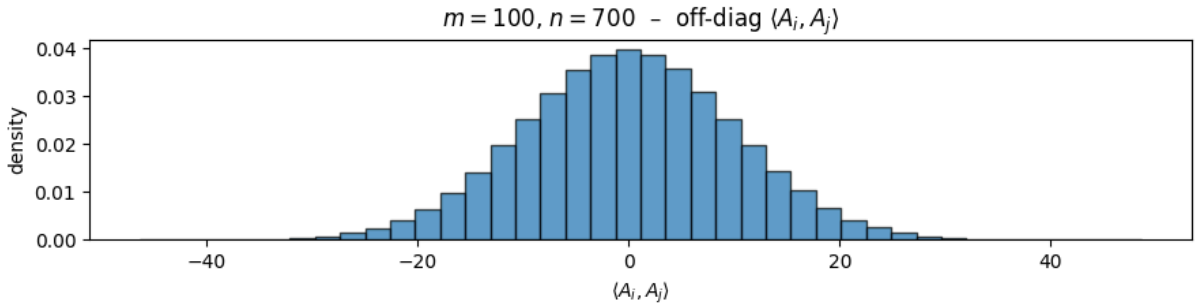












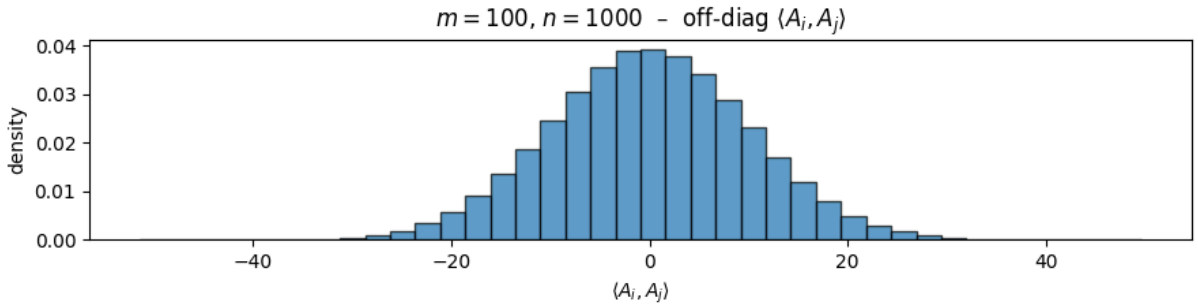


Figure 7 → Figure 14 shows that the distribution indeed approaches $N(0, 1)$

4. The Maximum Distribution (c)

In this section, we analyze the distribution of the maximum non-orthogonality between columns of a Gaussian matrix. This non-orthogonality is quantified by the maximum absolute value of the cosine similarity between any two distinct column vectors. Specifically, for a matrix $A \in \mathbb{R}^{m \times n}$, we study the distribution of the random variable:

$$M = \max_{i \neq j} \frac{|\langle A_i, A_j \rangle|}{\|A_i\| \|A_j\|} \tag{17}$$

Our experiment generates K independent realizations of this value, $M_1, M_2, ..., M_K$, by creating K different Gaussian matrices of size $m = 100, n = 300$. The histograms later shown in Section 4.2 display the empirical probability density function of this collection of maxima.

4.1. Theoretical Framework and the Gumbel Distribution

Let $C_{ij} = \frac{\langle A_i, A_j \rangle}{\|A_i\| \|A_j\|}$. For a given matrix A, we are examining the maximum of $N = \frac{n(n-1)}{2}$ random variables, $\left\{|C_{ij}|\right\}_{1 \leq i < j \leq n}$. For $m = 100$ and $n = 300$, this is the maximum of $N = 44850$ values.

We are interested in the maximum of $\left\{|C_{ij}|\right\}$. As established in previous sections:

- From part (a) (Section 2), for large m, $\|A_i\|$ concentrates around $\sqrt{m}$.
- From part (b) (Section 3), $Z_{ij} = \langle A_i, A_j \rangle$ is approximately $N(0, m)$.

Let's first characterize the distribution of a single variable C_{ij}.

$$C_{ij} = \frac{Z_{ij}}{\|A_i\| \|A_j\|} \approx \frac{N(0, m)}{\sqrt{m} \cdot \sqrt{m}} = \frac{N(0, m)}{m} \tag{18}$$

If a random variable $X \sim N(0, \sigma^2)$, then $\frac{X}{c} \sim N\left(0, \frac{\sigma^2}{c^2}\right)$. Thus:

$$C_{ij} \approx N\left(0, \frac{m}{m^2}\right) = N\left(0, \frac{1}{m}\right) \tag{19}$$

So, the individual correlation values are approximately drawn from a normal distribution with mean 0 and a small variance of $\frac{1}{m}$.

Our analysis, however, concerns the variable $M = \max_{i \neq j} |C_{ij}|$. The parent distribution is therefore not $N\left(0, \frac{1}{m}\right)$, but rather its absolute value, $|N\left(0, \frac{1}{m}\right)|$. This is known as a **folded normal distribution**.

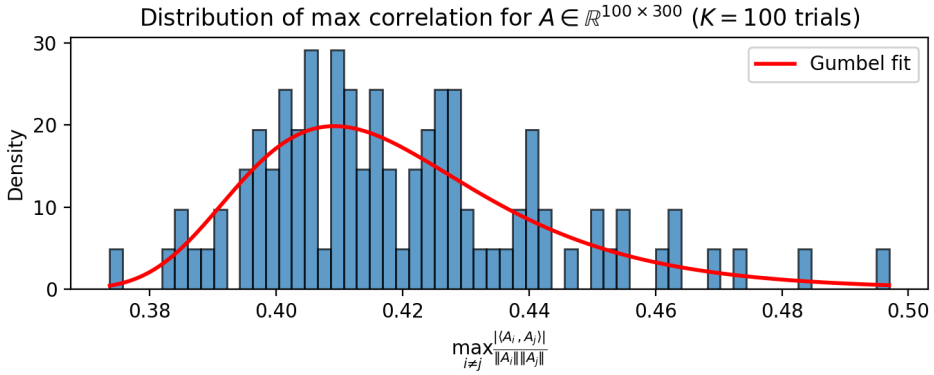
The tail of the folded normal distribution behaves identically to the tail of the underlying normal distribution. According to **Extreme Value Theory**, the limiting distribution for the maximum of many i.i.d. variables from a parent distribution with an exponential tail (like the normal distribution) is the **Gumbel distribution**.

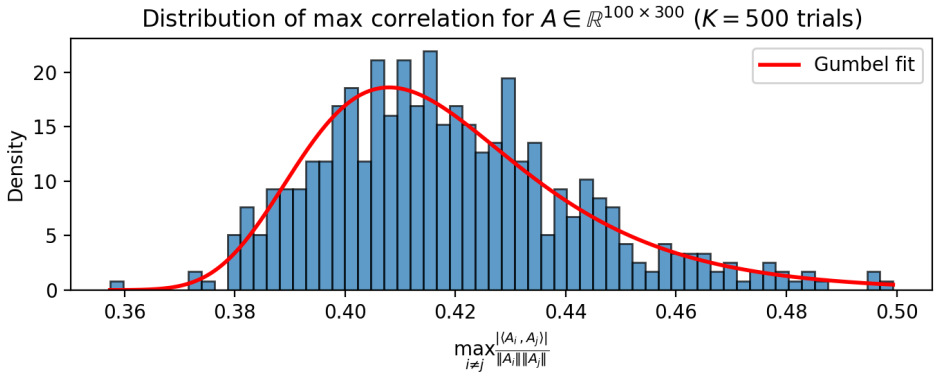
The probability density function (PDF) for the Gumbel distribution is given by:

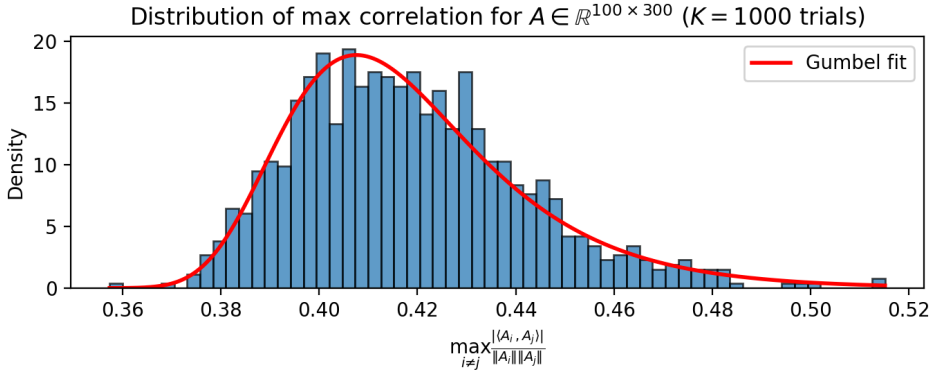
$$\begin{aligned} f(x; \mu, \beta) &= \frac{1}{\beta} e^{-(z + e^{-z})} \\ z &= \frac{x - \mu}{\beta} \end{aligned} \tag{20}$$

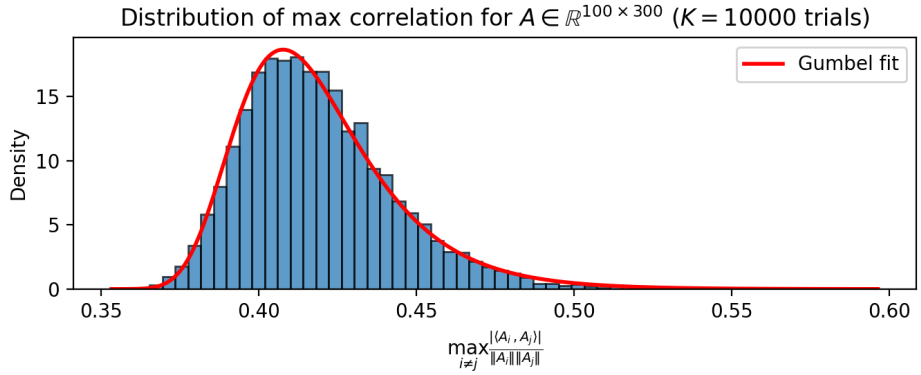
where μ is the mode of the distribution (location parameter) and β is the scale parameter (proportional to the standard deviation).

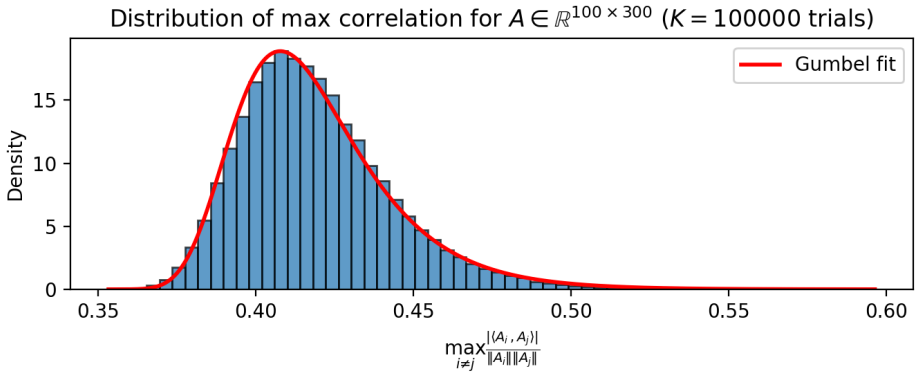
4.2. Analysis of the Histograms











The histograms generated, especially for large K (e.g., $K = 10000$ and $K = 100000$ as shown in Figure 18 and Figure 19, respectively), exhibit the distinct features of a Gumbel distribution:

- A single peak (unimodal).
- Asymmetry with a more extended tail on the right side.

As we can observe, growing K *(number of trials)* leads to a smoother plot and a clearer shape of the distribution, which aligns with the theoretical expectations of the Gumbel distribution.

The observed mode of the distribution is around 0.42, which is consistent with theoretical predictions. The location parameter μ can be approximated by:

$$\mu \approx \sqrt{\frac{2\ln(N)}{m}} = \sqrt{\frac{2\ln\left(\frac{n(n-1)}{2}\right)}{m}} \tag{21}$$

This formula arises from the well-known approximation for the expected maximum of N standard normal variables ($\sqrt{2\ln N}$), applied to our standardized variables $\left\{|\sqrt{m}C_{ij}|\right\}$.

For $m = 100$ and $n = 300$, we have $N = 44850$:

$$\mu \approx \sqrt{\frac{2\ln(44850)}{100}} \approx \sqrt{\frac{2 \cdot 10.71}{100}} = \sqrt{0.214} \approx 0.462 \tag{22}$$

This theoretical approximation gives a value in the general vicinity of the observed peak (around 0.42). The discrepancy arises, and will be more evident when discussing convergence at Section 5.2, because the variables $\left\{C_{ij}\right\}$ are not perfectly independent (for instance, $C_{1,2}$ and $C_{1,3}$ both depend on column A_1) and their distribution is only approximately normal. Nonetheless, this formula correctly shows that the peak of the distribution is determined by the dimensions m and n.

In conclusion, the observed distribution is a **Gumbel distribution**. This arises because we are plotting the maximum of a very large number of approximately independent, normally-distributed random variables (the cosine similarities).

5. Complexity

5.1. Algorithm Complexity and Runtime

The complexity of the algorithm is determined by the main operations within each of the K iterations.

The process begins by generating a Gaussian matrix of size $m \times n$, which has a time complexity of $O(mn)$. We then calculate the L2-norm for n columns of length m using `norms = np.linalg.norm(A, axis=0)`, an operation with $O(mn)$ complexity. The most computationally expensive step is the calculation of the Gram Matrix via `G = A.T @ A`. This matrix multiplication of an n x m matrix with

an m x n matrix has a complexity of $O(mn^2)$. Subsequent operations, including the outer product ($O(n^2)$), element-wise division ($O(n^2)$), and maximum extraction ($O(n^2)$), are less expensive.

The total complexity for a single iteration is the sum of these steps, dominated by the Gram matrix calculation:

$$O(\text{One Iteration}) = O(mn) + O(mn) + O(mn^2) + O(n^2) = O(mn^2) \tag{23}$$

Therefore, for K iterations, the total complexity of our algorithm is $O(Kmn^2)$. This implies that the runtime should scale linearly with K and m, and quadratically with n. We can verify this empirically.

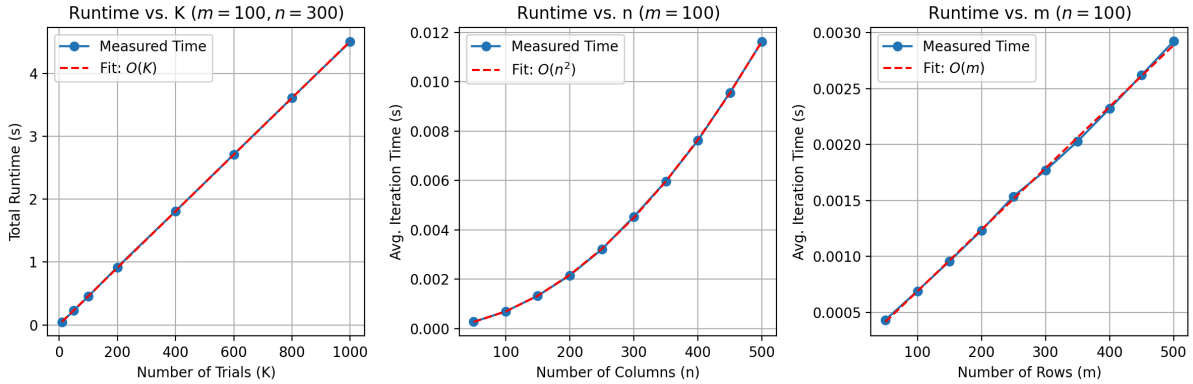


Figure 20: Runtime for varying K, n, and m

As predicted, Figure 20 confirms our theoretical model. The plots show that the runtime scales linearly with K and m, and quadratically with n. This empirically verifies the algorithm's overall complexity.

5.2. Algorithm Convergence and Choosing an Appropriate K

The question *"What value of K is good for a good estimate of the expected maximum?"* is about statistical convergence, not computational performance. K represents our sample size, which should be large enough to ensure our statistics (like the mean and the histogram's shape) are stable and reliable.

To compute these maxima over many iterations (up to $K = 10^5$), we used Multiprocessing. A simple way to visualize convergence is to plot the running average of the maximum correlation as K increases. We expect this average to fluctuate for small K and converge to a stable value as K grows.

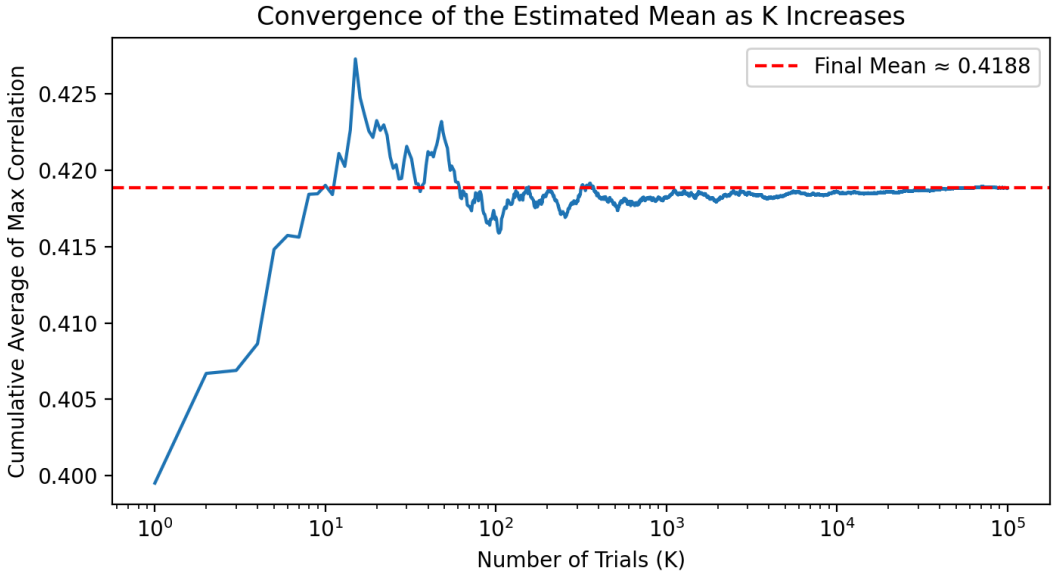


Figure 21: Convergence of eq. (17) as K grows

From Figure 21, we can observe that:

- For $K < 100$, the estimate is very noisy and unreliable.
- For $100 \leq K < 1000$, the estimate begins to stabilize, despite some minor yet visible fluctuations.
- For $K \geq 1000$, our estimate becomes very stable and converges smoothly to ≈ 0.42.

A K value in the range of 10^3 to 10^4 is a good choice for this problem, providing a balance between a reliable statistical estimate and computational cost. As seen in the plot, there is very little difference in the mean between $K = 10^4$ and $K = 10^5$, yet the computational cost is ten times greater, indicating that choosing $K = 10^5$ is likely unnecessary for the purpose of estimating the mean.

6. Another Maximum Distribution

7. Conclusion

Bibliography