Graduação em Matemática Aplicada ***FGV-EMAP***
Álgebra Linear Numérica
Prof: Bernardo Freitas Paulo da Costa
Aluno: Artur de Oliveira Teles

Projeto 4
Álgebra Linear Numérica

Introdução:

Este projeto foi feito usando novamente Rust e Linux, óbvio, como no primeiro projeto, pois cansei de usar python. Desta vez estou usando um biblioteca (crate) especializada em álgebra linear chamada **nalgebra**, sendo que para fazer plots estou usando **gnuplot**. Como o professor disse que o último projeto feito em Rust ficou muito pesado, o que de fato é verdade graças ao gerenciador de pacotes da linguagem chamado Cargo, estarei disponibilizando no eclass apenas o relatório do projeto e os arquivos de códigos estarão disponíveis aqui (ou https://github.com/nyoxon/fgv_aln_projeto4 caso o hiperlink não funcione), isto é, num repositório no github, o que permite a visualização dos códigos sem ter que rodá-los. As instruções para rodar os códigos estão no próprio repositório.
Em relação a eu estar fazendo o projeto sozinho: não acho que há algum problema substancial em o professor/monitor ter que corrigir, dado que todos os outros alunos façam duplas, um trabalho adicional, com peso 1, sendo que o professor poderia não ter passado esse projeto e ter colocado peso 4 para a A2, algo que eu preferia. Espero que lendo isso o professor/monitor não desconte nota 😀.

Questão 1:

a)

Teoria:

Seja c_i a i-coluna de uma matriz gaussiana $A \in \mathbb{R}^{m \times n}$, então:

$$||c_i||_2 = \sqrt{\sum_j^m X_{ji}^2} \sim \chi(m)$$

onde $\chi(m)$ é a distribuição chi com m graus de liberdade. Tal distribuição tem valor esperado e variância:

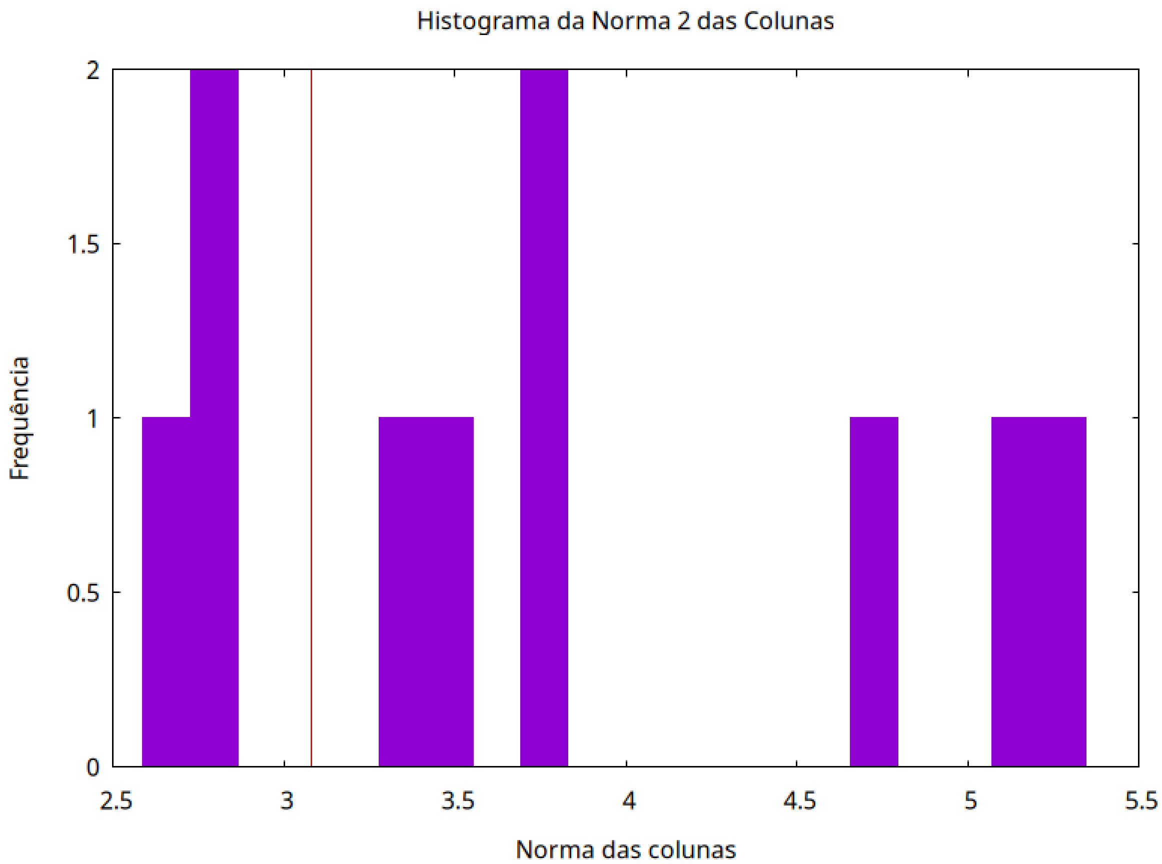
$$\mu = \sqrt{2}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \cong \sqrt{m - 0.5}$$

$$\sigma^2 = n - \mu^2$$

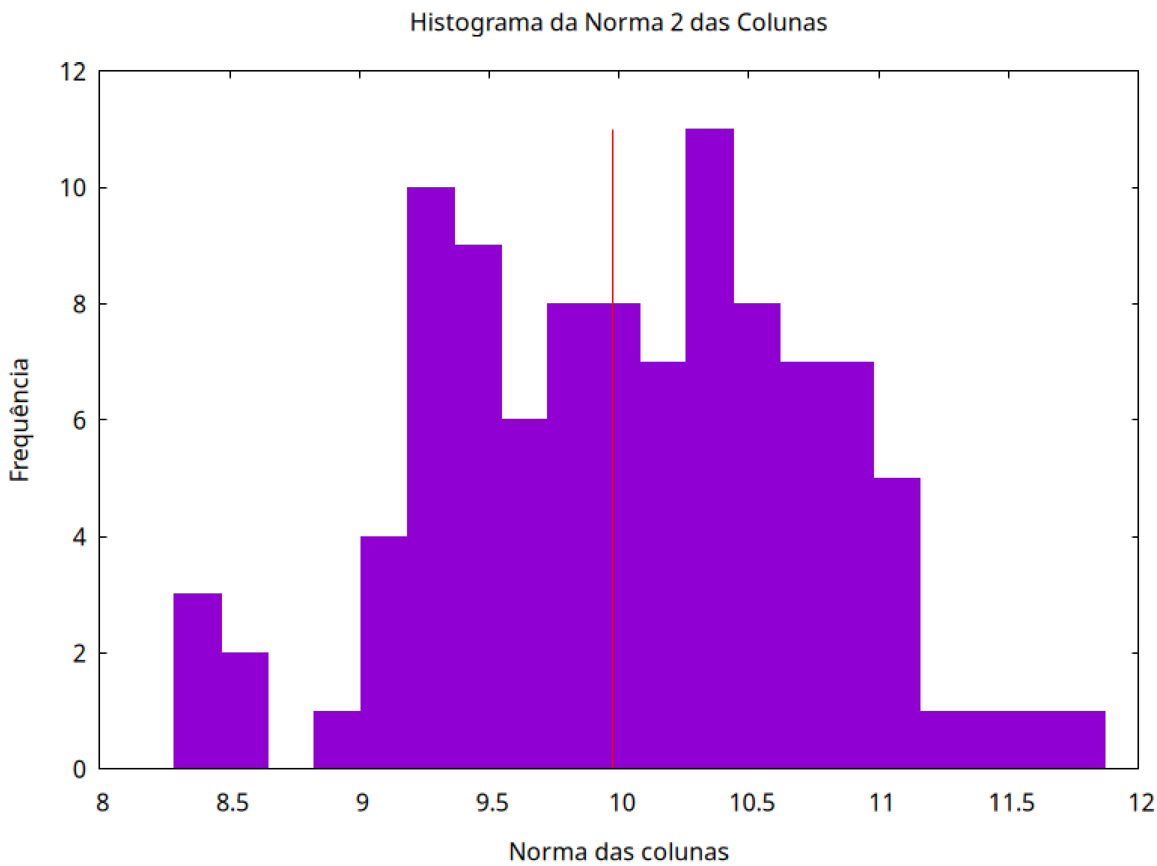
e, portanto, pela **Lei dos Grandes Números**, ela se aproxima de sua média o quanto se queira desde que m seja suficientemente grande.

<u>Prática:</u>

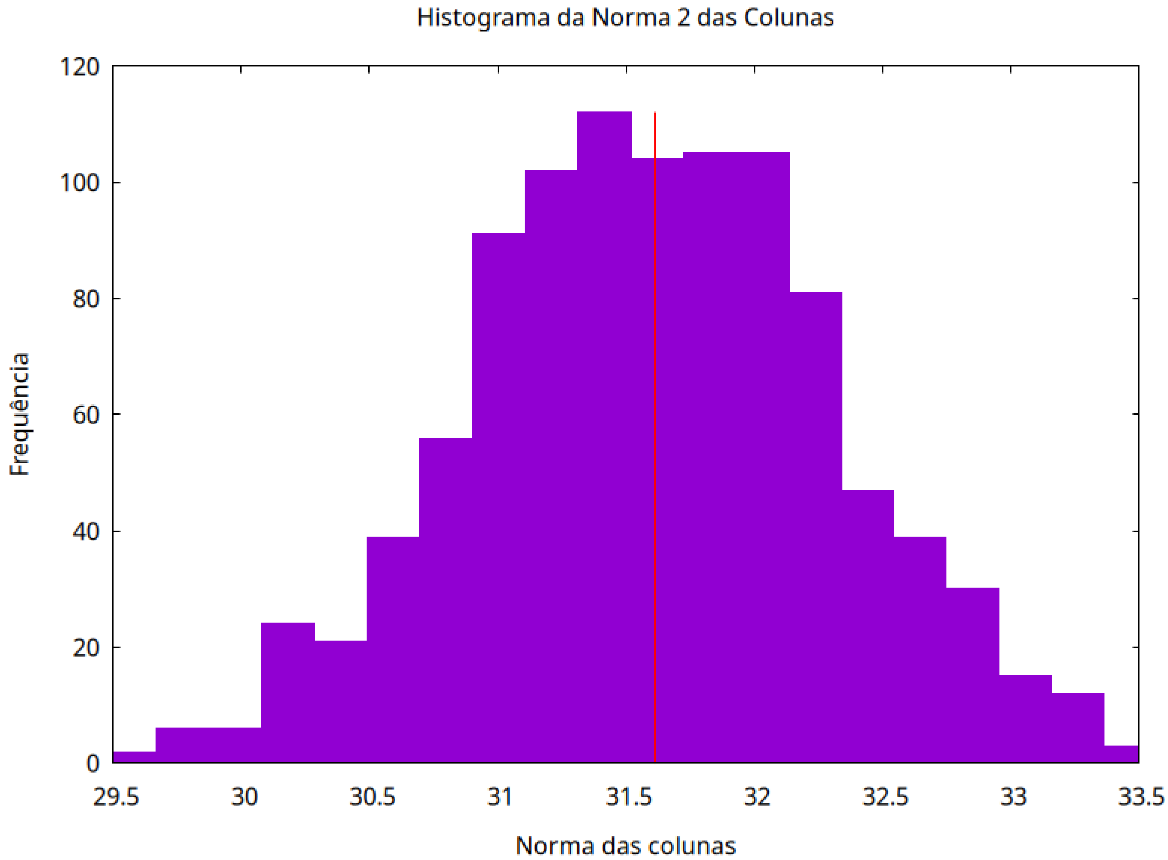
Plotando o histograma para valores variados de m e de n, podemos ver, de fato, esses valores se concentrando cada vez mais de μ (a linha vermelha vertical é o valor da média):



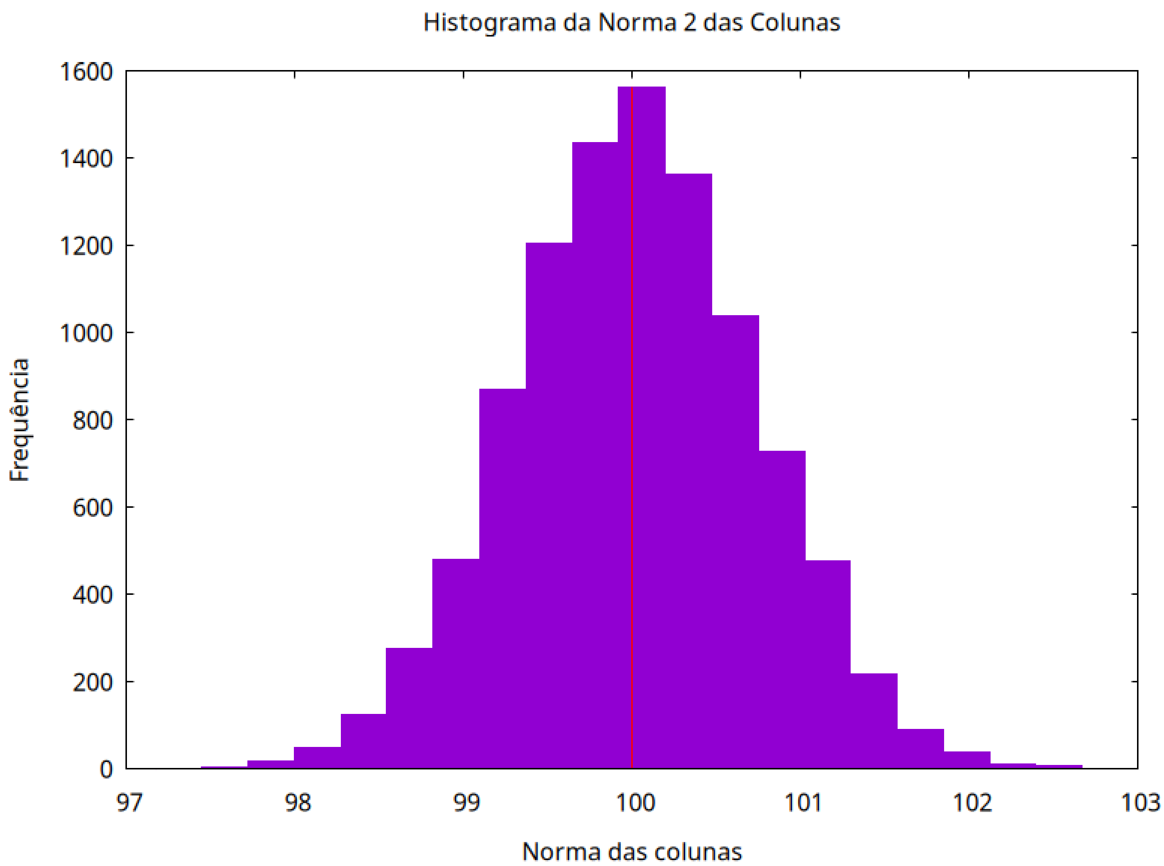
m = 10, n = 10


Histograma da Norma 2 das Colunas
Frequência
Norma das colunas
0
2
4
6
8
10
12
8
8.5
9
9.5
10
10.5
11
11.5
12

m = 100, n = 100


Histograma da Norma 2 das Colunas
Frequência
Norma das colunas
0
20
40
60
80
100
120
29.5
30
30.5
31
31.5
32
32.5
33
33.5

m = 1000, n = 1000



m = 10000, n = 10000

b)

<u>Teoria:</u>

Se $c_i,\, c_j\ \in \mathbb{R}^m$ são vetores aleatórios iid com distribuição normal padrão, então:

$$< c_i,\, c_j >= \sum_k^m c_{ik}c_{jk}$$

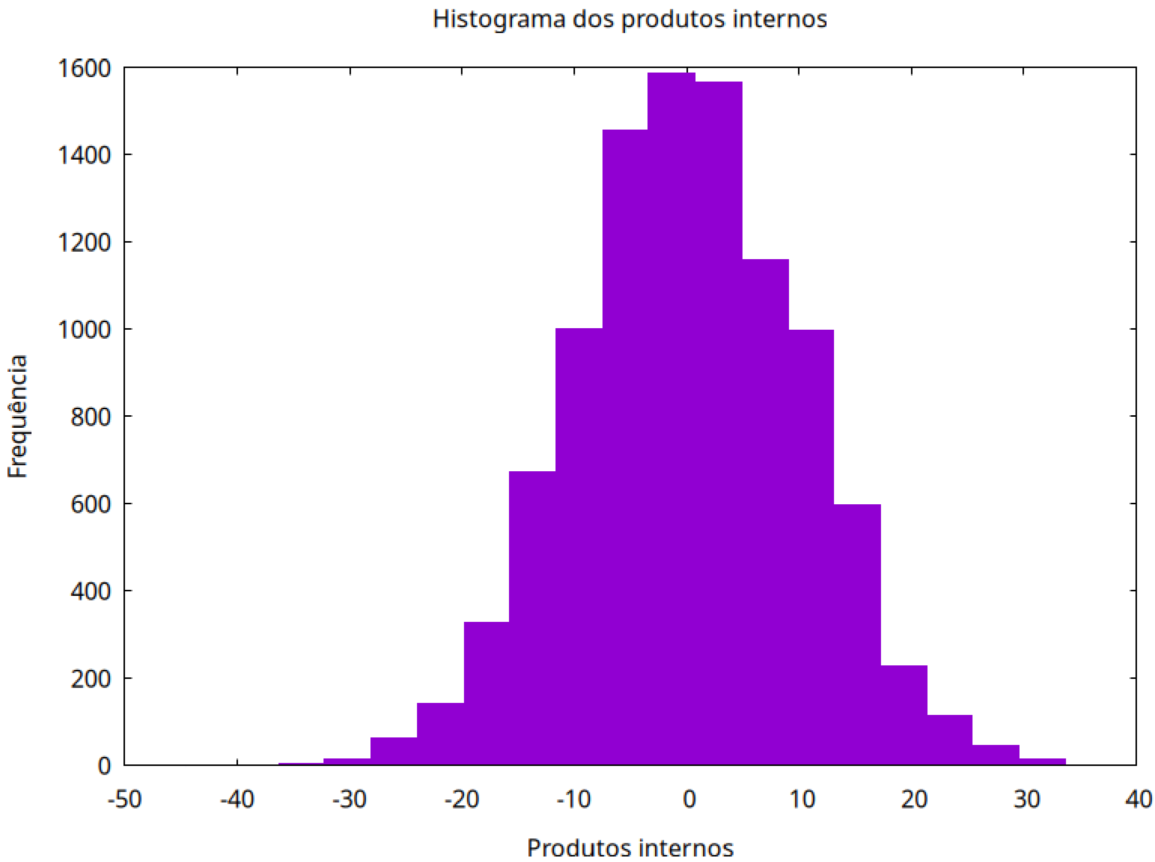
e, portanto:

$$E\left[< c_i,\, c_j >\right] = 0$$
$$Var\left[< c_i,\, c_j >\right] = m$$

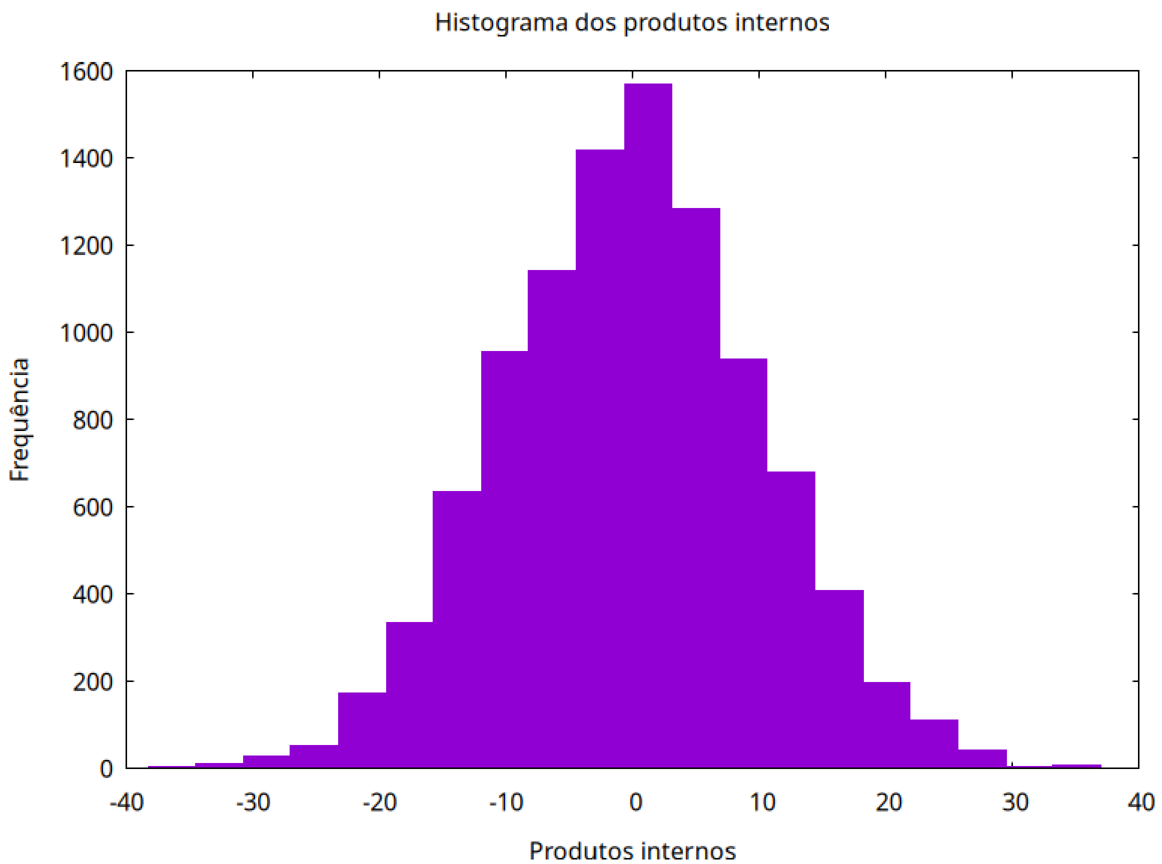
o que implica que os vetores $c_i,\, c_j$ tendem a ficar ortogonais entre si quando m cresce.

<u>Prática:</u>

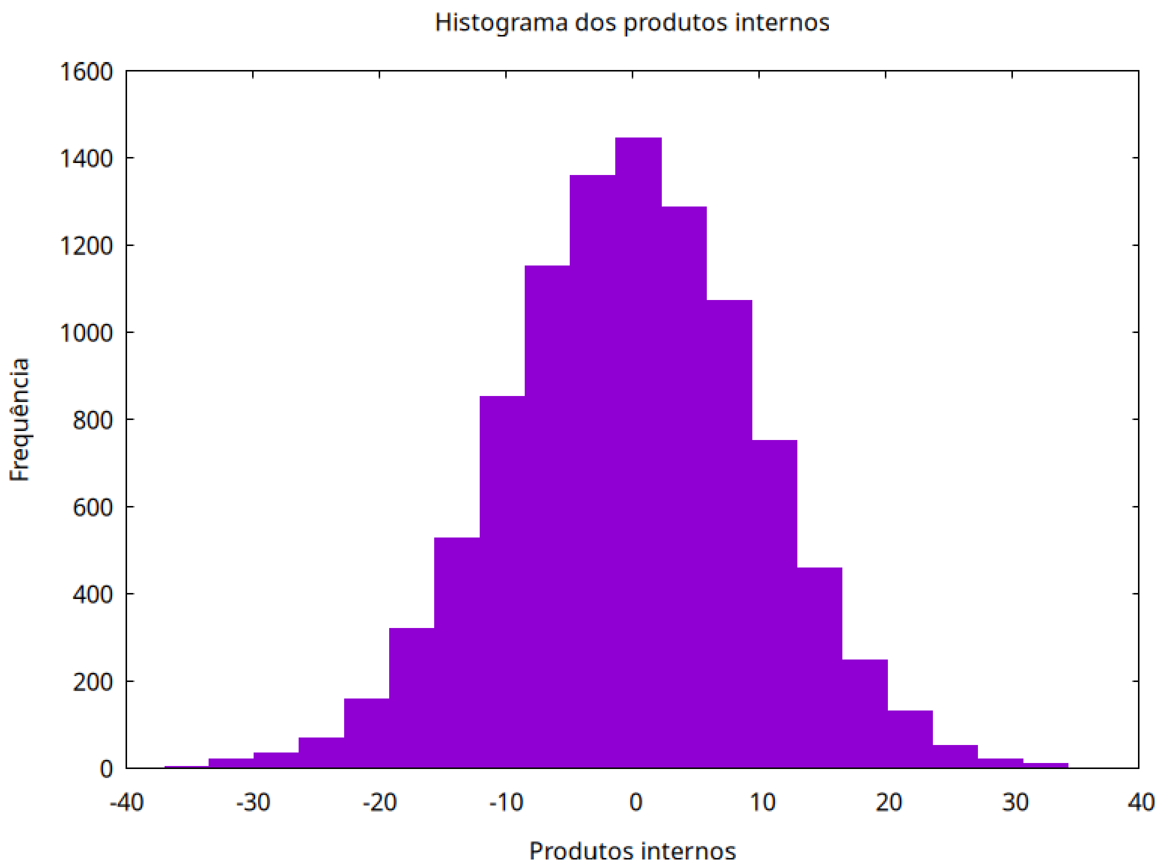
Plotando histogramas para o valor fixo de m e aumentando o valor de n, podemos perceber que os valores se concentram em torno do zero (o parâmetro *num_samples* diz quantos produtos internos devemos calcular de fato, pois em teoria deveríamos calcular $m.n$ deles, o que pode ser custoso. Aqui estou o deixando como 10000) :



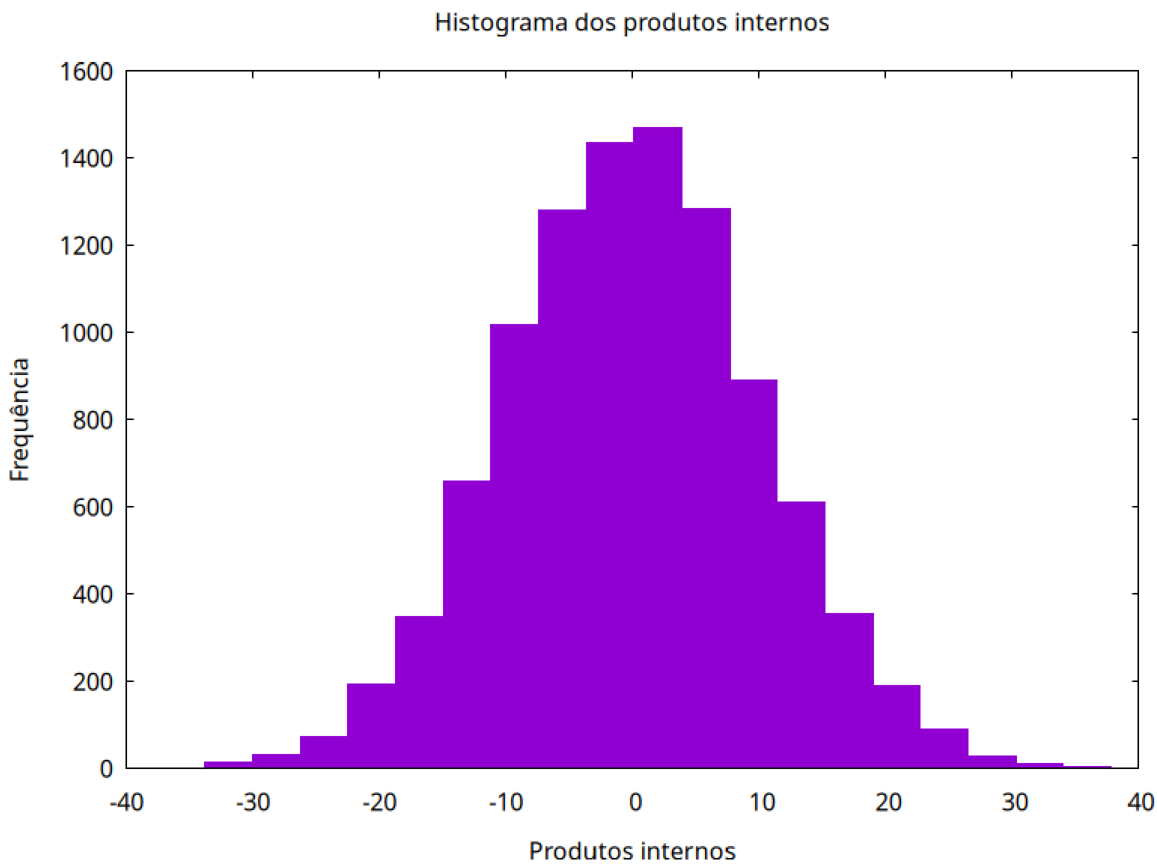
n = 100


Histograma dos produtos internos
Frequência
1600
1400
1200
1000
800
600
400
200
0
-40
-30
-20
-10
0
10
20
30
40
Produtos internos

n = 200


Histograma dos produtos internos
Frequência
Produtos internos
1600
1400
1200
1000
800
600
400
200
0
-40
-30
-20
-10
0
10
20
30
40

n = 500



n = 1000

c)

<u>Teoria:</u>

Queremos calcular o máximo de

$$|\cos\theta_{ij}| = \frac{< c_i,\, c_j >}{||c_i||\,||c_j||} \sim \left|Normal\left(0,\, \frac{1}{n}\right)\right|$$

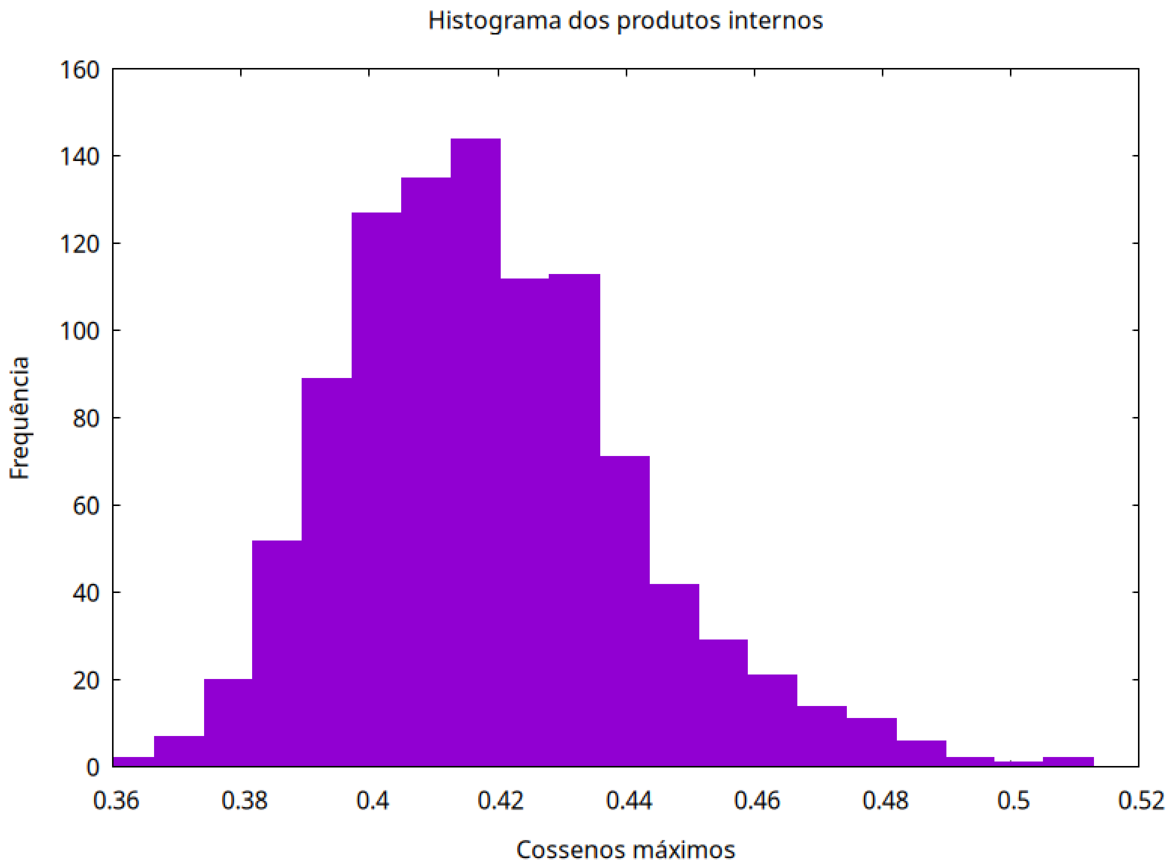
Como temos $N = \frac{300 \cdot 299}{2}$ dessa variáveis, o valor máximo tem comportamento esperado:

$E\left[\max\left(|\cos\theta_{ij}|\right)\right] \cong \sigma\sqrt{2\log N}$ onde $\sigma = \frac{1}{\sqrt{m}} = 0.1$, portanto o valor esperado se torna aproximadamente igual a 0.46

Então no histograma devemos observar valores em dentro de [0.3, 0.6] com pico aproximadamente em 0.46.

Prática:

Plotando o histograma, podemos observar que a análise teórica parece ser razoável:



d)

Queremos encontrar a complexidade de calcular

$$\max \left| \frac{< c_i, c_j >}{|c_i| \, |c_j|} \right|, i \neq j$$

Você tem $N = \frac{n(n-1)}{2}$ pares.

Calcular produtos internos tem complexidade $O(m)$
Normalizar com as normas já calculadas tem complexidade $O(1)$
Comparar e atualizar o máximo tem complexidade $O(1)$

Pré computar todas as normas tem complexidade $O\left(mn\right)$
Calcular todos os produtos internos normalizados tem complexidade $O\left(n^2m\right)$

Portanto a complexidade se dá quando:

- Pré-processamos as normas -> $O\left(mn\right)$
- Fazemos o loop principal $N\ \cdot\ O\left(m\right)\ =\ O\left(n^2m\right)$

Portanto a complexidade de calcular o máximo é igual a $O\left(n^2m\right)$

Agora, para a pergunta sobre K:

Seja Z o máximo do cosseno, então calculamos a média amostral:

$\mu_K = \frac{1}{K}\sum_{i=1}^{K} Z_k$, com variância:

$$Var\left(\mu_K\right) = \frac{Var\left(Z\right)}{K} \approx \frac{C}{K\log N} = \frac{C}{K\log n}$$

Então se queremos que a média seja menor que um erro ϵ, podemos tomar:

$$K \geq\ \frac{C}{\epsilon^2\log n}$$

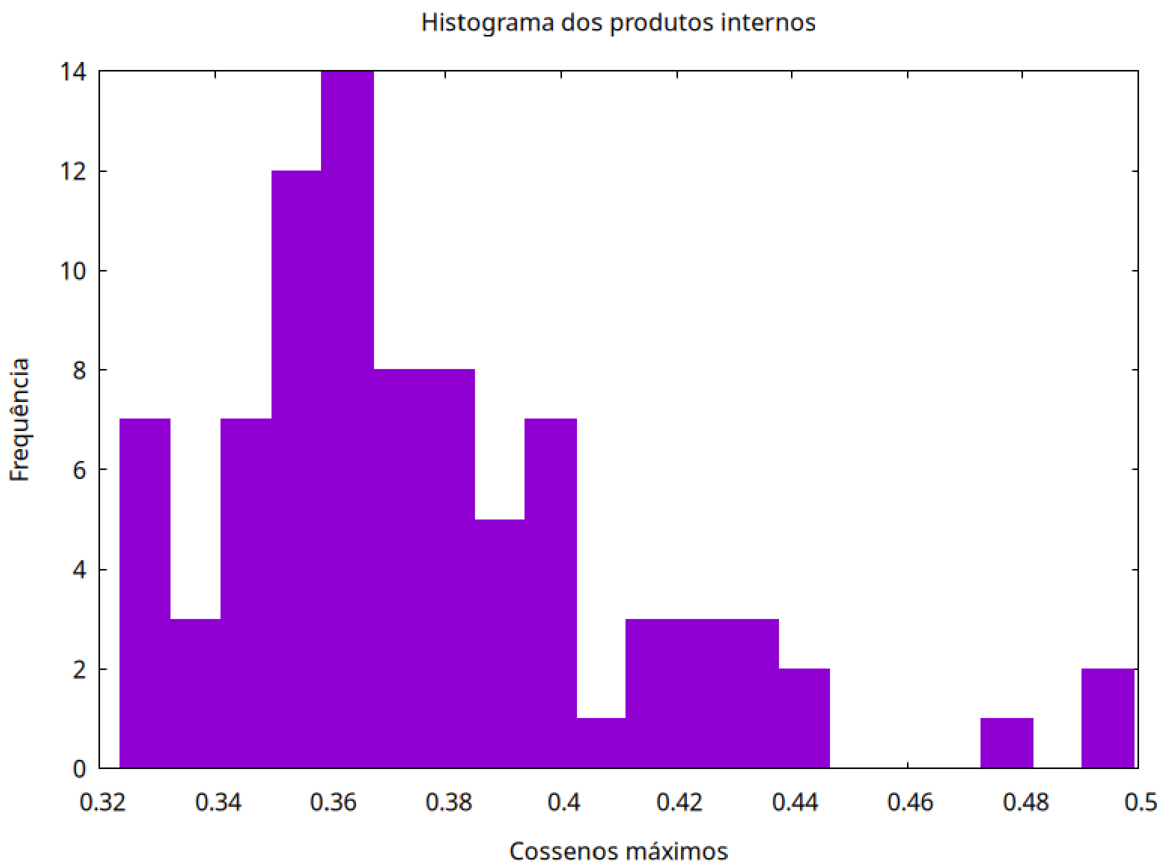
e)

Para fins humanitários, tomemos $C \approx 1$ e $\epsilon = 0.05$.

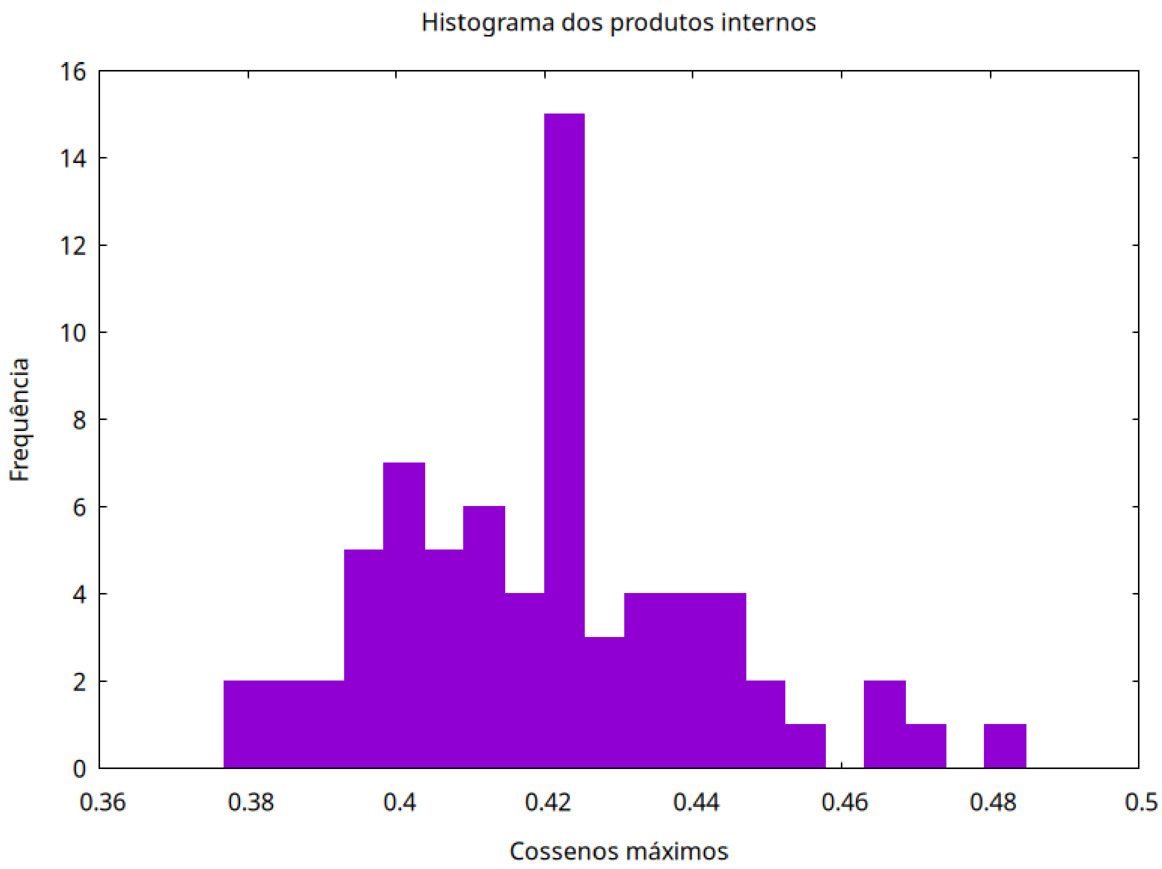
Então para n dado escolhemos

$$K = \frac{C}{\epsilon^2\log n}$$

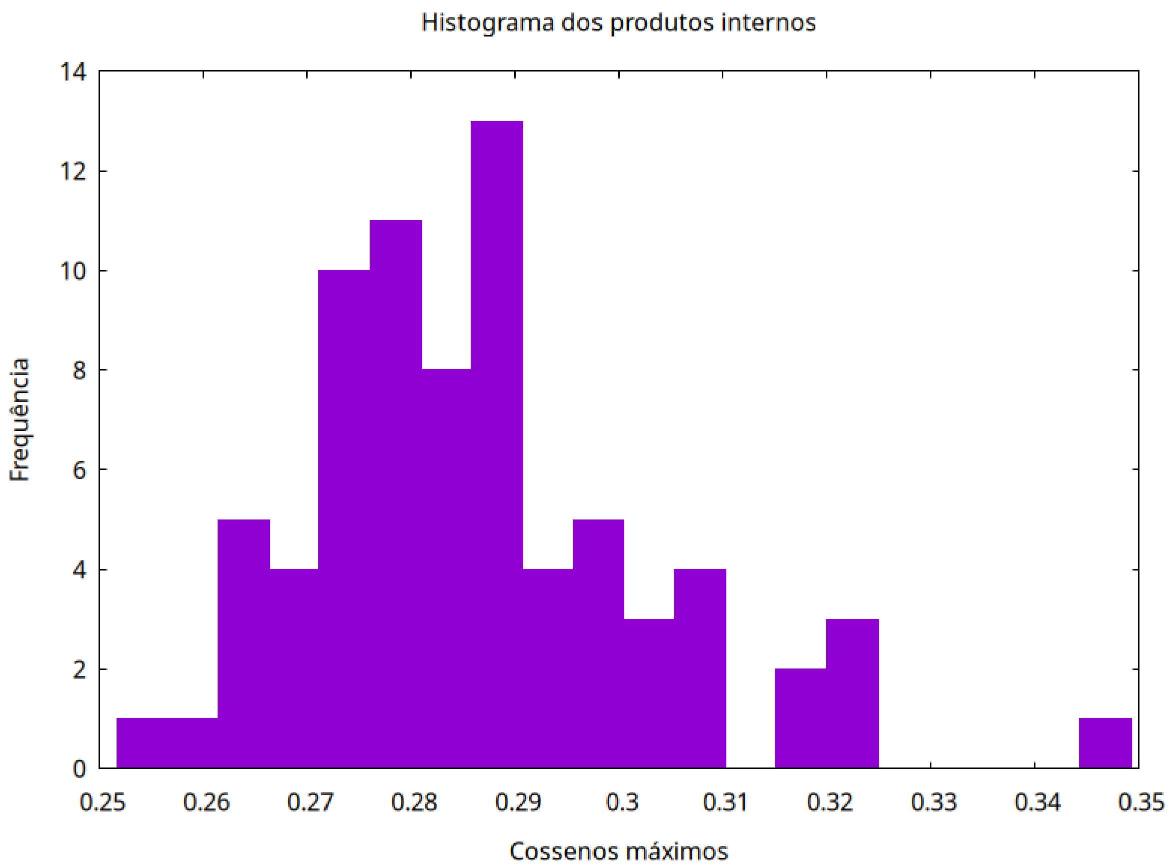
Aqui estão os histogramas:


Histograma dos produtos internos
Frequência
0
2
4
6
8
10
12
14
0.32
0.34
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5
Cossenos máximos

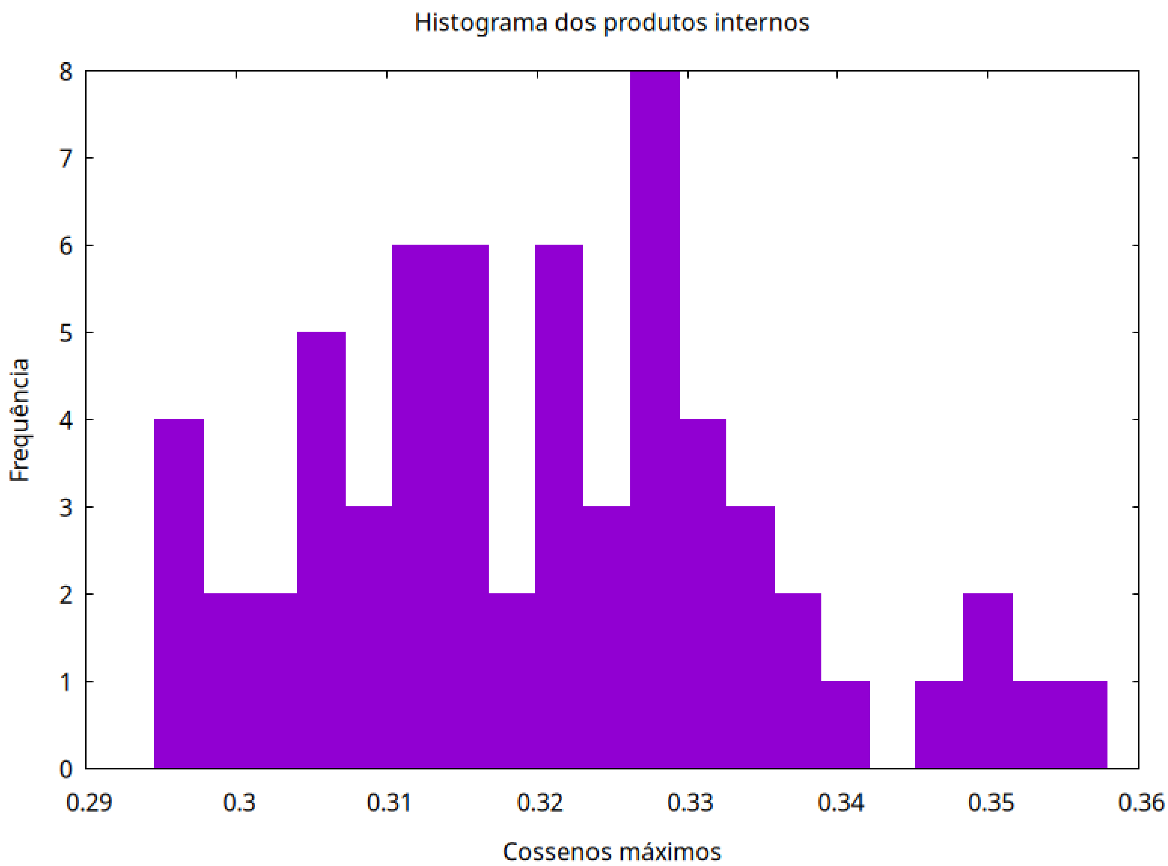
m = 100, n = 100


Histograma dos produtos internos
Frequência
Cossenos máximos
0
2
4
6
8
10
12
14
16
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5

m = 100, n = 300

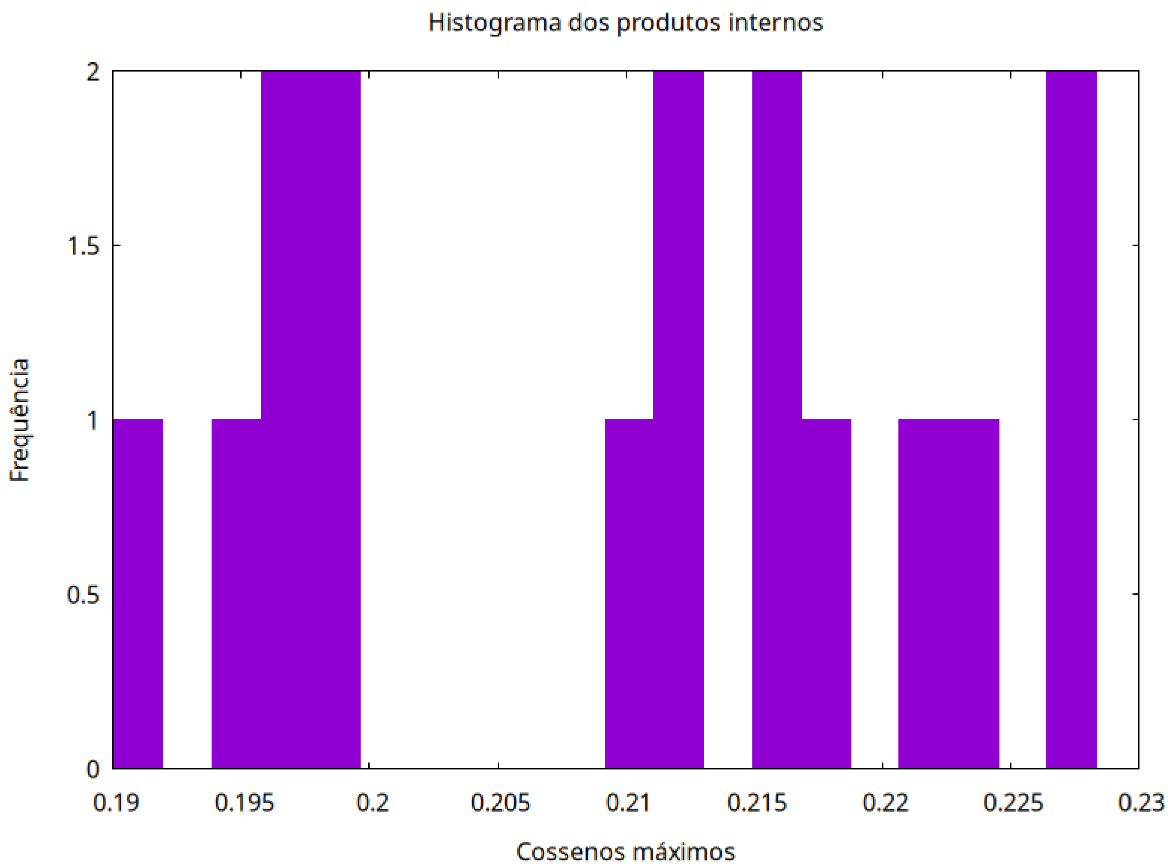

Histograma dos produtos internos
Frequência
Cossenos máximos
0
2
4
6
8
10
12
14
0.25
0.26
0.27
0.28
0.29
0.3
0.31
0.32
0.33
0.34
0.35

m = 200, n = 200

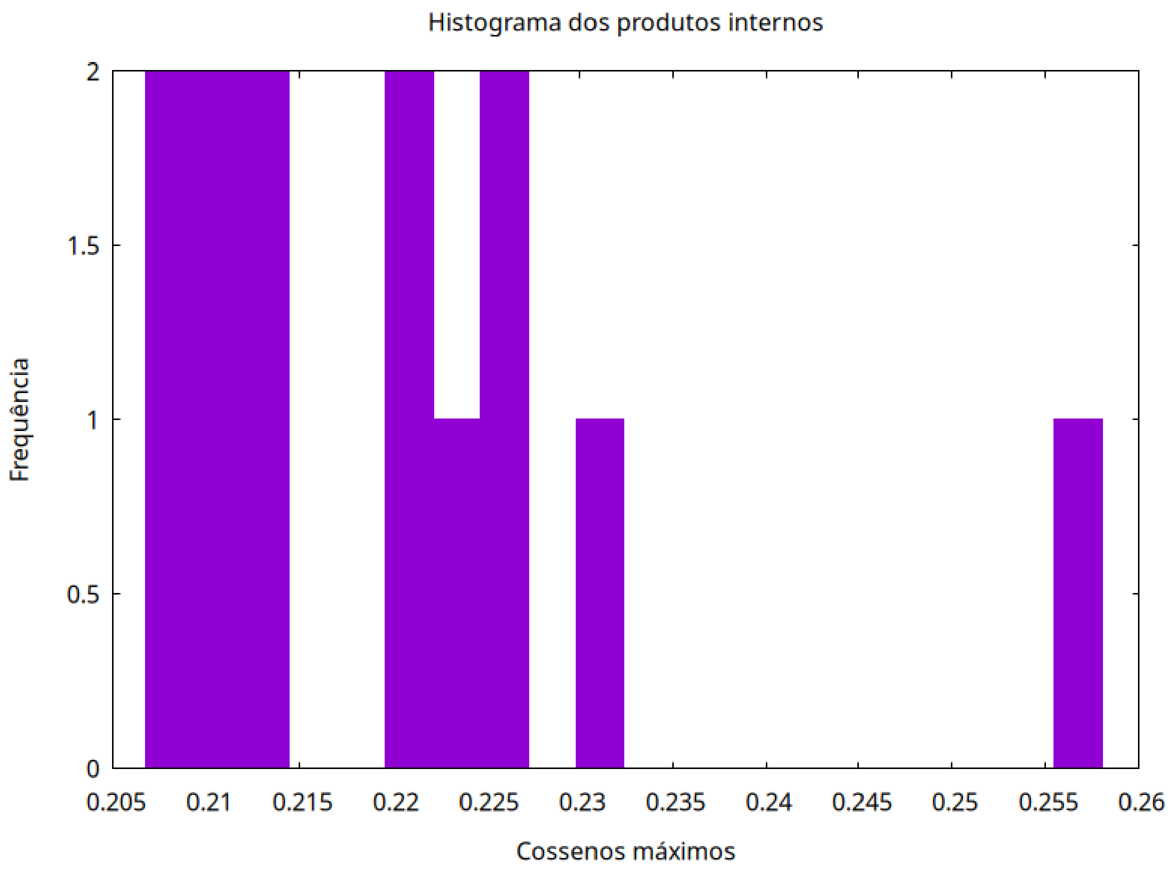


m = 200, n = 600

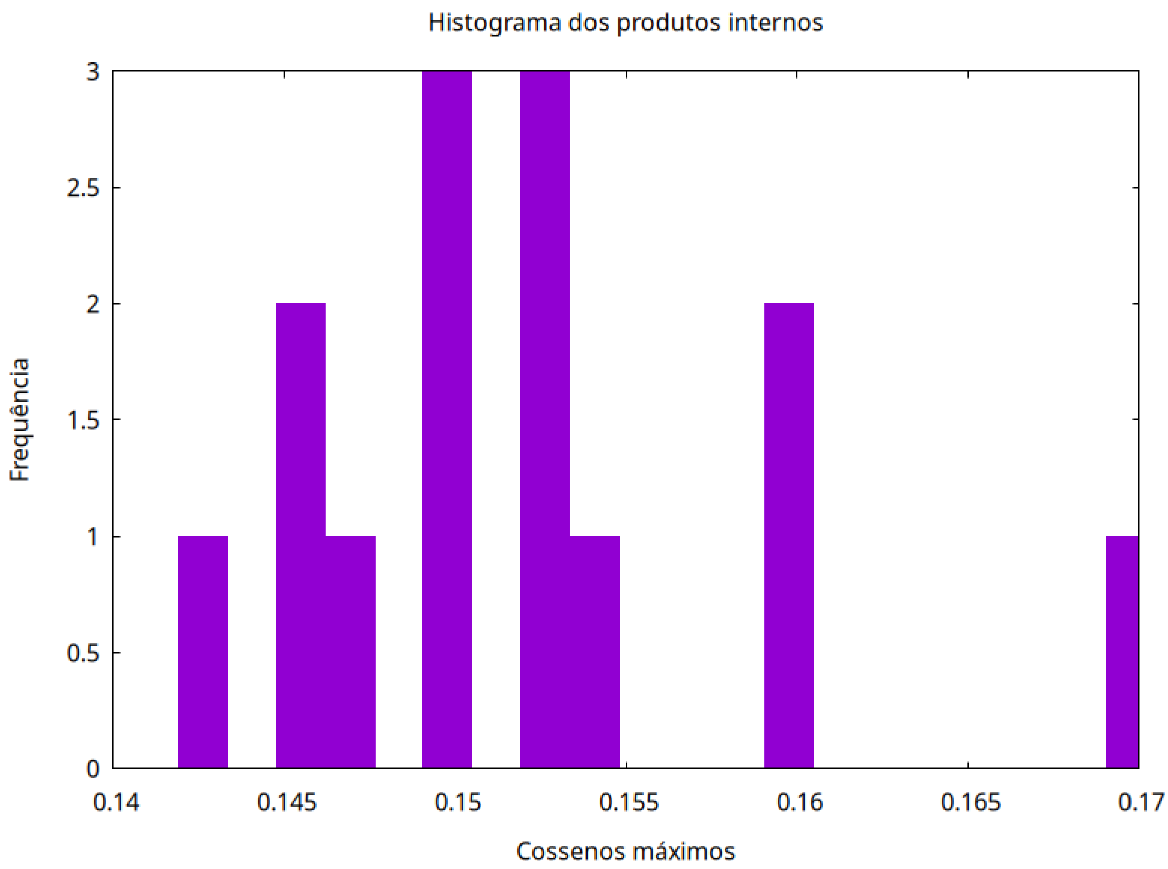
A partir de agora, para fins mais humanitários ainda, tomemos $\epsilon = 0.1$


Histograma dos produtos internos
Frequência
2
1.5
1
0.5
0
0.19
0.195
0.2
0.205
0.21
0.215
0.22
0.225
0.23
Cossenos máximos

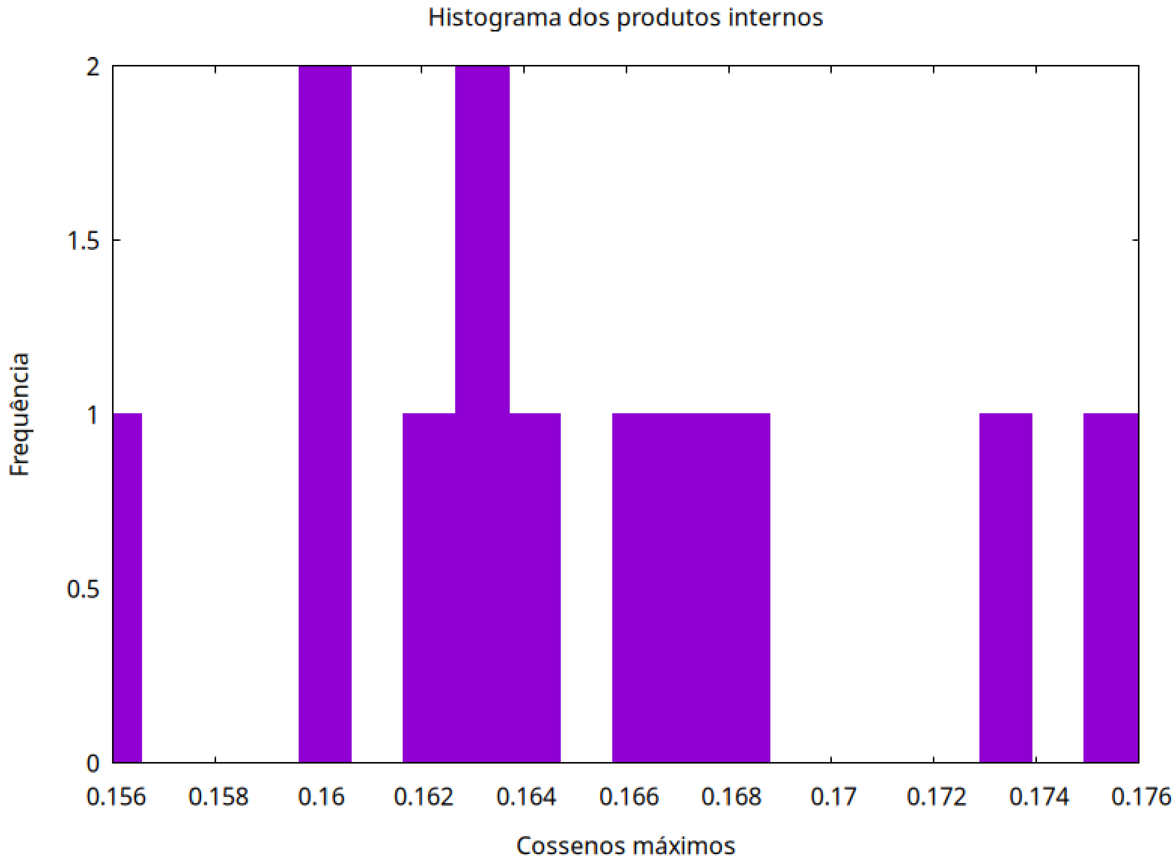
m = 500, n = 500


Histograma dos produtos internos
Frequência
2
1.5
1
0.5
0
0.205
0.21
0.215
0.22
0.225
0.23
0.235
0.24
0.245
0.25
0.255
0.26
Cossenos máximos

m = 500, n = 1500


Histograma dos produtos internos
Frequência
Cossenos máximos
0
0.5
1
1.5
2
2.5
3
0.14
0.145
0.15
0.155
0.16
0.165
0.17

m = 1000, n = 1000



m = 1000, n = 3000
(levou mais que 30 min para rodar)

O comportamento desses histogramas pode ser sintetizado nas seguintes regras, dado que

$$E[Z] \sim \sqrt{\frac{\log n}{m}}$$

- se m é fixo e n é aumentado, o máximo deve crescer e o histograma deve ser deslocado um pouco para a direita.
- se n é fixo e m é aumentado, o máximo deve diminuir e o histograma deve ser deslocado um pouco para a esquerda.