Factor de crescimento na eliminação de Gauss e pivotagem parcial

Trabalho realizado por Rui Filipe Mendes Alves da Costa, sob a orientação de Filomena Dias d'Almeida e Paulo Beleza Vasconcelos, no âmbito da Bolsa de Iniciação Cientifica "Factor de crescimento na eliminação de Gauss e pivotagem parcial" financiada pelo Centro de Matemática da Universidade do Porto.

Agradecimentos

O autor agradece ao Centro de Matemática da Universidade do Porto e aos Professores Filomena Dias d'Almeida e Paulo Beleza Vasconcelos pela oportunidade de realizar este trabalho, em especial aos Professores, pela orientação cuidada, pela motivação e apoio dispendido ao longo do trabalho. Agradecimentos especiais a todos os colegas, que deram apoio e ajuda fundamentais; à irmã, por todo o carinho; aos pais, por todo o esforço e dedicação. E à Carla, por todo o amor, por toda a força que deu e por toda a paciência demonstrada.

Índice

Introdução

No início da história da computação, matemáticos famosos, como, por exemplo, von Neumann e Hotelling, previam que o método de eliminação gaussiana fosse instável, devido à acumulação de erros de arredondamento. Pensava-se que o método seria estável apenas para matrizes de pequena dimensão. Mais tarde, com o evoluir da computação, veio a verificar-se que o método afinal era estável também para matrizes de grande dimensão. Desde então até aos dias de hoje, explicar esse fenómeno tornou-se um desafio teórico de grande importância. James Wilkinson foi um dos principais investigadores dedicados a esta questão. Num dos seus inúmeros trabalhos, concluiu que para estudar a estabilidade do método bastava estudar o crescimento das entradas da matriz ao longo do processo, mostrando que se esse crescimento não fosse elevado, o método seria estável; caso contrário, deveria esperar-se instabilidade. O crescimento dos elementos da matriz é medido através do factor de crescimento.

O factor de crescimento pode atingir valores elevadíssimos, mas em 50 anos de computação, não se conhece nenhuma matriz que tenha um factor de crescimento elevado e que seja proveniente de um problema real. Conhecem-se matrizes cujo factor de crescimento é bastante elevado, mas são matrizes não provenientes de problemas reais. Wilkinson e os seus contemporâneos não se dedicaram ao estudo deste fenómeno, mas, no seu livro *Algebraic Eigenvalue Problem*, Wilkinson afirma *"experience suggests that though such a bound is attainable it is quite irrelevant for practical purposes"*. Os primeiros a fornecerem um avanço nesta matéria foram L. N. Trefethen e R. S. Schreiber. No seu trabalho *Average-case stability of Gaussian elimination*, em 1990, mostram que o fenómeno é inteiramente estatístico.

O objectivo do nosso trabalho é estudar com detalhe o método de eliminação gaussiana, em particular o factor de crescimento, apresentando no final um estudo estatístico, elaborado com matrizes aleatórias, que evidencia o facto de, na pratica, não aparecerem matrizes que provoquem instabilidade explosiva no método de eliminação gaussiana.

1. Preliminares

1.1. Sistemas de numeração em vírgula flutuante

Uma operação aritmética realizada num computador é geralmente afectada por um erro de arredondamento. Isto acontece devido ao facto de um computador representar um número real com um número finito de dígitos, podendo, assim, representar apenas um subconjunto finito e discreto dos números reais. Esta limitação implica que os números representados não podem ser arbitrariamente grandes nem arbitrariamente pequenos. Além disso, tem de existir um "buraco" entre os números representados.

O sistema de números universalmente utilizado é o chamado **sistema de numeração em vírgula flutuante**. Neste sistema, a vírgula é guardada separadamente dos dígitos. Além disso, o buraco entre números adjacentes tem amplitude proporcional ao tamanho dos mesmos.

Existem diversas formas deste sistema. Apresentamos a que nos parece mais simples de compreender, visto não ser nosso objectivo tratar aprofundadamente este assunto.

O sistema de numeração em vírgula flutuante num computador é caracterizado por quatro inteiros: a base β (usualmente 2); a precisão t ($t = 24$ diz-se precisão simples e $t = 53$ diz-se precisão dupla; são as mais utilizadas); e o intervalo do expoente $[e_{\min}, e_{\max}]$. Sendo X o subconjunto dos números reais representado, um elemento $x \in X$ é da forma:

$$x = \pm .d_1 d_2 ... d_t \times \beta^e \text{, com } 0 \le d_i < \beta \text{ e } e \in [e_{\min}, e_{\max}].$$

Note-se que, para $\forall x \in X$, se tem $\beta^{e_{\min}-1} \le |x| \le \beta^{e_{\max}}(1-\beta^{-t})$.

A caracterização de X é tradicionalmente feita pelo número $\varepsilon_M = \frac{1}{2}\beta^{1-t}$.

A sua importância advém do facto de ter a seguinte propriedade:

$$\forall x \in \Re, \exists \hat{x} \in X : |x - \hat{x}| \le \varepsilon_M |x|$$

Ou seja, ε_M é erro relativo máximo com que cada número é representado.

Seja $fl : \Re \to X$ definida do seguinte modo: para $x \in \Re$, $fl(x)$ é o elemento do sistema de números em virgula flutuante mais próximo de x. Assim, temos que:

$$\forall x \in \Re, \quad \frac{|fl(x) - x|}{|x|} \le \varepsilon_M$$

A seguinte tabela mostra os valores dos parâmetros de sistemas de numeração em vírgula flutuante de algumas máquinas. O sistema que o Matlab utiliza é o IEEE duplo, criado pelo Institute of Electrical and Electronics Engineers, em 1985.

Máquina	β	T	$e_{\min}$	$e_{\max}$	ε_M
HP 28 e calculadoras 48G	10	12	-499	499	5×10^{-12}
IBM 3090 simples	16	6	-64	63	5×10^{-7}
IBM 3090 duplo	16	14	-64	63	1×10^{-16}
IBM 3090 expandido	16	28	-64	63	2×10^{-33}
IEEE simples	2	24	-125	128	6×10^{-8}
IEEE duplo	2	53	-1021	1024	1×10^{-16}
IEEE expandido	2	64	-16381	16384	5×10^{-20}

Passemos à **aritmética em vírgula flutuante**. As operações elementares em $\Re$, $+,\times$ e as sua inversas têm operações análogas definidas em X, que designaremos por $\oplus,\otimes$ (e suas inversas). A grande maioria dos computadores é construída seguindo o seguinte princípio:

Axioma fundamental da aritmética em virgula flutuante

Seja $\circ$ uma das operações elementares em $\Re$ e $\bullet$ a sua análoga em X. Então, $\forall x, y \in X, \exists \varepsilon, |\varepsilon| \le \varepsilon_M$ tal que $x \bullet y = (x \circ y)(1+\varepsilon)$.

Ou seja, o erro relativo cometido numa operação aritmética é $\dfrac{|x \bullet y - x \circ y|}{|x \circ y|} \le \varepsilon_M$.

1.2. Significado de $O(\varepsilon_M)$

Passemos à introdução de uma notação que nos acompanhará ao longo do nosso estudo. Usa-se a notação $g(t) \underset{t \to t_0}{=} O(f(t))$ quando, para todo o *t* suficientemente próximo de t_0, existe uma constante positiva *k* tal que $|g(t)| \le k|f(t)|$.

No nosso estudo, esta notação aparecerá de formas similares à seguinte: $\|x\| = O(\varepsilon_M)$. Aqui, o limite implícito é $\varepsilon_M \to 0$, omitindo este índice por comodidade.

É claro que ε_M é uma quantidade fixa em cada computador. Ao dizer $\varepsilon_M \to 0$, estamos a considerar a idealização de um computador. Isto é, $\|x\| = O(\varepsilon_M)$ significa que, ao executar o mesmo algoritmo numa sequência de computadores com ε_M a decrescer para zero, $\|x\|$ decresce para zero proporcionalmente a ε_M ou ainda mais rapidamente.

1.3. Normas de matrizes

As normas matriciais serão muito importantes no nosso estudo posterior, pois é o que nos permite, por exemplo, estudar erros. Mas antes de passarmos à sua introdução, é necessário introduzir o conceito de norma de um vector.

Definição: $\|\cdot\| : \Re^n \to \Re$ é uma norma em $\Re^n$ se verifica as seguintes propriedades:

i. $\|x\| \geq 0 \quad \forall x \in \Re^n$ e $\|x\| = 0 \Leftrightarrow x = 0$

ii. $\|\alpha x\| = |\alpha| \|x\| \quad \forall \alpha \in \Re, \forall x \in \Re^n$

iii. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \Re^n$

A classe de normas mais utilizada são as p-normas, definidas por $\|x\|_p = \left(\sum_{i=1}^{n} {x_i}^p \right)^{1/p}$. Dentro desta classe, as mais importantes são:

$$\|x\|_1 = |x_1| + \ldots + |x_n|$$
$$\|x\|_2 = \left(|x_1|^2 + \ldots + |x_n|^2 \right)^{1/2}$$
$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

É de fácil verificação que se tratam, de facto, de normas.

Um resultado importante é o facto de todas as normas em $\Re^n$ serem equivalentes, i.e., se $\|\cdot\|_i$ e $\|\cdot\|_j$ são normas em $\Re^n$, então existem constantes positivas c_1 e c_2 tais que $c_1 \|x\|_i \leq \|x\|_j \leq c_2 \|x\|_i \quad \forall x \in \Re^n$. Este resultado é consequência imediata do teorema que nos diz que, num espaço de dimensão finita, todas as normas são equivalentes. A demonstração deste resultado sai fora do âmbito deste trabalho, podendo ser encontrado em qualquer livro de álgebra linear.

Temos, por exemplo, que:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2$$
$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty$$
$$\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty$$

Passemos, então, a introduzir o conceito de **norma de uma matriz**. Este conceito é um dos mais utilizados na análise de algoritmos matriciais.

Definição: $\|\cdot\|:\Re^{m\times n}\to\Re$ é uma norma matricial se verifica as seguintes propriedades:

i. $\|A\|\geq 0 \quad \forall A\in\Re^{m\times n}$ e $\|A\|=0\Leftrightarrow A=0$

ii. $\|\alpha A\|=|\alpha|\|A\| \quad \forall\alpha\in\Re,\forall A\in\Re^{m\times n}$

iii. $\|A+B\|\leq\|A\|+\|B\| \quad \forall A,B\in\Re^{m\times n}$

As normas mais utilizadas em álgebra linear numérica são a **norma de Frobenius**, definida por $\|A\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{i,j}|^2}$, e as **p-normas**, definidas por $\|A\|_p=\sup_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}$.

É um resultado de fácil demonstração que $\forall p,\|AB\|_p\leq\|A\|_p\|B\|_p$.

Como $\Re^{m\times n}$ é um espaço de dimensão finita, temos:

Teorema: Todas as normas em $\Re^{m\times n}$ são equivalentes.

É um resultado muito importante para o estudo desenvolvido neste trabalho, pois permite-nos trabalhar numa norma à escolha e depois adaptar os resultados a outras normas, se necessário. Como anteriormente, a sua demonstração sai fora do âmbito deste trabalho.

De seguida, apresentamos algumas propriedades bastante utilizadas na análise de algoritmos matriciais.

- $\|A\|_2\leq\|A\|_F\leq\sqrt{n}\|A\|_2$
- $\max_{i,j}|a_{i,j}|\leq\|A\|_2\leq\sqrt{mn}\max_{i,j}|a_{i,j}|$
- $\|A\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{i,j}|$
- $\|A\|_\infty=\max_{1\leq j\leq m}\sum_{i=1}^{n}|a_{i,j}|$
- $\frac{1}{\sqrt{n}}\|A\|_\infty\leq\|A\|_2\leq\sqrt{m}\|A\|_\infty$
- $\frac{1}{\sqrt{m}}\|A\|_1\leq\|A\|_2\leq\sqrt{n}\|A\|_1$

1.4. Sistemas de equações equivalentes e matrizes elementares

Um sistema de equações lineares $\begin{cases} a_{11}x_1 + ... + a_{1n}x_n = b_1 \\ \quad\quad ... \\ a_{n1}x_1 + ... + a_{nn}x_n = b_n \end{cases}$ pode ser representado da forma $Ax = b$, onde $A = \begin{pmatrix} a_{11} & ... & a_{1n} \\ ... & ... & ... \\ a_{n1} & ... & a_{nn} \end{pmatrix}$, $x = (x_1,...,x_n)^T$ e $b = (b_1,...,b_n)^T$.

Supomos, por simplicidade, o sistema de equações definido em $\Re$.

Dois sistemas de equações lineares $Ax = b$ e $\hat{A}x = \hat{b}$ dizem-se **equivalentes** se qualquer solução de $Ax = b$ é solução de $\hat{A}x = \hat{b}$ e vice-versa.

Definição: Dizem-se elementares as seguintes operações numa matriz:

i. Multiplicação de uma linha por um escalar $\alpha \neq 0$
ii. Adição de um múltiplo escalar de uma linha a outra linha
iii. Permutação de duas linhas

Esta definição (bem como os seguintes resultados) pode ser descrita também para as mesmas operações nas colunas. Mas, visto que o método de eliminação gaussiana opera nas linhas, ficamos por esta definição.

Teorema: Seja $A = (A_1 \ \vdots \ b_1)$. Se $\widetilde{A} = (\widetilde{A}_1 \ \vdots \ \widetilde{b}_1)$ é obtida a partir de A por operações elementares, então os sistemas $A_1 x = b_1$ e $\widetilde{A}_1 x = \widetilde{b}_1$ são equivalentes.

A demonstração deste resultado pode ser encontrada em qualquer livro de Álgebra Linear.

As operações elementares numa matriz podem ser escritas numa forma matricial. Passemos então à definição de matrizes elementares.

Definição: Chamam-se **matrizes elementares** às matrizes que se obtêm da matriz identidade por uma e uma só operação elementar.

Utilizaremos a seguinte notação para as matrizes elementares:

- $M_i(\alpha)$: matriz identidade com $\alpha \neq 0$ na posição (i,i)
- $L_{ij}(\alpha)$ $(i \neq j)$: matriz identidade com $\alpha \neq 0$ na posição (i,j)
- P_{ij} : matriz identidade com zero nas posições (i,i) e (j,j), e 1 nas posições (i,j) e (j,i)

Apresentamos, agora, algumas propriedades das matrizes elementares:

- $M_i(\alpha) = I + (\alpha - 1)e_i e_i^T$, $\det(M_i(\alpha)) = \alpha$, $\left(M_i(\alpha)\right)^{-1} = M_i(\frac{1}{\alpha})$
- $L_{ij}(\alpha) = I + \alpha e_i e_j^T$, $\det(L_{ij}(\alpha)) = 1$, $\left(L_{ij}(\alpha)\right)^{-1} = L_{ij}(-\alpha)$
- $P_{ij} = I - (e_i - e_j)(e_i - e_j)^T$, $\det(P_{ij}) = \pm 1$, $\left(P_{ij}\right)^{-1} = P_{ij}$

É fácil verificar que uma operação elementar numa matriz é equivalente à pré-multiplicação dessa matriz por uma matriz elementar. Mais precisamente, a pré-multiplicação por:

- $M_i(\alpha)$ corresponde a multiplicar a *i*-ésima linha por α
- $L_{ij}(\alpha)$ corresponde a somar à *i*-ésima linha o produto de α pela *j*-ésima linha
- P_{ij} corresponde a trocar a linha *i* com a linha *j*.

2. Método de eliminação gaussiana

Daqui em diante, trabalharemos com a norma $\|.\|_\infty$ por uma questão de conveniência. Não há perda de generalidade, visto que todas as normas em $\Re^{m\times n}$ são equivalentes. Assim, omitiremos o índice ∞, escrevendo $\|.\|$ no lugar de $\|.\|_\infty$.

2.1. Sistemas triangulares

Antes de passarmos ao estudo da eliminação gaussiana, estudemos o caso particular de sistemas triangulares. Este caso é considerado em primeiro lugar porque o método de eliminação gaussiana pressupõe a capacidade de resolver sistemas triangulares.

Consideremos o sistema triangular superior (suponhamos invertível):

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ \qquad a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \qquad\qquad \ldots \\ \qquad\qquad\qquad a_{nn}x_n = b_n \end{cases}$$

Ora, $x_n = \dfrac{b_n}{a_{nn}}$. Substituindo x_n na penúltima equação, obtemos $x_{n-1} = \dfrac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$. Continuando este processo, obtemos $x_1 = \dfrac{b_1 - a_{12}x_2 - \ldots - a_{1n}x_n}{a_{11}}$. Este método é chamado **método de substituição para trás**. O algoritmo que o descreve é o seguinte:

Algoritmo do método de substituição para trás

$$x_n = \frac{b_n}{a_{nn}}$$

Para *i* de *n-1* até 1 fazer

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}$$

A resolução de um sistema triangular inferior pode ser feita por um método análogo, a que chamamos **método de substituição para a frente**:

Algoritmo do método de substituição para a frente

$$x_1 = \frac{b_1}{a_{11}}$$

Para *i* de 2 até *n* fazer

$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij} x_j}{a_{ii}}$$

Os seguintes resultados mostram-nos a qualidade da solução obtida computacionalmente na resolução de sistemas triangulares pelos métodos descritos atrás. A sua demonstração pode ser encontrada em [Hig].

Proposição 1: Seja $L \in \Re^{n\times n}$ uma matriz triangular inferior não-singular. Na resolução computacional do sistema $Lx = b$ por substituição para a frente, a solução obtida $\hat{x}$ satisfaz $(L+E)\hat{x} = b$, com $\|E\| \le n\varepsilon_M \|L\| + O(\varepsilon_M^2)$.

Proposição 2: Seja $U \in \Re^{n\times n}$ uma matriz triangular superior não-singular. Na resolução computacional do sistema $Ux = b$ por substituição para trás, a solução obtida $\hat{x}$ satisfaz $(U+E)\hat{x} = b$, com $\|E\| \le n\varepsilon_M \|U\| + O(\varepsilon_M^2)$.

2.2. Factorização A=LU

O método de eliminação gaussiana é utilizado para resolver sistemas de equações lineares $Ax = b$. Consiste em aplicar operações elementares à matriz $(A \;\vdots\; b)$ de modo a transformar o sistema num sistema equivalente $Ux = \tilde{b}$, com *U* triangular superior. Após este processo, aplicamos o método de substituição para trás para resolver $Ux = \tilde{b}$. Este método pode ser descrito pelo seguinte algoritmo:

Algoritmo do método de eliminação gaussiana

U=A, L=I
Para *j* de 1 até *n*-1 fazer
  Para i de *j*+1 até *n* fazer

$$l_{ij} = \frac{u_{ij}}{u_{jj}}$$

  Para *k* de *j* até *n* fazer

$$u_{ik} = u_{ik} - l_{ij} u_{jk}$$

Este método pode ser descrito por pré-multiplicações de matrizes elementares. O *k*-ésimo passo corresponde à pré-multiplicação da matriz *U* (obtida até esse passo) por um produto de matrizes elementares $L_{ik}\left(\frac{u_{ik}}{u_{kk}}\right)$, *i*=*k*+1,…,*n*. Como o produto de matrizes triangulares inferior é triangular inferior, temos que o *k*-ésimo passo corresponde à pré-multiplicação de *A* por

L_k, matriz triangular inferior (é fácil ver que a diagonal de L_k é constituída por 1's). Assim, o método de eliminação gaussiana pode ser descrito por:

$$L_{n-1}L_{n-2}...L_2L_1A = U$$

Tomando $L^{-1} = L_{n-1}L_{n-2}...L_2L_1$, temos $L^{-1}A = U \Leftrightarrow A = LU$. L^{-1} é triangular inferior com 1's ao longo da diagonal, pois é o produto de matrizes com a mesma propriedade. Logo, L também é triangular inferior com 1's ao longo da diagonal.

Concluímos, assim, que o método da eliminação gaussiana fornece a **factorização *A=LU***, com *U* triangular superior e *L* triangular inferior com 1's ao longo da diagonal. A interpretação de método de eliminação gaussiana como uma factorização de uma matriz foi introduzida por Dwyer, em 1944.

Após obter a factorização *A=LU*, resolver *Ax=b* é equivalente a resolver dois sistemas triangulares: $\begin{cases} Ly = b \\ Ux = y \end{cases}$.

Toda esta construção foi elaborada supondo que, no passo *k*, $u_{kk} \neq 0$. A u_{kk} chamaremos *k*-ésimo pivot. De facto, nem sempre existe factorização *A=LU*. Considerando, por exemplo, a matriz $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, vemos que o método falha logo no primeiro passo, pois o primeiro pivot é nulo. De facto, o único obstáculo à existência de factorização é o aparecimento de pivots nulos. O seguinte teorema dá-nos uma condição suficiente para a existência da factorização *A=LU* (e também para a unicidade). Neste teorema, *A(1:k,1:k)* representa a sub-matriz de *A* formada pelas suas primeiras *k* linhas e colunas.

Teorema 3: *A* tem factorização *LU* se $\det\big(A(1:k,1:k)\big) \neq 0$, $\forall k \in \{1,...,n-1\}$. Caso exista, essa factorização é única.

Demonstração: Suponhamos executados *k*-1 passos do método de eliminação gaussiana. Seja $A^{(k-1)} = L_{k-1}...L_1A$. Note-se que $a_{kk}^{(k-1)}$ é o *k*-ésimo pivot. Temos que $\det(A^{(k-1)}) = \det(L_{k-1})...\det(L_1)\det(A)$. Como vimos, para $i = 1,...,k-1$, L_i é o produto de matrizes elementares L_{ij}. Logo, como $\det(L_{ij}) = 1$, temos $\det(L_i) = 1$, para $i = 1,...,k-1$. Assim, $\det(A) = \det(A^{(k-1)})$. Logo, $\det(A(1:k,1:k)) = \det(A^{(k-1)}(1:k,1:k)) = a_{11}^{(k-1)}..a_{kk}^{(k-1)}$ e, assim, se $\det(A(1:k,1:k)) \neq 0$, temos que o *k*-ésimo pivot é não-nulo, existindo então factorização.

Quanto à unicidade, suponhamos que $A = L_1U_1 = L_2U_2$. Temos que $L_2^{-1}L_1 = U_2U_1^{-1}$. Ora, $L_2^{-1}L_1$ é triangular inferior e $U_2U_1^{-1}$ é triangular superior. Logo, $L_2^{-1}L_1 = I = U_2U_1^{-1} \Rightarrow L_1 = L_2, U_1 = U_2$.

É necessário analisar o que acontece no cálculo computacional da factorização *A=LU*. O aparecimento de pivots muito pequenos pode ter

consequências catastróficas a nível numérico. O seguinte resultado dá-nos um majorante para o erro cometido no cálculo numérico da factorização.

Teorema 4: Seja *A* uma matriz de dimensão *n.* Se não ocorrerem pivots nulos, então as matrizes $\hat{L}$ e $\hat{U}$ obtidas computacionalmente satisfazem $\hat{L}\hat{U} = A + H$, com $\|H\| \le 3(n-1)\varepsilon_M (\|A\| + \|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)$.

A demonstração deste resultado pode ser encontrada em [GoLo].

É importante analisar a qualidade da solução de *Ax=b* obtida computacionalmente pelo método de eliminação gaussiana. O seguinte resultado garante-nos que a solução obtida é a solução exacta, mas de um sistema de equações aproximado.

Teorema 5: Sejam $\hat{L}$ e $\hat{U}$ os elementos da factorização *A=LU* obtidos computacionalmente. Se os sistemas $\hat{L}y = b$ e $\hat{U}x = y$ forem resolvidos computacionalmente por substituição para a frente e para trás, respectivamente, então a solução obtida $\hat{x}$ satisfaz $(A+E)\hat{x} = b$, com $\|E\| \le n\varepsilon_M \left(3\|A\| + 5\|\hat{L}\|.\|\hat{U}\|\right) + O(\varepsilon_M^2)$.

Demonstração: Sabemos que:

$$\left(\hat{L} + E_1\right)\hat{y} = b, \quad \|E_1\| \le n\varepsilon_M \|\hat{L}\| + O(\varepsilon_M^2)$$

$$\left(\hat{U} + E_2\right)\hat{x} = \hat{y}, \quad \|E_2\| \le n\varepsilon_M \|\hat{U}\| + O(\varepsilon_M^2)$$

Assim,

$\left(\hat{L} + E_1\right)\left(\hat{U} + E_2\right)\hat{x} = b \Leftrightarrow \left(\hat{L}\hat{U} + \hat{L}E_2 + E_1\hat{U} + E_1E_2\right)\hat{x} = b$. Pelo teorema anterior, sabemos que $\hat{L}\hat{U} = A + H$, com $\|H\| \le 3(n-1)\varepsilon_M (\|A\| + \|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)$. Logo, $\left(A+H+\hat{L}E_2+E_1\hat{U}+E_1E_2\right)\hat{x}=b$.

Sendo $E = H + \hat{L}E_2 + E_1\hat{U} + E_1E_2$, temos que $(A+E)\hat{x} = b$.

$$\begin{aligned}\|E\| &\le \|H\| + \|\hat{L}\|.\|E_2\| + \|E_1\|.\|\hat{U}\| + \|E_1\|.\|E_2\| \\ &\le 3(n-1)\varepsilon_M \|A\| + 3(n-1)\varepsilon_M \|\hat{L}\|.\|\hat{U}\| + n\varepsilon_M \|\hat{L}\|.\|\hat{U}\| + n\varepsilon_M \|\hat{L}\|.\|\hat{U}\| + O(\varepsilon_M^2) \\ &\le 3n\varepsilon_M (\|A\| + \|\hat{L}\|.\|\hat{U}\|) + 2n\varepsilon_M \|\hat{L}\|.\|\hat{U}\| + O(\varepsilon_M^2) = n\varepsilon_M (3\|A\| + 5\|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)\end{aligned}$$

2.3. Pivotagem parcial

O aparecimento de pivots nulos ou muito pequenos no método de eliminação gaussiana tem efeitos catastróficos a nível numérico, levando em imensos casos a elevada instabilidade. Esta instabilidade pode (na maior parte dos casos) ser eliminada escolhendo, em cada passo, para pivot o maior elemento disponível. Para tal, depois de identificar tal elemento, basta efectuar

uma simples troca de linhas e/ou colunas. A esta técnica chamamos **pivotagem**.

Se, no *k*-ésimo passo, considerarmos como candidatos a pivot todos os elementos da matriz, temos $O((n-k)^2)$ elementos a examinar de modo a determinar o máximo. Isto leva a que o custo total da escolha de pivots seja $O(n^3)$ operações. Esta técnica de pivotagem chama-se **pivotagem total**.

Na prática, é possível escolher bons pivots de entre número muito menor de pivots. A técnica de pivotagem mais utilizada é a chamada **pivotagem parcial**. Consiste em, no passo *k*, escolher o pivot entre os elementos da coluna *k* que se encontram abaixo da diagonal, ou seja, os elementos a_{ik}, $i = k,...,n$. Esta técnica tem um custo de *O(n-k)* operações em cada passo, tendo um custo total de $O(n^2)$ operações.

O processo descrito anteriormente corresponde a, em cada passo, após determinar o pivot, pré-multiplicar a matriz por uma matriz elementar P_{ij} e só depois passar à eliminação gaussiana.

O método de eliminação gaussiana com pivotagem parcial pode ser descrito pelo seguinte algoritmo:

Algoritmo do método de eliminação gaussiana com pivotagem parcial

U=A, L=I, P=I

Para *j* de 1 até *n*-1 fazer

Escolher $i \geq j$ tal que $\left|u_{i,j}\right| = \max_{j \leq i \leq n} \left|u_{ij}\right|$

$u_{j,j:n} \leftrightarrow u_{i,j:n}$

$l_{j,1:j-1} \leftrightarrow l_{i,1:j-1}$

$p_{j,:} \leftrightarrow p_{i,:}$

Para *k* de *j*+1 até *n* fazer

$$l_{kj} = \frac{u_{kj}}{u_{jj}}$$

$$u_{k,j:n} = u_{k,j:n} - l_{kj} u_{j.j:n}$$

O método de eliminação gaussiana com pivotagem parcial não fornece uma factorização *A=LU*, mas uma factorização análoga:

Proposição 6: A eliminação gaussiana com pivotagem parcial, quando aplicada a uma matriz *A*, fornece uma factorização *PA=LU*, com:

- *P* uma matriz de permutação
- *L* uma matriz triangular inferior com 1's ao longo da diagonal e $\left|l_{ij}\right| \leq 1, \forall i,j$
- *U* uma matriz triangular superior

Demonstração: A eliminação gaussiana pode ser descrita do seguinte modo: $L_{n-1}P_{n-1}...L_1P_1A = U$. Definamos:

$\hat{L}_{n-1} = L_{n-1}$

$\hat{L}_k = P_{n-1}...P_{k+1}L_kP_{k+1}...P_{n-1}$, $k \leq n-2$

Temos que:

$$\hat{L}_{n-1}...\hat{L}_1P_{n-1}...P_1A = L_{n-1}(P_{n-1}L_{n-2}P_{n-1})...(P_{n-1}...P_2L_1P_2...P_{n-1})P_{n-1}...P_1A$$
$$= L_{n-1}P_{n-1}...L_1P_1A$$

$$\Leftrightarrow \hat{L}_{n-1}...\hat{L}_1P_{n-1}...P_1A = U$$

Sendo $L^{-1} = \hat{L}_{n-1}...\hat{L}_1$ e $P = P_{n-1}...P_1$, temos que $L^{-1}PA = U$, logo *PA=LU*.

Cada $\hat{L}_k$ é obviamente uma matriz triangular inferior, logo L^{-1} e, portanto, *L* também o são. Como utilizamos pivotagem parcial, as entradas de $\hat{L}_k$ são, em valor absoluto, não superiores a 1. Como inverter L_k é trocar o sinal às entradas que se encontram abaixo da diagonal, temos que as entradas de L_k^{-1} são, em valor absoluto, não superiores a 1. Logo, o mesmo se passa para $\hat{L}_k^{-1}$ e, portanto, para *L*.

Como consequência deste resultado, vemos que aplicar o método de eliminação gaussiana com pivotagem parcial a uma matriz *A* é equivalente a aplicar o método de eliminação gaussiana sem pivotagem à matriz *PA*, resolvendo depois o sistema *PAx=Pb*. É claro que na prática isto não é aplicável, visto não termos acesso à matriz P à priori. Mas do ponto de vista teórico é bastante útil, como nos mostra o seguinte resultado, consequência deste facto e do teorema 5.

Proposição 7: Assumindo que não ocorrem erros de arredondamento na permutação de linhas, a solução $\hat{x}$ de *Ax=b* obtida computacionalmente pelo método de eliminação gaussiana com pivotagem parcial satisfaz $(A+E)\hat{x} = b$, com $\|E\| \leq n\varepsilon_M(3\|A\| + 5n\|\hat{U}\|) + O(\varepsilon_M^2)$.

Demonstração: Como vimos, aplicar o método de eliminação gaussiana com pivotagem parcial a uma matriz *A* é equivalente a aplicar o método de eliminação gaussiana sem pivotagem à matriz *PA*. Assim, pelo teorema 3, temos que $\hat{x}$ satisfaz $(PA+E_1)\hat{x} = Pb$, com $\|E_1\| \leq n\varepsilon_M(3\|PA\| + 5\|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)$. Ora, $\|P\| = 1$. Logo, $\|E_1\| \leq n\varepsilon_M(3\|A\| + 5\|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)$.

$(PA+E_1)\hat{x} = Pb \Leftrightarrow (A+P^TE_1)\hat{x} = b$. Sendo $E = P^TE_1$, temos $(A+E)\hat{x} = b$.

$\|E\| \le \|P^T\|.\|E_1\| = \|E_1\| \le n\varepsilon_M (3\|A\| + 5\|\hat{L}\|.\|\hat{U}\|) + O(\varepsilon_M^2)$. Como as entradas de $\hat{L}$ são, em valor absoluto, não superiores a 1, temos que $\|\hat{L}\| \le n$, obtendo, assim, o resultado pretendido.

2.4. Factor de crescimento

James Wilkinson proporcionou um grande avanço no estudo da estabilidade do método de eliminação gaussiana com pivotagem parcial ao reduzir esse estudo à análise do crescimento dos elementos da matriz ao longo do processo. É um resultado importantíssimo, que veremos mais adiante. Antes de mais, introduzimos a definição do objecto que mede o referido crescimento.

Definição 8: O **factor de crescimento** de uma matriz *A* com factorização $PA=LU$ é

$$\rho(A) = \frac{\max_{i,j} |u_{ij}|}{\max_{i,j} |a_{ij}|}$$

O facto de estarmos a trabalhar com pivotagem parcial implica que os multiplicadores são, em valor absoluto, não superiores a 1, o que leva a que o factor de crescimento seja limitado:

Teorema 9: Seja $A \in \Re^{n\times n}$ com factorização $PA = LU$. Então $\rho \le 2^{n-1}$. Além disso, sendo $\hat{A}(k), k = 1,...,n-1$, a matriz obtida a partir de A após o k-ésimo passo (incluindo a k-ésima pivotagem) da eliminação gaussiana com pivotagem parcial, temos que $\max_{i,j} |\hat{a}(k)_{ij}| \le 2^k \max_{i,j} |a_{ij}|$.

Demonstração: Seja $A(k)$ a matriz formada pelas últimas $n-k$ linhas e colunas de $\hat{A}(k)$, i.e., $a_{i,j}(k) = \hat{a}_{i+k,j+k}(k), i,j = 1,...,n-k$.

O *k*-ésimo passo da eliminação gaussiana (após efectuar a pivotagem) mantém inalterada a k-ésima linha da matriz obtida após os passos anteriores. Vemos, assim, que as entradas não nulas da k-ésima linha de U são as entradas da primeira linha de $A(k-1)$. Logo, temos que $\max_{i,j} |u_{i,j}| = \max\{\max_{1\le i\le n-k} |a_{1,i}(k)|, k = 0,...,n-1\}$.

Consideremos, agora, a transformação T que ao input $X \in C_{m\times m}$ faz corresponder o output $Y \in C_{m-1\times m-1}$ obtido do seguinte modo: Y é a matriz formada pelas últimas $m-1$ linhas e colunas da matriz obtida aplicando a X o primeiro passo da eliminação gaussiana com pivotagem parcial. Note-se que $A(k) = T(A(k-1)), k = 1,...n-1$.

Mostremos que, se $Y = T(X)$, então $\max_{i,j}|y_{i,j}| \le 2\max_{i,j}|x_{i,j}|$.

Suponhamos, sem perda de generalidade, que $\max_{i}|x_{i,1}| = |x_{1,1}|$.

Sendo $\alpha_i = \dfrac{x_{i,1}}{x_{1,1}} \le 1, i = 2,..,m$, temos que:

$y_{i,j} = x_{i+1,j+1} - \alpha_{i+1}x_{1,j+1}$, $i, j = 1,...,m-1$.

Assim, $\max|y_{i,j}| = \max|x_{i+1,j+1} - \alpha_{i+1}x_{1,j+1}|$

$\Rightarrow \max|y_{i,j}| \le \max|x_{i+1,j+1}| + \max\{|\alpha_{i+1}| \cdot |x_{1,j+1}|\}$

Temos que $|\alpha_i| \le 1$. Logo:

$\Rightarrow \max|y_{i,j}| \le \max|x_{i+1,j+1}| + \max|x_{1,j+1}|$

$\Rightarrow \max|y_{i,j}| \le 2\max|x_{i,j}|$

Como $A(k) = T(A(k-1))$, temos

$\max|a_{i,j}(k)| \le 2\max|a_{i,j}(k-1)|$, $i = 1,...,n-1$. Então:

$\max|a_{i,j}(n-1)| \le 2\max|a_{i,j}(n-2)| \le 4\max|a_{i,j}(n-3)| \le ... \le 2^{n-1}\max|a_{i,j}(0)| = 2^{n-1}\max|a_{i,j}|$

Logo, temos que $\max|a_{i,j}(k)| \le 2^k \max|a_{i,j}|$, $k = 1,...,n-1$.

Assim, como $\max_{i,j}|u_{i,j}| = \max\{\max_{1\le i\le n-k}|a_{1,i}(k)|, k = 0,...,n-1\}$, temos:

$\max|u_{i,j}| \le 2^{n-1}\max|a_{i,j}| \Rightarrow \rho(A) \le 2^{n-1}$.

Note-se que este resultado diz-nos que se $\rho(A) = 2^{n-1}$, então $\max_{i,j}|u_{ij}| = |u_{nn}| = 2^{n-1}\max_{i,j}|a_{i,j}|$.

Um factor de crescimento da ordem de 2^n leva a uma perda de *n* bits de precisão, o que é intolerável, visto que um computador representa um número com 64 bits em precisão dupla, e é muitas vezes necessário resolver sistemas de dimensão da ordem das centenas e milhares.

O que acontece é que a majoração obtida anteriormente para o factor de crescimento não pode ser melhorada. Vejamos: considere-se a seguinte matriz:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix}$$

No primeiro passo do método, obtemos $\begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & -1 & 1 & 0 & 2 \\ 0 & -1 & -1 & 1 & 2 \\ 0 & -1 & -1 & -1 & 2 \end{pmatrix}$.

Prosseguindo com o método (note-se que não ocorre pivotagem em nenhum passo, logo *P*=*I*), obtemos: $U=\begin{pmatrix}1&0&0&0&1\\0&1&0&0&2\\0&0&1&0&4\\0&0&0&1&8\\0&0&0&0&16\end{pmatrix}$ e $L=\begin{pmatrix}1&0&0&0&0\\-1&1&0&0&0\\-1&-1&1&0&0\\-1&-1&-1&1&0\\-1&-1&-1&-1&1\end{pmatrix}$.

O factor de crescimento é o máximo possível: $\rho(A)=2^4$. De facto, é fácil ver que, se *A* é uma matriz de dimensão *n* da forma $\begin{pmatrix}c&0&\dots&0&c\\-c&c&\dots&0&c\\\dots&\dots&\dots&\dots&\dots\\-c&-c&\dots&c&c\\-c&-c&\dots&-c&c\end{pmatrix}$, então

$U=\begin{pmatrix}c&0&\dots&0&c\\0&c&\dots&0&2c\\\dots&\dots&\dots&\dots&\dots\\0&0&\dots&c&2^{n-2}c\\0&0&\dots&0&2^{n-1}c\end{pmatrix}$ e $L=\begin{pmatrix}1&0&\dots&0&0\\-1&1&\dots&0&0\\\dots&\dots&\dots&\dots&\dots\\-1&-1&\dots&1&0\\-1&-1&\dots&-1&1\end{pmatrix}$, obtendo-se $\rho(A)=2^{n-1}$.

Vejamos o que acontece na resolução computacional de *Ax*=*b* pelo método de eliminação gaussiana com pivotagem parcial quando *A* tem esta forma (com *c*=1 e *dim(A)*=60).

Quando $b=\vec{0}$, a solução obtida computacionalmente é $\hat{x}=\vec{0}$. Apesar do factor de crescimento máximo, esta é a solução exacta. Vejamos: *b* não é alterado ao longo do processo, pois *b* é o vector nulo. Logo, o sistema triangular superior a resolver é:

$$Ux=\vec{0}\Leftrightarrow\begin{cases}x_1+x_{60}=0\\x_2+2x_{60}=0\\\quad\dots\\2^{59}x_{60}=0\end{cases}\Leftrightarrow x=\vec{0}.$$

De facto, é de esperar que não ocorram erros: *b* é o vector nulo e, ao longo do processo, às entradas de *b* é somado 0, mantendo-se, assim, sempre o mesmo vector, ou seja, sem erros. Assim, o sistema triangular superior a resolver por substituição para trás é $\hat{U}x=\vec{0}$. Assim, $\hat{x}_{60}=fl\left(\frac{0}{\hat{U}_{60,60}}\right)=0$ e, para *i*=1,...,59, $\hat{x}_i=fl\left(\frac{-\sum_{j=i+1}^{60}\hat{U}_{ij}\hat{x}_j}{\hat{U}_{ii}}\right)=fl\left(\frac{-\hat{U}_{i,60}\hat{x}_{60}}{\hat{U}_{ii}}\right)=fl\left(\frac{0}{\hat{U}_{ii}}\right)=0$.

Quando $b=\vec{1}$ ($\vec{1}=(1,\dots,1)^T$), a solução obtida computacionalmente é $\hat{x}=(0,\dots,0,1)^T$. Novamente, trata-se da solução exacta. Vejamos: a solução de

Ly=b é $y_i = 2^{i-1}$. Mostremos isto por indução. Ora, obviamente $y_1 = 1$. Suponhamos que $y_{i-1} = 2^{i-2}$. Temos que $y_i = b_i + \sum_{k=1}^{i-1} y_k = b_{i-1} + y_{i-1} + \sum_{k=1}^{i-2} y_k$. Como $y_{i-1} = b_{i-1} + \sum_{k=1}^{i-2} y_k$, temos $y_i = 2y_{i-1} = 2^{i-1}$. Logo, o sistema *Ux=y* é

$$\begin{cases} x_1 + x_{60} = 1 \\ x_2 + 2x_{60} = 2 \\ \quad \ldots \\ 2^{59} x_{60} = 2^{59} \end{cases}$$, donde $x_{60} = 1$ e, para *i*=1,...,59, $x_i = 2^{i-1} - 2^{i-1} x_{60} = 0$.

Novamente, a não ocorrência de erros é natural: *b* é a ultima coluna de *A*, logo, ao longo do processo, *b* será sempre igual à última coluna de *A.* Assim, o sistema triangular superior a resolver por substituição para trás é $\hat{U}x = \hat{d}$, onde $\hat{d}$ é a última coluna de $\hat{U}$. Assim, $\hat{x}_{60} = fl\left(\frac{\hat{U}_{60,60}}{\hat{d}_{60,60}}\right) = fl\left(\frac{\hat{U}_{60,60}}{\hat{U}_{60,60}}\right) = 1$

e $\hat{x}_i = fl\left(\frac{\hat{d}_i - \sum_{j=i+1}^{60} \hat{U}_{ij}\hat{x}_j}{\hat{U}_{ii}}\right) = fl\left(\frac{\hat{U}_{i,60} - \hat{U}_{i,60}\hat{x}_{60}}{\hat{U}_{ii}}\right) = fl\left(\frac{\hat{U}_{i,60} - \hat{U}_{i,60}}{\hat{U}_{ii}}\right) = fl\left(\frac{0}{\hat{U}_{ii}}\right) = 0$,

i=1,...,59.

No primeiro caso, o vector *b* não é alterado ao longo do processo, não ocorrendo, assim, acumulação de erros e, como se trata do vector nulo, não ocorrem erros na substituição para trás. No segundo caso, o vector *b* é a última coluna de *A*, logo a acumulação de erros em *b* é igual à acumulação de erros na última coluna de $\hat{U}$, sendo essa acumulação anulada ao efectuar substituição para trás, como vimos anteriormente. É, assim, natural ver o que acontece quando *b* não é de nenhuma destas formas.

Consideremos $b = A \times \vec{1}$. A solução obtida computacionalmente é $x_i = \begin{cases} 0, \, se \; i \in \{54,...,59\} \\ 1, \, se \; i \notin \{54,...,59\} \end{cases}$. Agora, ocorreram erros. De facto, a solução exacta é $x = (1,...,1)^T$. Vejamos: a *i*-ésima entrada de *b* é a soma dos elementos da *i*-ésima linha de *A.* Assim, $b_i = \begin{cases} 3 - i, \, i = 1,...,59 \\ -58, \, i = 60 \end{cases}$. Mostremos, por indução, que a solução de *Ly=b* é $y_i = \begin{cases} 2^{i-1} + 1, \, i = 1,..,59 \\ 2^{59}, \, i = 60 \end{cases}$. Obviamente, $y_1 = 2$. Suponhamos que $y_{i-1} = 2^{i-2} + 1$. Temos que $y_i = b_i + \sum_{k=1}^{i-1} y_k = b_{i-1} - 1 + y_{i-1} + \sum_{k=1}^{i-2} y_k$. Como $y_{i-1} = b_{i-1} + \sum_{k=1}^{i-2} y_k$, temos $y_i = 2y_{i-1} - 1 = 2(2^{i-2} + 1) - 1 = 2^{i-1} + 1$. Quanto a y_{60}, temos que $y_{60} = b_{60} + \sum_{k=1}^{59} y_k = b_{59} - 2 + y_{59} + \sum_{k=1}^{58} y_k$. Como $y_{59} = b_{59} + \sum_{k=1}^{58} y_k$, temos

$y_{60} = 2y_{59} - 2 = 2^{59}$. Quanto à solução de *Ux=y*, temos $2^{59} x_{60} = y_{60} \Leftrightarrow x_{60} = 1$ e, $\forall i = 1,...,59$, $x_i + 2^{i-1} x_{60} = y_i \Leftrightarrow x_i = 2^{i-1} + 1 - 2^{i-1} = 1$. Assim, a solução exacta é $x = (1,...,1)^T$

Passemos, então, ao teorema de Wilkinson. É o resultado que permitiu reduzir o estudo da estabilidade do método ao estudo do factor de crescimento. Foi apresentado por Wilkinson num artigo pioneiro em 1961.

Teorema de Wilkinson: A solução $\hat{x}$ de *Ax=b* obtida computacionalmente pelo método de eliminação gaussiana com pivotagem parcial satisfaz $(A+E)\hat{x} = b$, com $\|E\| \leq 8n^3 \rho(A) \|A\| \varepsilon_M + O(\varepsilon_M^2)$.

Demonstração: Antes de mais, note-se que $\|U\| = \|U\|_\infty \leq n \max_{i,j} |u_{ij}|$ e $\|A\| = \|A\|_\infty \geq \max_{i,j} |a_{ij}|$. Assim, temos que $n\rho(A) \geq \frac{\|U\|}{\|A\|} \Rightarrow \|U\| \leq n\rho(A)\|A\|$. Logo, pela proposição anterior, temos que $\hat{x}$ verifica $(A+E)\hat{x} = b$, com $\|E\| \leq n\varepsilon_M (3\|A\| + 5n^2 \rho(A)\|A\|) + O(\varepsilon_M^2) \Rightarrow \|E\| \leq 8n^3 \rho(A)\|A\|\varepsilon_M + O(\varepsilon_M^2)$

Podemos, assim, ver que a solução $\hat{x}$ de $Ax=b$ obtida computacionalmente é a solução exacta de um sistema aproximado com erro relativo $\frac{\|E\|}{\|A\|} \leq 8n^3 \rho(A)\varepsilon_M + O(\varepsilon_M^2)$. Vemos, assim, que se o factor de crescimento não for muito elevado, a solução obtida é bastante boa. Em contrapartida, se o factor de crescimento for muito elevado, devemos esperar instabilidade.

O que acontece é que, apesar de exemplos como o analisado anteriormente, em 50 anos de computação, não se conhece nenhuma matriz com factor de crescimento muito elevado que seja proveniente de um problema real. Isto é, todas as matrizes que se conhece que provoquem instabilidade explosiva são exemplos construídos, mas não observáveis na prática. Isto é ainda hoje motivo de discussão, visto não estar completamente explicado. Mais adiante, veremos um estudo estatístico que evidencia este mesmo facto.

Os seguintes resultados são bastante recentes, ambos da autoria de Desmond J. Higham e Nicholas J. Higham. O primeiro mostra a forma que uma matriz deve ter para se obter o factor de crescimento máximo. Já o segundo, dá-nos um limite inferior para o factor de crescimento. Ambas as demonstrações podem ser encontradas em [Hig], nas páginas 178 e 179.

Teorema 11: Todas as matrizes A de dimensão *n* tais que $\rho(A) = 2^{n-1}$ são da forma $A = DM\begin{pmatrix} T & \\ 0 & \alpha d \end{pmatrix}$, onde $D = diag(\pm 1)$, *M* é triangular inferior

com $m_{ii} = 1$ e $m_{ij} = -1$, para *i>j*, *T* é uma matriz arbitrária não-singular de dimensão *n*-1, $d = (1,2,...,2^{n-1})^T$ e $\alpha = \max_{i,j} |a_{ij}| = |a_{1n}|$.

Teorema 12: Seja $A \in C^{n \times n}$ uma matriz não-singular e sejam $\alpha = \max_{i,j} |a_{ij}|, \beta = \max_{i,j} |(A^{-1})_{ij}|$ e $\theta = (\alpha\beta)^{-1}$. Então, $\theta \leq n$ e, para quaisquer matrizes de permutação *P* e *Q* tais que *PAQ* tem factorização *LU*, o factor de crescimento para a eliminação gaussiana sem pivotagem satisfaz $\rho(PAQ) \geq \theta$.

3. Estudo estatístico

O estudo estatístico que aqui apresentamos tem como objectivo mostrar o facto de a ocorrência de factores de crescimento elevados ser rara, de acordo com o que se vem conjecturando. Trata-se de uma tentativa de explicar o fenómeno de um modo estatístico.

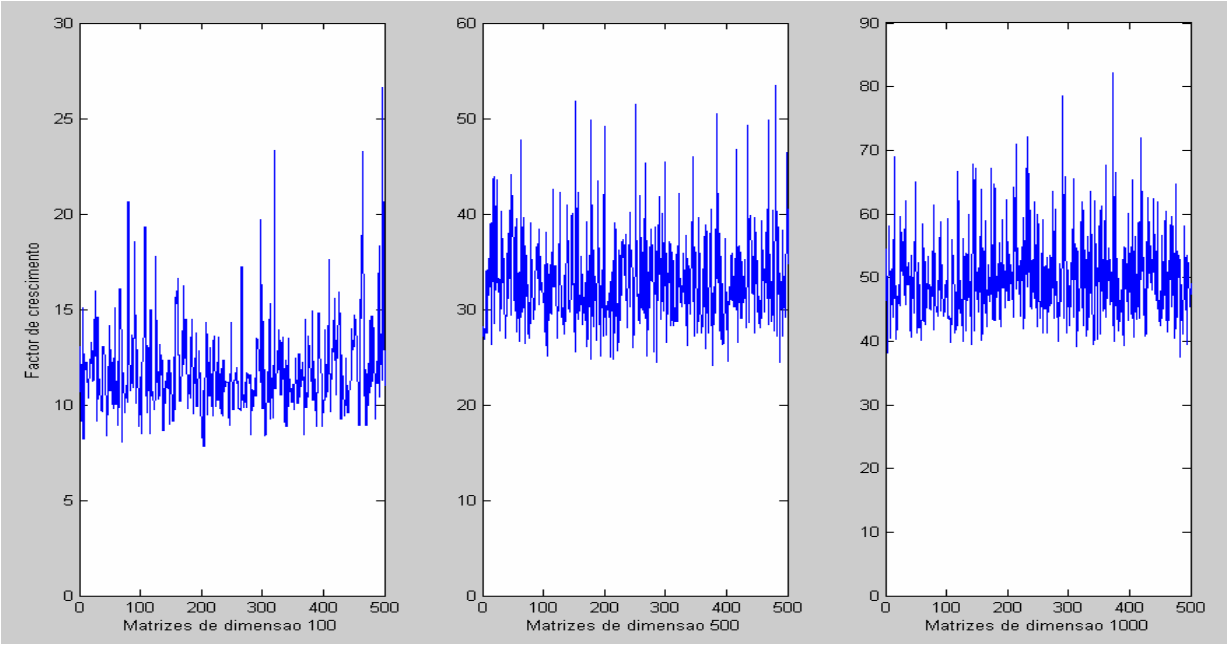
Este estudo consistiu na observação do que acontece com o factor de crescimento quando as matrizes são matrizes aleatórias. Com este objectivo, e utilizando o gerador de números aleatórios do Matlab, procedemos à geração de um número elevado de matrizes com diferentes distribuições.

É claro que as matrizes provenientes de problemas reais não têm nada de aleatório. Têm todo o tipo de propriedades especiais e, se tentássemos descrevê-las através de uma variável aleatória, teria de ser uma distribuição muito curiosa. Deste modo, estudam-se diferentes tipos de matrizes aleatórias, e tiram-se conclusões a partir daí.

Todas as experiências foram realizadas com uma amostra de 500 matrizes.

Começamos por apresentar os resultados obtidos quando utilizamos a distribuição $U(-1,1)$. Testamos matrizes de três dimensões diferentes: 100, 500 e 1000. O gráfico seguinte mostra os factores de crescimento obtidos (de modo a tornar mais visível o comportamento dos factores de crescimento, utilizamos escalas diferentes para diferentes dimensões, mantendo a mesma escala quando estudamos diferentes distribuições e dimensões iguais).

$U(-1,1)$



Obtém-se o que se esperava. O factor de crescimento é sempre bastante inferior ao máximo teórico. Este facto torna-se mais evidente analisando a seguinte tabela:

$U(-1,1)$

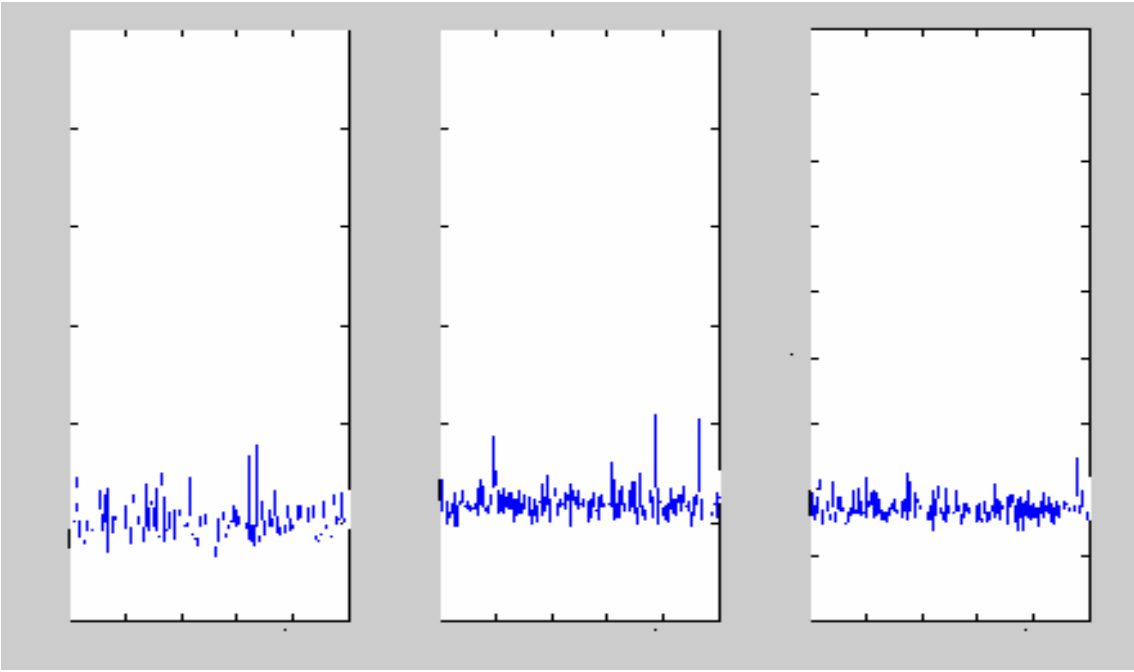
Dimensão	Máximo teórico	Máximo obtido	Mínimo obtido	Média	Desvio padrão
100	2^{99}	26.6106	7.8289	11.7080	2.4351
500	2^{499}	53.5071	24.1096	33.0806	4.8343
1000	2^{999}	82.2481	37.5197	50.0537	6.8148

A diferença entre o máximo teórico e o máximo obtido é abismal, em todas as dimensões! A média dos valores obtidos, em cada dimensão, é muita baixa e com um desvio padrão também pequeno.

Daqui podemos concluir que, em 500 matrizes de cada dimensão, não ocorreu uma que fosse provocar instabilidade explosiva no método de eliminação gaussiana com pivotagem parcial.

Passemos agora aos resultados obtidos com a distribuição $N(0,1)$. As dimensões testadas são, novamente, 100, 500 e 1000. O gráfico seguinte mostra os factores de crescimento obtidos.

$N(0,1)$



Os resultados obtidos são ainda melhores do que no caso da distribuição $U(-1,1)$. É bem visível um decréscimo nos valores dos factores de crescimento. A tabela seguinte evidencia este facto:

$N(0,1)$

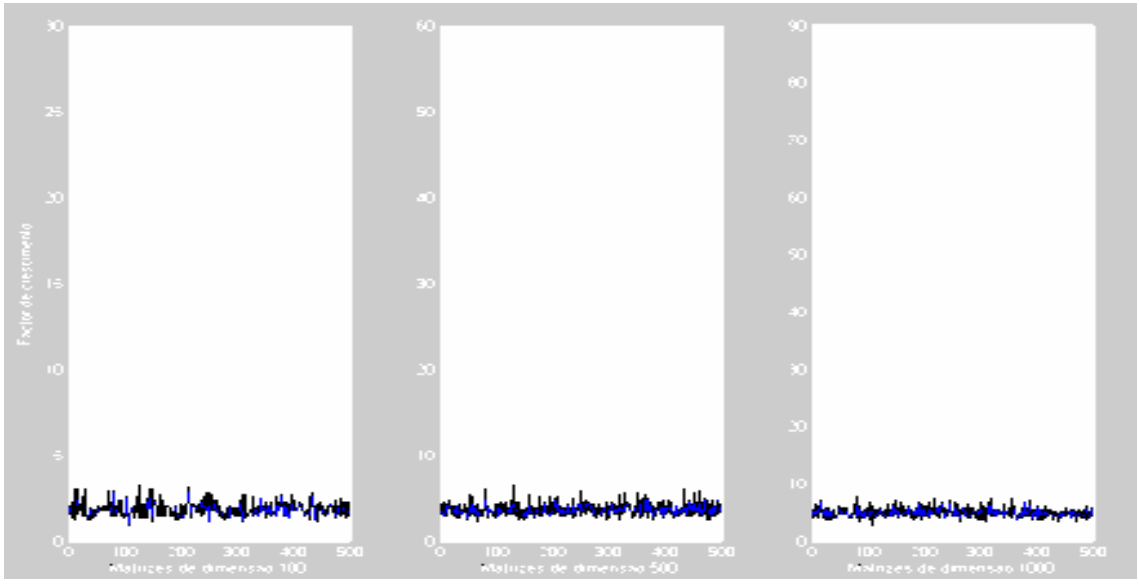
Dimensão	Máximo teórico	Máximo obtido	Mínimo obtido	Média	Desvio padrão
100	2^{99}	8.9307	3.2731	5.1061	1.0734
500	2^{499}	26.9607	8.0363	12.0885	2.1401
1000	2^{999}	32.7669	11.9791	17.2977	2.5098

De facto, todos os valores são inferiores aos observados para a distribuição $U(-1,1)$, o que leva a uma diferença ainda maior entre o máximo obtido e o máximo teórico. Note-se que a média dos valores é bastante baixa, assim como o seu desvio padrão. O que indica que o método de eliminação gaussiana com pivotagem parcial se revelou estável para todas as matrizes aleatoriamente geradas.

Novamente, em 500 matrizes de cada dimensão, não ocorre uma que leva a uma instabilidade intolerável.

Os resultados obtidos com a distribuição χ_1^2 (para as mesmas dimensões) são ainda melhores. Vejamos os resultados obtidos:

χ_1^2



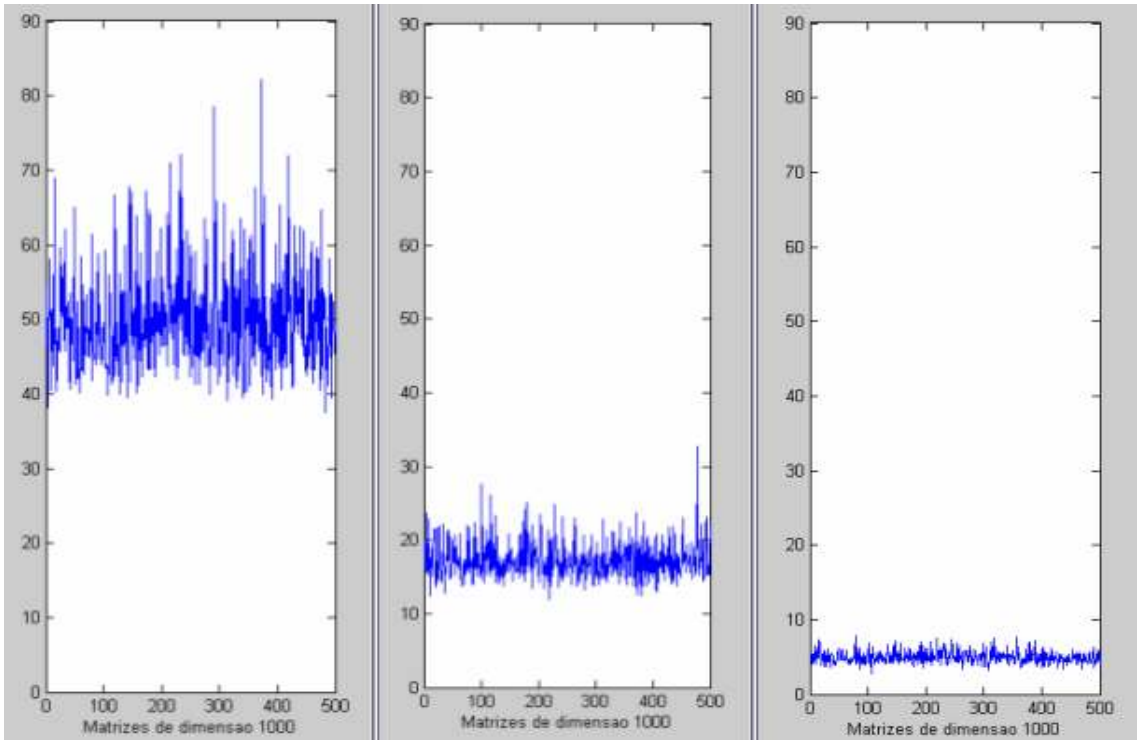
Os resultados obtidos são ainda melhores do que nos dois casos anteriores, observando-se valores muito inferiores. A seguinte tabela resume os valores obtidos para a distribuição χ_1^2 .

χ_1^2

Dimensão	Máximo teórico	Máximo obtido	Mínimo obtido	Média	Desvio padrão
100	2^{99}	3.2593	0.9677	1.8503	0.4361
500	2^{499}	6.5698	2.1060	3.6643	0.6856
1000	2^{999}	7.8743	2.7350	4.9593	0.8247

Podemos comparar os resultados obtidos para as diferentes distribuições, por exemplo, para a dimensão 1000:

$U(-1,1)$ $N(0,1)$ χ_1^2



Podemos ver que os valores obtidos mais elevados foram para a distribuição Uniforme, e que a distribuição χ_1^2 apresenta os valores mais baixos.

A seguinte tabela resume todos os resultados obtidos:

Dimensão	Máximo teórico	Distribuição	Máximo obtido	Mínimo obtido	Média	Desvio padrão
100	2^{99}	$U(-1,1)$	26.6106	7.8289	11.7890	2.4351
		$N(0,1)$	8.9307	3.2731	5.1061	1.0734
		χ^2_1	3.2593	0.9677	1.8503	0.4361
500	2^{499}	$U(-1,1)$	53.5071	24.1096	33.0806	4.8343
		$N(0,1)$	26.9607	8.0363	12.0885	2.1401
		χ^2_1	6.5698	2.1060	3.6643	0.6856
1000	2^{999}	$U(-1,1)$	82.2481	37.5197	50.0537	6.8148
		$N(0,1)$	32.7669	11.9791	17.2977	2.5098
		χ^2_1	7.8743	2.7350	4.9593	0.8247

Como podemos ver, para nenhuma distribuição nem nenhuma dimensão, ocorreu uma matriz que levasse a uma instabilidade explosiva no método de eliminação gaussiana com pivotagem parcial. O que vem de acordo com o que estávamos à espera. Actualmente, pensa-se que o conjunto de matrizes para as quais o método de eliminação gaussiana com pivotagem parcial é drasticamente instável é extraordinariamente pequeno. Ou, usando um abuso de linguagem, o conjunto de matrizes para as quais o método de eliminação gaussiana com pivotagem parcial é drasticamente instável é um conjunto de medida nula no conjunto de todas as matrizes.

Nota: Foram testadas outras distribuições, como, por exemplo, uniformes em intervalos de diferente amplitude e normais com diferentes médias e variâncias, mas os resultados são bastante semelhantes aos aqui apresentados, optando-se, deste modo, por não os exibir.

4. Conclusão

Este estudo estatístico mostra-nos que, para todas as distribuições testadas, o método de eliminação gaussiana com pivotagem parcial se revelou estável para todas as matrizes geradas aleatoriamente. De realçar o facto de, no total de 4500 matrizes, não aparecer uma única cujo factor de crescimento fosse elevado. Estes resultados estatísticos apontam na direcção do que se vem pensando acerca do factor de crescimento: as matrizes com factor de crescimento elevado (e, portanto, para as quais o método de eliminação gaussiana se revela instável) formam um conjunto extraordinariamente pequeno no conjunto de todas as matrizes.

Contudo, considerando apenas matrizes que são computáveis (isto é, matrizes cujas entradas são elementos do sistema de numeração utilizado), não é verdade que o conjunto de matrizes para as quais o método de eliminação gaussiana com pivotagem parcial se revela instável tenha medida nula: o conjunto das matrizes computáveis é finito e conhecemos matrizes para as quais o método se revela instável. De qualquer modo, o nosso estudo leva-nos a conjecturar que a probabilidade de ocorrer uma matriz para a qual o método se revela instável, apesar de não ser nula, é bastante reduzida.

5. Bibliografia

- **[BaTr]** – Bau III, D; Trefethen, L.N.; “Numerical Linear Algebra”; SIAM; 1997
- **[GoLo]** – Golub, H.G.; Van Loan, C.F.; “Matrix Computations”; The John Hopkins University Press; 1996
- **[Hig]** – Higham, N.J.; “Accuracy and Stability of Numerical Algorithms”; SIAM; 1996
- **[ReWil] –** Reinsch, C.; Wilkinson, J.H.; “Linear Algebra”; Springer;1971
- **[Val]** – Valença, M.R.; “Métodos Numéricos”; Instituto Nacional de Investigação Cientifica; 1988
- **[Wil]** – Wilkinson, J.H.; “Rounding errors in algebraic processes”; National Physical Laboratory; 1963