Universidade Federal da Paraíba
Centro de Ciências e Tecnologia
Coordenação de Pós-Graduação em Informática

Estudo e Implementação de Métodos Diretos para Solução Exata de Sistemas Lineares

Adeilton Fernandes da Costa

Campina Grande - PB
1998

Adeilton Fernandes da Costa

Estudo e Implementação de Métodos Diretos para Solução Exata de Sistemas de Equações Lineares

Dissertação apresentada ao Curso de Mestrado em Informática da Universidade Federal da Paraíba, como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: Ciência da Computação

Linha de pesquisa: Matemática Computacional

Orientador: Mário Toyotaro Hattori
Co-Orientador: João Marques de Carvalho

Campina Grande
Universidade Federal da Paraíba

ESTUDO E IMPLEMENTAÇÃO DE MÉTODOS DIRETOS PARA SOLUÇÃO EXATA DE SISTEMAS LINEARES

ADEILTON FERNANDES DA COSTA

DISSERTAÇÃO APROVADA EM 17.07.1998

p/ PROF. MÁRIO TOYOTARO HATTORI, M.Sc
Orientador

PROF. JOÃO MARQUES DE CARVALHO, Ph.D
Co-Orientador

PROF. BRUNO CORREIA DA NÓBREGA QUEIRÓZ, M.Sc.
Examinador

PROF. MAURO CAVALCANTE PEQUENO, D.Sc
Examinador

CAMPINA GRANDE - PB

Ficha Catalográfica

Costa, Adeilton Fernandes

C837E

Estudo e Implementação de Métodos Diretos para Solução Exata de Sistemas Lineares - Campina Grande: CCT/COPIN da UFPB, 1998, 73 p.

Dissertação (mestrado) - Universidade Federal da Paraíba, Centro de Ciências e Tecnologia, Coordenação de Pós-graduação em Informática, Campina Grande, 1998.

Orientador: Mário Toyotaro Hattori
Co-Orientador: João Marques de Carvalho

1- Sistemas de Equações Lineares 2- Fatoração *LU* 3- Eliminação de Gauss 4 - Eliminação de Jordan 5- Solução Exata de Sistemas Lineares 6- Aritmética em Múltipla Precisão I -Título.

CDU - 519.612

Dedicatória

A Deus, por ter lançado o desafio e me fortalecido nos momentos de dificuldades;

Aos meus pais, Augusto e Adelzuita, exemplos de perseverança;

A minha esposa, Socorro, pela demonstração de carinho, compreensão e confiança;

Aos meus filhos, Ícaro e Iuane, continuação da minha vida.

"Bem aventurado o homem que põe no
Senhor a sua confiança ..."

(Sl 40:4)

Agradecimentos

Aos meus irmãos e cunhadas pelos poucos momentos felizes em que estivemos juntos.

Ao meu orientador prof. Mário Toyotaro Hattori (in memorian), pela confiança, critica, paciência e todo o esforço feito para que esse trabalho se tornasse uma realidade.

Ao prof. Bruno pela enorme contribuição durante a fase final desse trabalho.

Aos funcionários do Departamento de Sistemas e Computação. Alberto, Aninha, Manuela, Vera e Zeneide.

Aos professores João Marques, Agammenon, Fernedá, Homero, Jacques, Peter, e Valfredo.

Aos colegas Carlos, Jean, Kátia, Kíssia, Mário Ernesto, Salete, etc. pela amizade conquistada.

Ao Departamento de Ciências Exatas da Universidade Federal de Rondônia e a todos que durante esses anos colaboraram direta e indiretamente na realização desse trabalho.

RESUMO

A proposta desse trabalho é implementar um método para solução exata de sistemas lineares, nos racionais, cujos coeficientes são números inteiros, utilizando um pacote de aritmética de ponto flutuante em multipla precisão.

O trabalho apresenta um estudo dos métodos diretos de fatoração *LU*, eliminação de Gauss e eliminação de Jordan e em seguida um estudo detalhado do método proposto por Fox para solução de sistemas lineares, sem erros de arredondamento.

Finalmente são apresentados alguns resultados obtidos pelos métodos diretos e proposto por Fox.

ABSTRACT

The purpose of this work is to implement method for exact solution of linear systems, over Q, with integer coefficients, using the package of arithmetic floating point on precision multiple.

The work present a study of the direct methods of *LU* fatoraction, Jordan elimination and Gauss elimination followed by a detailed study of the method proposed by Fox for solution of linear systems, without rounding errors.

Finally some results obtained by the direct methods and proposed by Fox are presented.

SUMÁRIO

RESUMO		**vi**
ABSTRACT		**vii**
Lista de Tabelas		**x**
CAPÍTULO I	**Introdução**	**01**
	1.1 Introdução	01
	1.2 Objetivo do trabalho	02
	1.3 Estrutura da dissertação	03
CAPÍTULO II	**Sistemas Lineares e Matrizes**	**04**
	2.1 Introdução	04
	2.2 Sistema Linear	04
	2.2.1 Solução de um sistema linear	05
	2.2.2 Discusão de um sistema linear	06
	2.2.3 Interpretação geométrica de um sistema linear	06
CAPÍTULO III	**Erros em Computação**	**08**
	3.1 Introdução	08
	3.2 Origem e conceito dos erros	09
	3.3 Propagação dos erros	10
	3.4 Aritmética de ponto flutuante	11
	3.5 Overflow e Underflow	13
	3.6 Operações em ponto flutuante	14
CAPÍTULO IV	**Métodos Diretos**	**16**
	4.1 Introdução	16
	4.2 Fatoração *LU*	17

4.2.1 Algoritmo da fatoração LU 20
4.2.2 Algoritmo para substituição progressiva 20
4.2.3 Algoritmo para substituição regressiva 21
4.3 Eliminação de Gauss 21
4.3.1 Algoritmo da Eliminação de Gauss 24
4.4 Eliminação de Jordan 25
4.4.1 Algoritmo da Eliminação de Jordan 28
4.5 Erros na solução de sistemas lineares 29
4.6 Pivotamento 33
CAPÍTULO V Solução Exata de Sistemas Lineares 36
5.1 Introdução 36
5.2 Solução Exata de Sistemas Lineares 37
5.2.1 Algoritmo do método proposto por Fox 47
5.2.2 Algoritmo do Vetor Auxiliar e do Vetor Solução do Sistema Linear 48
5.2.3 Algoritmo de Euclides 48
5.3 Aritmética em múltipla precisão 51
5.4 Cálculo da Precisão 53
CAPÍTULO VI Testes e Resultados 57
6.1 Introdução 57
6.2 Descrição das matrizes 57
6.3 Observações sobre implementação 59
6.4 Resultados dos testes 59
CAPÍTULO VII Conclusões 66
Referências Bibliográficas 70

Lista de Tabelas

5.1	Número de operações realizadas pelo algoritmo (5.2.1)	49
5.2	Número de operações realizadas pelos métodos	50
5.3	Estimativa de precisão requerida para resolver o sistema linear (5.4.5)	54
6.3	Solução do sistema linear (5.2.2) pelos métodos diretos e proposto por Fox	61
6.4	Erros absolutos cometidos na obtenção da solução do sistema linear 5.2.2	62
6.5	Soluções dos sistemas lineares com estrutura (6.2.2).	63
6.6	Erros absolutos cometidos na obtenção das soluções da tabela 6.5	64
6.7	Tempo de execução, em segundos, em sistemas lineares com estrutura 6.2.2	65
7.1	N° máximo de equações de um sistema linear simétrico a ser resolvido numa precisão de 50 dígitos segundo o número m	67

Capítulo I

Introdução

1.1 Introdução

Na computação científica existem problemas cujos dados são exatos e sabe-se que existe uma solução exata. Usando aritmética convencional (disponível no computador), a solução computada desses problemas nem sempre é exata. Quando se sabe que um problema tem solução exata e há necessidade de obtê-la o tempo de processamento passa a ter importância secundária. Agora, como obter essa solução exata? [Schreiner92].

O interesse em encontrar uma solução exata de sistemas lineares, via computador, tem crescido [Lipson81]. Algoritmos para solução exata de sistemas lineares com coeficientes inteiros ou racionais são apresentados em [McClellan73, McClellan77, Cabay77]. Os métodos diretos de Fatoração *LU*, Eliminação de Gauss e Eliminação de Jordan [Golub96, Hattori94, Dorn72] ainda são inadequados na computação exata.

A dificuldade na obtenção da solução exata por esses métodos surge quando o resultado de uma operação aritmética supera o limite da palavra no computador. A dificuldade maior é na execução da operação de divisão, pois o resultado desta operação pode não ser exato. Uma maneira de superar essa dificuldade seria utilizar um método que evite divisões ou que os resultados das divisões sejam exatos [Fox64].

Se a operação de divisão for indispensável, deve-se usar *aritmética racional* ou *aritmética dos resíduos* (escala os números e mapeia no intervalo [0, m-1] em que m é o módulo). Se o resultado de uma operação aritmética for exata mas supera o limite do computador, pode-se usar uma *aritmética inteira em múltipla precisão* ou *aritmética de ponto flutuante em múltipla precisão*, nesta, se o computador dispuser de precisão suficiente, a conversão de um número inteiro para ponto flutuante será exata [Knuth69, Bailey93, Figueiredo89].

O cálculo em múltipla precisão, que extrapola a precisão disponível em hardware, vem sendo estudado na ciência da computação desde que foram introduzidos os primeiros modelos de computadores. O estudo de constantes matemáticas é uma das áreas em que a múltipla precisão é amplamente utilizada [Bailey93, Brent78, Smith91]. A solução exata de sistemas lineares é outra área de aplicação [Schreiner92].

Como exemplos de áreas que envolvem problemas de solução de sistemas lineares podem ser citados os estudos de difusão de nêutrons, de escoamento de fluidos, de elasticidade, de transmissão de calor em sólidos, de previsão do tempo, além de problemas de engenharia [Young71].

1.2 Objetivo do trabalho

O objetivo deste trabalho é estudar a viabilidade do método proposto por Fox para obtenção da solução exata de sistemas lineares, implementar esse método e comparar os resultados obtidos com os resultados obtidos pelos métodos diretos de fatoração *LU*, eliminação de Gauss e eliminação de Jordan.

Este método proposto por Fox [Fox64] é indicado para sistemas lineares em que os elementos da matriz completa do sistema linear sejam números inteiros. Todas as operações

realizadas neste método, inclusive a divisão, são realizadas sem erros de arredondamentos, isso justifica a obtenção da solução exata do sistema linear.

1.3 Estrutura da Dissertação

Este trabalho está organizado em 7 capítulos.

O capítulo I apresenta a introdução.

O capítulo II apresenta uma revisão de sistemas lineares e matrizes.

O capitulo III apresenta um estudo sobre a origem e propagação dos erros em computação e uma revisão de aritmética em ponto flutuante .

O capítulo IV apresenta uma revisão dos métodos diretos de fatoração LU, eliminação de Gauss e eliminação de Jordan com os respectivos algoritmos, um estudo dos erros na obtenção da solução exata por esses métodos e um estudo sobre pivotamento parcial e total.

O capítulo V apresenta como obter a solução exata de um sistema linear pelos métodos diretos, uma descrição do método proposto por Fox e a necessidade da utilização da aritmética de múltipla precisão para obtenção da solução exata de um sistema linear por esse método.

O capítulo VI apresenta a descrição das matrizes usadas nos testes e os resultados experimentais obtidos por esses métodos.

O capítulo VII apresenta a conclusão e sugestões para trabalhos futuros.

Capítulo II

Sistemas Lineares e Matrizes

2.1 Introdução

Muitos problemas de análise numérica podem ser resolvidos utilizando sistemas lineares. Entre esses problemas encontra-se a solução de equações diferenciais ordinárias ou parciais pelo método das diferenças finitas, os problemas de autovalores da física matemática e a aproximação polinomial. Este capítulo apresenta alguns conceitos de sistemas lineares e matrizes, os quais serão utilizadas na obtenção da solução de sistemas lineares, capítulos IV, V e VI.

2.2 Sistema Linear

É um conjunto de n equações lineares a n incógnitas $x_1, x_2, x_3, \ldots, x_n$. Assim, o sistema

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2,\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3n}x_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \ldots + a_{in}x_n = b_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n,
\end{aligned}
\qquad (2.2.1)
$$

é linear.

O sistema linear (2.2.1) pode ser escrito na forma matricial

$$A\mathbf{x} = \mathbf{b}, \tag{2.2.2}$$

em que A é a matriz dos coeficientes representados por a_{ij} com $i = 1, 2, ..., n$, $\mathbf{x}$ é o vetor incógnita e $\mathbf{b}$ é o vetor independente cujos componentes são, respectivamente, x_i e b_i, $i = 1, 2, ..., n$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} \\ \vdots & \vdots & \vdots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{pmatrix} \quad \text{e} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}. \tag{2.2.3}$$

A matriz A é chamada de *matriz incompleta* do sistema e a matriz que se obtém acrescentando a matriz A o vetor $\mathbf{b}$ é chamada de *matriz completa* do sistema e é denotada por A^*. A matriz completa do sistema (2.2.1) é a matriz de ordem $n \times (n + 1)$ dada por

$$A^* = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} & b_i \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{mi3} & \dots & a_{mn} & b_n \end{pmatrix} \tag{2.2.4}$$

2.2.1 Solução de um Sistema Linear

Diz-se que o vetor $(\propto_1, \propto_2, \propto_3, \dots, \propto_n)$ é solução de um sistema linear, se for solução de todas as equações do sistema linear,

$$
\begin{aligned}
&a_{11}\propto_1 + a_{12}\propto_2 + a_{13}\propto_3 + \ldots + a_{1n}\propto_n = b_1,\\
&a_{21}\propto_1 + a_{22}\propto_2 + a_{23}\propto_3 + \ldots + a_{2n}\propto_n = b_2,\\
&a_{31}\propto_1 + a_{32}\propto_2 + a_{33}\propto_3 + \ldots + a_{3n}\propto_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{i1}\propto_1 + a_{i2}\propto_2 + a_{i3}\propto_3 + \ldots + a_{in}\propto_n = b_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}\propto_1 + a_{n2}\propto_2 + a_{n3}\propto_3 + \ldots + a_{nn}\propto_n = b_n.
\end{aligned}
\qquad (2.2.5)
$$

2.2.2 Discussão de um sistema linear

Quanto a solução o sistema linear (2.2.1), pode ser:

Compatível (possui solução)
- *Determinado* (possui uma única solução)
- *Indeterminado* (possui infinitas soluções)

Incompatível: não possui solução

2.2.3 Interpretação geométrica de um sistema linear

No sistema linear (2.2.1) se $n = 2$, cada equação representa uma reta no $\Re^2$, se $n = 3$, cada equação representa um plano no $\Re^3$ e para $n > 3$, cada equação representa um hiperplano no $\Re^n$. Considerando

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \ldots + a_{in}x_n = b_i,$$

como sendo uma equação de um hiperplano, então, geometricamente a solução do sistema linear (2.2.1) é a inteseção desses n hiperplanos.

Esse conjunto de hiperplanos podem ser *concorrentes*, *paralelos* ou *coincidentes*, três possibilidades que correspondem respectivamente aos casos de compatibilidade, incompatibilidade e indeteminação, visto na seção 2.2.2.

No sistema linear (2.2.1) se o vetor **b** for nulo diz-se que o sistema linear é *homogêneo*, caso contrário, diz-se *não-homogêneo.*

Teorema 2.3.1 - Um sistema homogêneo de n equações a n incógnitas existe solução diferente da trivial ($x_1 = x_2 = \ldots = x_n = 0$) se, e somente se, o determinante da matriz A do sistema (2.2.1) for nulo [Steinberg74].

Teorema 2.3.2 - Um sistema não-homogêneo de n equações lineares a n incógnitas existe solução única se, e somente se, a matriz A do sistema (2.2.1) for *não singular*, ou seja, o det $A \neq 0$ [Steinberg74].

Se A é não singular, então A^{-1} (inversa da matriz A) existe e a solução do sistema (2.2.1) pode ser expressa, formalmente, como

$$\mathbf{x} = A^{-1}\,\mathbf{b}. \qquad (2.2.6)$$

O interesse na resolução de sistemas lineares é especialmente os casos em que exista solução e única.

Capítulo III

Erros em Computação

3.1 Introdução

O objetivo desse capítulo é procurar entender porque muitos resultado numéricos fornecidos pelo computador nem sempre são os esperados de acordo com a matemática. Para isso será apresentado a origem de vários tipos de erros em computação e a conseqüente propagação dos mesmos de uma operação para outra, segundo [Albrecht73, Steinberg74, Dorn72, Hattori94].

Para entender o comportamento dos computadores quando operam com números reais, também será apresentado como esses números podem ser representado no computador através do modelo em ponto flutuante, uma rápida revisão da aritmética com os números representados nesse modelo e os problemas causados quando se opera em ponto flutuante, segundo [Figueiredo89, Hattori94].

3.2 Origem e conceito dos erros

Na Computação Numérica existem três fontes de erros:

(a) Nos dados: quando os dados são obtidos experimentalmente, isto é, através de medidas. Toda medida está sujeita às limitações dos instrumentos usados e nela está embutido um *erro inerente* que é inevitável.

(b) Nos métodos: um método numérico é uma aproximação da solução de um problema de matemática. Nessa aproximação é comum introduzir o chamado *erro de truncamento*.

(c) No computador: o computador pode introduzir dois tipos de erros:

- Erro de conversão: quando um dado é fornecido num sistema de base diferente do sistema de base utilizada pelo computador. Por exemplo, um número que tem representação finita no sistema decimal pode não tê-la no sistema utilizado pelo computador.

- Erro de arredondamento: quando um número a ser representado tem d dígitos e a precisão disponível no computador é de t dígitos, com t < d. Por exemplo, multiplicando-se dois números x e y de t dígitos, o resultado precisa de $2t$ dígitos. Representando esse resultado com t dígitos comete-se o erro de arredondamento.

Definição 3.1 O *erro absoluto* (E) é a diferença entre x (valor exato) e x' (valor aproximado de x)

$$E = x - x'. \tag{3.2.1}$$

Na maioria dos casos estamos interessados no valor absoluto de E, $|E|$. Dizemos assim que $x = x' \pm |E|$, em que $|E|$ representa uma incerteza no valor de x'.

Definição 3.2 O *erro relativo* (e) é a razão entre o erro absoluto e o valor exato.

$$e = \frac{|E|}{|x|} = \frac{|x - x'|}{|x|} \cong \frac{|x - x'|}{|x'|}. \tag{3.2.2}$$

Definição 3.2 O *erro porcentual* (e^p) é obtido multiplicando-se o erro relativo por 100.

Por exemplo, para $x = 0{,}9995$ e $x' = 0{,}9994$ tem-se

$$E = 0{,}0001\ ,\ e \cong 0{,}0001 \quad \text{e} \quad e^p = 0{,}01\% \tag{3.2.3}$$

3.3 Propagação dos erros

Esta seção mostra como os erros em um dado ponto de um cálculo propagam-se, isto é, até que limite esses erros que contaminam os números utilizados numa operação afetam o resultado.

Seja $y = f(x_1, x_2, \ldots, x_n)$ o valor exato de uma operação f e $y' = f(x'_1, x'_2, \ldots, x'_n)$ o valor de f obtido quando opera-se com os dados x_i sujeitos a erro. $E_i = x_i - x'_i$ é o erro absoluto de cada operando. Deseja-se obter

$$E = y - y' = f(x_1,\ldots,x_n) - f(x'_1,\ldots,x'_n) = f(x_1,\ldots,x_n) - f(x_1 - E_1,\ldots,x_n - E_n). \tag{3.3.1}$$

Desenvolvendo $f(x_1 - E_1,\ldots,x_n - E_n)$ em série de Taylor, tem-se

$$f(x_1 - E_1,\ldots,x_n - E_n) = f(x_1,\ldots,x_n) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} E_i + \frac{1}{2!} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} E_i^2 - \ldots \tag{3.3.2}$$

Assim

$$E = y - y' = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} E_i - \frac{1}{2!} \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} E_i^2 - \ldots \tag{3.3.3}$$

Supondo que E_i seja muito menor que 1, uma hipótese razoável, podemos desprezar as derivadas de ordem superior a 1 para obter

$$E \equiv \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} E_i \ . \tag{3.3.4}$$

Tomando os valores absolutos, obtém-se uma estimativa do limite do erro.

$$|E| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} E_i \right| . \tag{3.3.5}$$

3.4 Aritmética de ponto flutuante

O computador representa um número real x usando um modelo aproximado denominado sistema de números em ponto flutuante.

Na forma de ponto flutuante, um número é representado como uma *fração*, também chamada *mantissa*, e um inteiro, chamado *expoente* ou *característica*. O número assim representado tem a forma

$$f \times b^e \tag{3.4.1}$$

em que, b é a base do sistema de números, $b \geq 2$,
e é o expoente (inteiro qualquer) e
f é a mantissa.

A representação interna de um número em ponto flutuante no computador tem a forma

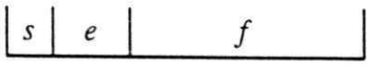

figura 3.1

em que, s é o sinal de f que é representado em alguma base B,

e é o expoente representado em uma base b, não necessariamente igual a B e

f é a mantissa de (3.1)

Como o número de dígitos de f e de e é limitado, nem todo número real pode ser representado em ponto flutuante. Logo, o conjunto de números reais representáveis em ponto flutuante num computador é finito.

Quanto mais próximo o dígito estiver do ponto decimal da mantissa, mais esse dígito é significativo. Um número em ponto flutuante está *normalizado* se o dígito mais significativo for diferente de zero

Em decimal flutuante um número x é representado por $x = f \times 10^e$, se ele for normalizado f não pode ser menor do que 1/10 pois seu primeiro dígito deve ser não nulo. Devido a que, por hipótese, f é uma fração própria o valor de f é absoluto e não pode chegar a 1. Resumindo

$$\frac{1}{10} \leq |f| < 1. \qquad (3.4.2)$$

Por exemplo, o número inteiro 3246 seria representado por $.3246 \times 10^4$.

Como nem todo número real pode ser representado em ponto flutuante, então o interessante é saber qual o menor e o maior positivos desse conjunto.

Seja o sistema em ponto flutuante,

$$PF(B, t, m, M), \tag{3.4.3}$$

em que: B é a base utilizada na representação da mantissa;

t o número de dígitos, na base B, da mantissa, chamada precisão;

m o menor expoente representável;

M o maior expoente representável.

O menor número positivo q deverá ter menor expoente, m e a menor mantissa possível. Devido a normalização, a menor mantissa será a = 100...0 e assim o número terá valor $0{,}1\times B^m$, ou seja,

$$q = B^{m-1} . \tag{3.4.4}$$

Expressando B e $m-1$ em decimal obtém-se o equivalente decimal deste menor número.

O maior positivo Q terá o maior expoente possível M, e todos os dígitos da mantissa iguais a $B-1$. O valor numérico da mantissa é obtido efetuando-se a soma

$$(B-1)\,B^{-1} + (B-1)\,B^{-2} + \ldots + (B-1)\,B^{-t} = (B-1)(B^{-1} + B^{-2} + \ldots + B^{-t}) = (1-B^{-t}). \tag{3.4.5}$$

Então o maior número positivo será

$$Q = (1-B^{-t})\,B^M . \tag{3.4.6}$$

3.5 Overflow e Underflow

Seja o sistema (3.4.3) e $\Re'$ o conjunto de números que podem ser representados

nesse sistema. Nesta seção será mostrada as diferenças entre efetuar cálculos em $\mathfrak{R}'$ e em $\mathfrak{R}$ (conjunto dos números reais).

A diferença entre executar operações em $\mathfrak{R}'$ e $\mathfrak{R}$ decorre do fato de que em $\mathfrak{R}'$ as operações não são fechadas, isto é, se $s_1 + s_2$, $s_1 - s_2$, $s_1 \times s_2$ e s_1/s_2 podem não pertencer a $\mathfrak{R}'$. Um número s resultante de uma operação não pertence a $\mathfrak{R}'$ nos seguintes casos:

(a) $| s | > Q$, de (3.4.6), neste caso diz-se que ocorre overflow, essa ocorrência é considerada irremediável.

(b) $0 < | s | \leq q$, de (3.4.4), neste caso diz-se que ocorre underflow, essa ocorrência é remediada adotando $s' = 0$. Essa prática pode invalidar um resultado, caso em que o underflow é chamado *destrutivo.*

3.6 Operações em ponto flutuante

Para somar dois números em ponto flutuante $x_1 = .1246 \times 10^1$ e $x_2 = .3290 \times 10^{-1}$. Desloca-se a mantissa do número de menor expoente para direita, em número de casas igual à diferença nos expoentes:

$$x_1 + x_2 = .1246 \times 10^1 + .003290 \times 10^1 = (.1246 + .003290) \times 10^1 = .1279 \times 10^1.$$

Note que os dígitos deslocadas para fora da precisão do sistema (quatro dígitos nesse caso) foram desprezados devido o arredondamento. Uma maneira de evitar a perda desses dígitos é fornecer mais dígitos na mantissa.

A subtração é realizada truncando-se o sinal do subtraendo e em seguida procedendo exatamente como na adição. A multiplicação e divisão em ponto flutuante são

realmente algo mais simples de realizar e explicar do que a soma. Na multiplicação as mantissas são multiplicadas como aparecem. Se ambos os fatores fossem normalizados de início, o produto ou já estaria normalizado ou teria quando muito um zero à esquerda, de modo que a normalização do resultado é fácil. O arredondamento não é feito usualmente. O expoente do resultado é simplesmente a soma dos expoentes dos fatores, modificados pela normalização, se esta foi realizada.

A divisão em ponto flutuante está inteiramente relacionada à multiplicação. O quociente ou já está normalizado ou tem quando muito um zero à esquerda, supondo que o dividendo e o divisor estejam normalizados, o expoente do quociente é a diferença dos dois expoentes.

Qualquer operação aritmética pode produzir overflow ou underflow. O caso extremo de overflow na operação de divisão ocorre ao se tentar dividir por zero, mas tal tentativa causa uma parada no programa.

Capítulo IV

Métodos Diretos

4.1 Introdução

Os métodos numéricos para a solução de sistemas lineares podem ser divididos em dois tipos: *diretos* e *iterativos*. Métodos diretos são aqueles que, na ausência de erros de arredondamento ou outros erros, conduzem à solução exata num número finito de operações aritméticas. Métodos iterativos são aqueles que a partir de uma aproximação inicial da solução tenta obter a solução por um processo de aproximações sucessivas. Neste, quando as aproximações tendem para a solução diz-se que o processo é *convergente*, caso contrário, diz-se *divergente*. Exemplos: método de Jacobi, gradientes conjugados, etc., [Ramos96].

Na prática, em virtude do computador operar com uma palavra de comprimento finito, os métodos diretos não conduzem a soluções exatas. Na verdade, erros que surgem de arredondamentos podem conduzir a resultados extremamente pobres ou, mesmo, sem

aplicação. Uma grande parte da análise numérica está dedicada ao estudo das razões do surgimento desses erros, as maneiras como os mesmos ocorrem e à pesquisa de métodos para minimizar a totalidade de tais erros. O método fundamental usado para obter uma solução direta é a *eliminação de Gauss*, mas, mesmo dentro dessa classe, há uma variedade de escolhas de métodos que diferem em eficiência computacional e precisão. Alguns desses métodos diretos serão estudados nas seções 4.2, 4.3 e 4.4, segundo [Hattori94, Dorn72, Golub96].

4.2 Fatoração *LU*

O objetivo desta seção é fatorar a matriz A do sistema linear

$$A\mathbf{x} = \mathbf{b}, \tag{4.2.1}$$

em que A é de ordem n, não singular e sem estrutura especial, num produto

$$A = LU, \tag{4.2.2}$$

em que L é uma matriz triangular inferior com diagonal principal unitária e U é uma matriz triangular superior, segundo [Hattori94, Ketter69, Fox64, Hopkins88].

Admitindo que as matrizes L e U tenham sido encontradas na forma

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{pmatrix} \quad \text{e} \quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{pmatrix} \tag{4.2.3}$$

tal que $A = LU$. O sistema linear (4.2.1) torna-se então

$$(LU)\mathbf{x} = \mathbf{b} \tag{4.2.4}$$

Estabelecendo

$$U\mathbf{x} = \mathbf{y} \tag{4.2.5}$$

o sistema linear (4.2.4) torna-se

$$L\mathbf{y} = \mathbf{b}. \tag{4.2.6}$$

A seguir será apresentado como obter os elementos das matrizes L e U a partir da matriz A.

1° passo: Obter a primeira linha de U a partir de

$$u_{1j} = a_{1j}\ , \qquad j = 1, 2, \ldots, n \tag{4.2.7}$$

e a primeira coluna de L a partir de

$$l_{i1} = \frac{a_{i1}}{a_{11}}\ , \qquad i = 2, 3, \ldots, n \tag{4.2.8}$$

Passo k: Após obter as $k - 1$ primeiras colunas de L, as $k - 1$ primeiras linhas de U e os elementos da diagonal principal da matriz L, que são todos iguais a 1, obtém-se

$$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj}\ , \qquad j = k, \ldots, n, \tag{4.2.9}$$

e analogamente obtém-se

$$l_{ik} = \frac{1}{u_{kk}}\left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk}\right), \qquad i = k+1, \ldots, n. \tag{4.2.10}$$

Os elementos da matriz $A = LU$ são obtidos por

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir}u_{rj}, \qquad i, j = 1, 2, \ldots, n. \tag{4.2.11}$$

A fatoração $A = LU$ só existe se os elementos (pivôs) $u_{ii} \neq 0$, $i = 1, \ldots, n$. O caso em que, pelo menos, um dos pivôs for nulo será tratado na seção 4.6.

Para resolver o sistema linear (4.2.1) por esse método, deve-se:

(a) fatorar A na forma $A = LU$,
(b) resolver o sistema triangular inferior $L\mathbf{y} = \mathbf{b}$ por substituição progressiva, e
(c) obter o vetor solução $\mathbf{x}$ a partir do sistema triangular superior $U\mathbf{x} = \mathbf{y}$ por substituição regressiva.

Com o subproduto da fatoração LU pode-se calcular o determinante da matriz A do sistema linear (4.2.1).

$$\det A = \det(LU) = \det L \times \det U. \tag{4.2.12}$$

Os determinantes das matrizes triangulares L e U são dados por

$$\det L = 1 \quad \text{e} \quad \det U = u_{11} \times u_{22} \times u_{33} \times \ldots \times u_{nn} = \prod_{i=1}^{n} u_{ii}. \tag{4.2.13}$$

Das equações (4.2.12) e (4.2.13) obtém-se

$$\det A = \det U. \tag{4.2.14}$$

Se $u_{ii} = 0$ para qualquer i, então

$$\det A = 0, \tag{4.2.15}$$

isso mostra que a matriz A é *singular*.

A seguir será apresentado os algoritmos para obter a fatoração LU, o vetor **y** e o vetor solução **x**.

4.2.1 Algoritmo da fatoração LU - Seja a matriz A do sistema linear (4.2.1), L uma matriz triangular inferior e U uma matriz triangular superior. Esse algoritmo constrói as matrizes L e U tal que $A = LU$, armazenando U no triângulo superior de A e L no triângulo inferior de A, sem armazenar os elementos da diagonal principal de L que são todos iguais a 1. A matriz A será destruída.

for $k = 1, \ldots, n\text{-}1$
  for $i = k+1, \ldots, n$
    $\eta = a_{ik} / a_{kk}$
    $a_{ik} := \eta$
    for $j = k+1, \ldots, n$
      $a_{ij} = a_{ij} - \eta a_{kj}$
    end for
  end for
end for

4.2.2 Algoritmo para substituição progressiva - Dada uma matriz triangular superior L de ordem n e o vetor independente **b**. Este algoritmo determina o vetor **y** de modo que $L\mathbf{y} = \mathbf{b}$.

for $i = 1, \ldots, n$

$y_i := b_i$

for $j = 1, \ldots, i-1$

$y_i := y_i - l_{ij} y_j$

$y_i := y_i / l_{ii}$

end for

end for

4.2.3 Algoritmo para substituição regressiva - Dada uma matriz triangular inferior U de ordem n. e o vetor independente **y**. Este algoritmo determina o vetor **x** de modo que $U\mathbf{x} = \mathbf{y}$.

for $i = n, \ldots, 1$

$x_i := y_i / u_{nn}$

for $j = i + 1, \ldots, n$

$x_i := x_i - u_{ij} x_j$

$x_i := x_i / u_{ii}$

end for

end for

4.3 Eliminação de Gauss

O objetivo desta seção é transformar a matriz completa A^* do sistema linear (4.2.1), de modo que, após a transformação a matriz incompleta A fique triangular superior, segundo [Hattori94, Ketter69, Fox64].

Seja o sistema linear (4.2.1) na forma

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2j}x_j + \ldots + a_{2n}x_n = b_2,\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3j}x_j + \ldots + a_{3n}x_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n = b_i, \qquad (4.3.1)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nj}x_j + \ldots + a_{nn}x_n = b_n.
\end{aligned}
$$

1° passo: De $a_{11} \neq 0$ (primeiro pivô) e m_i (i-ésimo multiplicador da i-ésima equação de (4.3.1)) dado por

$$m_i = a_{i1}/a_{11}, \qquad (4.3.2)$$

com $i = 2, 3, \ldots, n$. Obtém-se

$$a'_{ij} = a_{ij} - m_i a_{1j} \text{ e } b'_i = b_i - m_i b_1, \quad i = 2, \ldots, n \text{ e } j = 1, \ldots, n. \qquad (4.3.3)$$

O sistema linear após a eliminação de x_1 das $n-1$ últimas equações de (4.3.1) fica

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n = b_1,\\
&a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2j}x_j + \ldots + a'_{2n}x_n = b'_2,\\
&a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3j}x_j + \ldots + a'_{3n}x_n = b'_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a'_{i2}x_2 + a'_{i3}x_3 + \ldots + a'_{ij}x_j + \ldots + a'_{in}x_n = b'_i, \qquad (4.3.4)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nj}x_j + \ldots + a'_{nn}x_n = b'_n,
\end{aligned}
$$

2° passo: De $a'_{22} \neq 0$ (segundo pivô) e m'_i (i-ésimo multiplicador da i-ésima equação de (4.3.4)) dado por

$$m'_i = a'_{i1} / a'_{22}\,, \tag{4.3.5}$$

com $i = 3, \ldots, n$. Obtém-se

$$a''_{ij} = a'_{ij} - m'_i\, a'_{2j} \quad \text{e} \quad b''_i = b'_i - m'_i\, b'_{12}, \tag{4.3.6}$$

para $i = 3, \ldots, n$ e $j = 2, \ldots, n$.

O sistema linear após a eliminação de x_2 das $n - 2$ últimas equações de (4.3.4) fica

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n &= b_1,\\
a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2j}x + \ldots + a'_{2n}x_n &= b'_2,\\
a''_{33}x_3 + \ldots + a''_{3j}x_j + \ldots + a''_{3n}x_n &= b''_3,\\
&\ldots\ldots\ldots\ldots\\
a''_{i3}x_3 + \ldots + a''_{ij}x_j + \ldots + a''_{nn}x_n &= b''_i,\\
&\ldots\ldots\ldots\ldots\\
a''_{n3}x_3 + \ldots + a''_{nj}x_j + \ldots + a''_{nn}x_n &= b''_n.
\end{aligned} \tag{4.3.7}$$

Continuando analogamente, no k-ésimo passo, de $a_{kk}^{k-1} \neq 0$ (k-ésimo pivô) e

$$m_i^{k-1} = a_{ik}^{k-1} / a_{kk}^{k-1}\,, \tag{4.3.8}$$

com $i = k+1, \ldots, n$, obtém-se as k-ésimas equações:

$$a_{ij}^{k} = a_{ij}^{k-1} - m_i^{k-1} a_{kj}^{k-1} \quad \text{e} \quad b_i^{k} = b_i^{k-1} - m_i^{k-1} b_k^{k-1} \tag{4.3.9}$$

para $i = k+1, \ldots, n$ e $j = k, \ldots, n$ e $k = 1, \ldots, n-1$.

No passo $k = n - 1$ elimina-se x_k da última equação, obtendo a equação

$$a_{nn}^{n-1} x_n = b_n^{n-1}. \qquad (4.3.10)$$

O sistema linear (4.3.1) após a transformação fica

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2, \\ a''_{33}x_3 + \ldots + a''_{3n}x_n &= b''_3, \\ \ldots\ldots\ldots\ldots & \\ a_{nn}^{n-1} x_n &= b_n^{n-1}. \end{aligned} \qquad (4.3.11)$$

A solução do sistema linear (4.3.11), que também é solução do sistema linear (4.3.1), é obtida por substituições regressivas. Considerando os pivôs $a_{11}, a_{22}, \ldots, a_{kk}$ diferentes de zero, a solução do sistema triangular é dada por

$$\begin{aligned} x_n &= b_n^{n-1} / a_{nn}^{n-1}, \\ &\ldots\ldots\ldots\ldots \\ x_2 &= (b'_2 - (a'_{23}x_3 + \ldots + a'_{2n}x_n)) / a'_{22}, \\ x_1 &= (b_1 - (a_{12}x_2 + \ldots + a_{1n}x_n)) / a_{11}. \end{aligned} \qquad (4.3.12)$$

4.3.1 Algoritmo da Eliminação de Gauss - Dada a matriz completa A^*, do sistema linear (4.3.1). Este algoritmo computa a transformação (4.3.11).

for $k=1,n-1$

  for $i = k+1, \ldots, n$

    $\eta := a_{ik} / a_{kk}$

    $a_{ik} := \eta$

    for $j = k+1, \ldots, n+1$

      $a_{ij} := a_{ij} - \eta a_{kj}$.

end for

end for

end for

Após a transformação da matriz A na matriz triangular superior U, obtém-se o vetor solução **x** a partir das equações (4.3.12), isso é feito pelo algoritmo 4.2.3.

4.4 Eliminação de Jordan

O objetivo deste método é transformar a matriz completa A^* do sistema linear (4.2.1), de modo que, após a transformação a matriz incompleta A fique identidade, segundo [Hattori94, Ketter69, Fox64].

Este método é similar aos métodos de fatoração LU e eliminação de Gauss, no entanto não utiliza substituições regressivas nem progressivas para obter a solução do sistema linear.

Seja o sistema linear (4.2.1) na forma

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + \ldots + a_{2j}x_j + \ldots + a_{2n}x_n = b_2,\\
&a_{31}x_1 + a_{32}x_2 + \ldots + a_{3j}x_j + \ldots + a_{3n}x_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n = b_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nj}x_j + \ldots + a_{nn}x_n = b_n.
\end{aligned} \tag{4.4.1}$$

1° passo: Da primeira equação de (4.4.1) dividida por $a_{11} \neq 0$ e de m_i (i-ésimo multiplicador da i-ésima equação de (4.4.1)) dado por

$$m_i = a_{i1}, \tag{4.4.2}$$

com $i = 2, 3, \ldots, n$. Obtém-se

$$a'_{ij} = a_{ij} - m_i a_{1j}, \qquad \text{e} \qquad b'_i = b_i - m_i b_1, \tag{4.4.3}$$

para $i = 2, \ldots, n$ e $j = 1, \ldots, n$.

O sistema linear após a eliminação de x_1 nas $n-1$ últimas equações de (4.4.1) fica

$$\begin{aligned}
x_1 \; + \; & \frac{a_{12}}{a_{11}} x_2 + \ldots + \frac{a_{1j}}{a_{11}} x_j + \ldots + \frac{a_{1n}}{a_{11}} x_n = \frac{b_1}{a_{11}}, \\
& a'_{22} x_2 + \ldots + a'_{2j} x_j + \ldots + a'_{2n} x_n = b'_2, \\
& a'_{32} x_2 + \ldots + a'_{3j} x_j + \ldots + a'_{3n} x_n = b'_3, \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
& a'_{i2} x_2 + \ldots + a'_{ij} x_j + \ldots + a'_{in} x_n = b'_i, \\
& \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
& a'_{n2} x_2 + \ldots + a'_{nj} x_j + \ldots + a'_{nn} x_n = b'_n,
\end{aligned} \tag{4.4.4}$$

2° passo: Da segunda equação de (4.4.4) dividida por $a'_{22} \neq 0$ e de m'_i (i-ésimo multiplicador da i-ésima equação de (4.4.4)) dado por

$$m'_i = a_{i2}, \tag{4.4.5}$$

com $i = 1, 3, 4, \ldots, n$. Obtém-se

$$a''_{ij} = a'_{ij} - m'_i a'_{2j}, \qquad \text{e} \qquad b''_j = b'_i - m'_i b'_2, \tag{4.4.6}$$

com $i = 1, 3, 4, \ldots, n$ e $j = 1, \ldots, n$.

O sistema linear após a eliminação de x_2 na primeira e nas $n-2$ últimas equações de (4.4.4) fica

$$
\begin{aligned}
x_1 \quad &+ \quad a''_{13}x_3 + \ldots + a''_{1j}x_j + \ldots + a''_{1n}x_n = b''_1, \\
x_2 + \frac{a''_{23}}{a''_{22}}x_3 &+ \ldots + \frac{a''_{2j}}{a''_{22}}x_j + \ldots + \frac{a''_{2n}}{a''_{22}}x_n = \frac{b''_2}{a''_{22}}, \\
a''_{33}x_3 &+ \ldots + a''_{3j}x_j + \ldots + a''_{3n}x_n = b''_3, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a''_{i3}x_3 &+ \ldots + a''_{ij}x_j + \ldots + a''_{in}x_n = b''_i, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a''_{n3}x_3 &+ \ldots + a''_{nj}x_j + \ldots + a''_{nn}x_n = b''_n.
\end{aligned}
\qquad (4.4.7)
$$

Continuando, analogamente, no k-ésimo passo: Da k-ésima equação dividida por $a_{kk}^{k-1} \neq 0$ (k-ésimo pivô) e de

$$m_i^{k-1} = a_{ik}^{k-1}, \qquad (4.4.8)$$

com $i = 1, 2, \ldots, k-1, k+1, \ldots, n$ e $k = 2, \ldots, n$. Obtém-se

$$a_{ik}^{k} = a_{ij}^{k-1}{}_{j} - m_i^{k-1}\, a_{kj}^{k-1} \qquad \text{e} \qquad b_i^k = b_i^{k-1} - m_i^{k-1}\, b_k^{k-1}\, b_k \qquad (4.4.9)$$

com $i = 1,2, \ldots, k-1, k+1, \ldots, n$ e $j = k, \ldots, n$.

Quando $k = n$ os elementos acima e abaixo da diagonal principal da matriz A são todos eliminados. O sistema linear (4.4.1) finalmente fica

$$
\begin{aligned}
x_1 \qquad\qquad\qquad &= b_1^n , \\
x_2 \qquad\qquad\qquad &= b_2^n , \\
x_3 \qquad\qquad\qquad &= b_3^n , \\
\cdots\cdots\cdots\cdots\cdots\cdots & \qquad\qquad (4.4.10) \\
x_n = b_n^n .
\end{aligned}
$$

Assim $\mathbf{x}^T = (b_1^n , b_2^n , b_3^n , \ldots , b_n^n)$ de (4.4.10) é a solução do sistema linear (4.4.1).

4.4.1 Algoritmo da Eliminação de Jordan - Dada a matriz A^*, matriz aumentada do sistema linear (4.4.1). Este algoritmo computa a transformação (4.4.10).

for $i = 1, \ldots, n$
  for $k = 1, \ldots, n$
    $\eta := a_{ik} / a_{kk}$
    $a_{ik} := \eta$
    for $j = 1, \ldots, n+1$
      if ($k \neq i$) **then**
        $a_{ij} = a_{ij} - \eta a_{kj}$
    end for
  end for
end for
for j = 1, . . ., n
  $x_j = a_{j, n+1}$
end for

4.5 Erros na solução de sistemas lineares

Nos métodos apresentados neste capítulo, erros de arredondamentos na obtenção da solução de sistemas lineares via computador são inevitáveis devido a limitação da precisão. Além da introdução de erros nos cálculos, os elementos da matriz A e do vetor $\mathbf{b}$ poderão estar afetados por erro, seja porque foram calculados, seja porque são resultados de medidas (seção 3.2). Na verdade os erros em A e $\mathbf{b}$ muitas vezes não são conhecidos. Ao usar o computador obtém-se a solução computada $\mathbf{x}'$ que pode ser uma boa solução ou não. Nesta seção será mostrado critérios para verificar se uma solução computada de um sistema linear é aceitavel ou não, segundo [Hattori94, Dorn72, Albrecht73, Steinberg74].

Antes será apresentado algumas definições necessárias:

Definição 4.5.1 - A *norma vetorial*, denotada por $\| \mathbf{x} \|$, é uma função $f : \Re^n \rightarrow \Re$ que satisfaz as seguintes propriedades

i) $\| \mathbf{x} \| \geq 0$; $\| \mathbf{x} \| = 0$ implica $\mathbf{x} = 0$ (vetor nulo),
ii) $\| k\mathbf{x} \| = | k | \| \mathbf{x} \|$, $k \in \Re$ e $\mathbf{x} \in \Re^n$,
iii) $\| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|$, $\mathbf{x}, \mathbf{y} \in \Re^n$.

Definição 4.5.2 - Seja $\| \mathbf{x} \|$ uma norma vetorial e $\mathbf{x} \in \Re^n$. A *norma da matriz A* induzida pela norma vetorial é

$$\| A \| = \max_{\|\mathbf{x}\|=1} \| A\mathbf{x} \| \qquad (4.5.1)$$

Propriedades:

Para quaisquer matrizes A e B quadradas e qualquer vetor $\mathbf{x}$, tem-se

i) $\|AB\| \leq \|A\| \times \|B\|$,

ii) $\|A\mathbf{x}\| \leq \|A\| \times \|\mathbf{x}\|$.

Definição 4.5.3 - A *norma* ∞ (ou de Chebyschev) de um vetor $\mathbf{x}$, denotada por $\| \mathbf{x} \|_\infty$, é dada por

$$\| \mathbf{x} \|_\infty = \max_{1 \le i \le n} |x_i| \tag{4.5.2}$$

Um dos critérios é, ao resolver $A\mathbf{x} = \mathbf{b}$ e obter a solução computada $\mathbf{x}_c$ substituí-la no sistema linear e verificar se resíduo

$$\mathbf{r} = \mathbf{b} - A\,\mathbf{x}_c \tag{4.5.3}$$

é pequeno em alguma norma. A solução mais significativa é a que apresentar o menor resíduo, espera-se.

Um outro critério é recomputar a solução introduzindo uma pequena mudança no problema, alterando um ou mais dados (computar a solução em precisão dupla, caso seja possível, é uma prática comum). Considerando o sistema linear

$$A\mathbf{x} = \mathbf{b}, \tag{4.5.4}$$

na forma:

$$\begin{pmatrix} 41 & 40 \\ 40 & 39 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 81 \\ 79 \end{pmatrix}, \tag{4.5.5}$$

supondo que exista um pequeno erro no vetor $\mathbf{b}$ e que esse vetor passe a ser

$$\mathbf{b}' = (80.99, 79.01)^T \tag{4.5.6}$$

(uma pequena mudança de 0,012%), resolvendo o sistema

$$A\mathbf{x}' = \mathbf{b}' \tag{4.5.7}$$

pelos métodos de fatoração *LU*, eliminação de Gauss e eliminação de Jordan obtém-se os esultados aproximados

$$\mathbf{x}' = (1.7902, 0.1899)^T,\ \mathbf{x}' = (1.9799, 0.1899)^T \text{ e } \mathbf{x}' = (1.7902, 0.1899)^T, \tag{4.5.8}$$

respectivamente. Esses resultados levam a crer que as soluções de (4.5.8) são boas aproximações da solução do sistema (4.5.5). Mas estas aproximações estão longe da solução exata que é

$$\mathbf{x} = (1, 1)^T. \tag{4.5.9}$$

Para entender o que ocorre no exemplo (4.5.5) deve-se entender as variações relativas do vetor solução e do vetor **b**, seção (3.2). Da fatoração *LU* obtém-se os erros absolutos

$$E\mathbf{x} = (1, 1)^T - (1.7902, 0.1899)^T = (-0.7902, 0.8101)^T, \tag{4.5.10}$$

$$E\mathbf{b} = (81, 79)^T - (80.99, 79.01)^T = (0.01, -0.01)^T. \tag{4.5.11}$$

Usando norma ∞ obtém-se os erros relativos

$$e_{\mathbf{x}} = \frac{\| E\mathbf{x} \|}{\| \mathbf{x}' \|} = \frac{0.8101}{1.7902} = 0.4525, \tag{4.5.12}$$

$$e_{\mathbf{b}} = \frac{\| E\mathbf{b} \|}{\| \mathbf{b}' \|} \cong \frac{0.01}{80.99} = 0.000125 = 1.25 \times 10^{-4}. \tag{4.5.13}$$

Tem-se então

$$\frac{\| e_{\mathbf{x}} \|}{\| e_{\mathbf{b}} \|} = \frac{0.4525}{0.000125} = 3620 \tag{4.5.14}$$

ou seja, a variação do erro relativo na solução foi cerca de 3620 vezes a variação do erro relativo em **b**.

Observa-se que pequenos erros em **b** causam erros maiores, na solução **x**. Isto é um desastre.

A seguir será apresentada uma estimativa da relação entre os erros relativos nos dados ($\mathbf{b}$ e A) e na solução. Começando introduzindo um erro $E\mathbf{b}$ e calculando o erro $E\mathbf{x}$ no sistema linear $A\mathbf{x} = \mathbf{b}$. Seja

$$A(\mathbf{x} + E\mathbf{x}) = \mathbf{b} + E\mathbf{b}$$

o sistema após a introdução do erro. Tem-se

$$A\mathbf{x} + AE\mathbf{x} = \mathbf{b} + E\mathbf{b},$$

$$AE\mathbf{x} = E\mathbf{b},$$

e em alguma norma,

$$\| E\mathbf{x} \| \leq \| A^{-1} \| \| E\mathbf{x} \|. \tag{4.5.15}$$

Introduzindo erro em A

$$(A + EA)(\mathbf{x} + E\mathbf{x}) = \mathbf{b},$$

$$A\mathbf{x} + AE\mathbf{x} + EA\mathbf{x} + EA\, E\mathbf{x} = \mathbf{b},$$

$$AE\mathbf{x} + EA(\mathbf{x} + E\mathbf{x}) = 0,$$

onde se retira

$$E\mathbf{x} = -A^{-1}\, EA(\mathbf{x} + E\mathbf{x}),$$

e, em alguma norma

$$\|E\mathbf{x}\| \leq \|A^{-1}\| \, \|EA\| \, \|\mathbf{x}+E\mathbf{x}\|,$$

$$\frac{\|E\mathbf{x}\|}{\|\mathbf{x} + E\mathbf{x}\|} \leq \|A^{-1}\| \, \|A\| \, \frac{\|EA\|}{\|A\|}, \tag{4.5.16}$$

que expressa o erro relativo na solução $\mathbf{x}$ devido o erro relativo em A

O número

$$K(A) = \|A^{-1}\| \, \|A\|, \tag{4.5.17}$$

é chamado *número de condição* da matriz A em relação a alguma norma dada.

Da desigualdade

$$\|\mathbf{b}\| = \| A\mathbf{x} \| \leq \| A \| \, \| \mathbf{x} \|$$

obtém-se

$$\frac{1}{\|\mathbf{x}\|} \leq \frac{\|A\|}{\|\mathbf{b}\|}. \tag{4.5.18}$$

Multiplicando ambos os membros de (4.5.15) por $1/\|\mathbf{x}\|$ e utilizando a desigualdade (4.5.18), obtém-se

$$\frac{\|E\mathbf{x}\|}{\|\mathbf{x}\|} \leq K(A) \frac{\|E\mathbf{b}\|}{\|\mathbf{b}\|}, \tag{4.5.19}$$

que dá uma estimativa de erro relativo em **x** devido o erro relativo em **b**.

A interpretação que se dá a (4.5.16) e (4.5.19) é a seguinte: se $K(A) >> 1$, então o sistema linear $A\mathbf{x} = \mathbf{b}$ é dito *mal-condicionado*, isto é, para matrizes cujo número de condição seja elevado, um pequeno erro relativo em A ou **b**, pode causar um grande erro relativo na solução **x**.

4.6 Pivotamento

Nos métodos apresentados neste capítulo, as transformações foram executadas ignorando a possibilidade de parada no processo de transformação do sistema linear, essa

parada é devido a um dos elementos da diagonal da matriz U ser nulo, pois tendo um valor nulo a divisão por este valor é impossível, caso de overflow.

Em (4.2.15) foi mostrado que, se um dos elementos da diagonal principal da matriz U (pivôs) se anular o determinante da matriz A é zero e a matriz A é dita singular. Portanto, neste caso, a matriz A não admite inversa e o sistema linear não pode ser resolvido.

No método da eliminação de Gauss (seção 4.3), sabendo-se que o sistema linear tem solução, no k-ésimo passo da obtenção do multiplicador

$$m_i^{k-1} = a_{ik}^{k-1} / a_{kk}^{k-1} , \tag{4.6.1}$$

com $i = k+1, \ldots, n.$, mesmo que o pivô a_{kk}^{k-1} seja zero, o sistema linear pode ser resolvido fazendo uma permuta (ou troca) de linhas para obter um pivô não nulo, se o valor for próximo de zero e não for efetuada uma troca de linhas ou colunas, os erros de arredondamentos (surgidos durante a transformação do sistema) podem provocar grandes erros nos resultados (seção 4.5). Por exemplo, seja o sistema linear

$$\begin{pmatrix} 0 & 5 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10 \\ 4 \end{pmatrix}, \tag{4.6.2}$$

um dos pivôs é zero, pelo algoritmo 4.3.1 o sistema linear (4.6.2) não pode ser resolvido, mas o sistema linear tem solução $\mathbf{x}^T = (1, 2)$, o fato de um dos pivôs ser zero não indica que o sistema não tenha solução, para evitar pivôs nulos deve-se incluir no algoritmo 4.3.1 uma permuta de linhas ou colunas para obter um pivô não nulo.

Após a troca da primeira linha pela segunda linha o sistema linear (4.6.2) fica

$$\begin{pmatrix} 2 & 1 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \end{pmatrix} . \tag{4.6.3}$$

Para evitar essa parada e a propagação desses erros durante a transformação do sistema linear, em cada passo k da eliminação de Gauss escolhe-se o pivô $a_{kk}^{k-1} \neq 0$. Esta escolha e feita pelas seguintes estratégias:

Pivotamento parcial - Em cada passo k da eliminação de Gauss, troca-se a linha k com a linha r, tal que

$$|a_{rk}^{k-1}| = \max_{i=k,\ldots,n} |a_{ik}^{k-1}|, \tag{4.6.4}$$

em que $k \neq r$.

Pivotamento total - Em cada passo k da eliminação de Gauss, troca-se a linha k com a linha r e a coluna k, tal que

$$|a_{rs}^{k-1}| = \max_{i,j=k,\ldots,n} |a_{ij}^{k-1}|, \tag{4.6.5}$$

em que $k \neq r$ e $k \neq s$.

Essas estratégias de pivotamento também são aplicáveis nos método de fatoração *LU* e eliminação de Jordan.

Capítulo V

Solução Exata de Sistemas Lineares

5.1 Introdução

No capítulo IV foi mostrado como se obter a solução de um sistema linear pelos métodos diretos de fatoração *LU*, eliminação de Gauss e eliminação de Jordan, a interferência dos erros de arredondamentos na obtenção da solução exata de um sistema linear e as estratégias de pivotamento. Este capítulo mostra as maneiras de se obter, caso exista, a solução exata de um sistema linear por esses métodos diretos, um estudo do método proposto por Fox para obter a solução exata de um sistema linear sem erros de arredondamentos, a construção do algoritmo e uma comparação entre esse método e os métodos diretos quanto ao número de operações realizadas. Finalmente, mostra a necessidade da aritmética em múltipla precisão na obtenção da solução exata de um sistema linear pelo método proposto por Fox.

5.2 Solução Exata de Sistemas Lineares

Seja o sistema linear

$$A\mathbf{x} = \mathbf{b}, \tag{5.2.1}$$

em que A é uma matriz, não singular, de ordem n cujos elementos são números inteiros e $\mathbf{b}$ um vetor coluna de n componentes inteiros. Todas as operações realizadas nesse método são exatas, inclusive a divisão.

Para ilustrar o método, seja o sistema linear (5.2.1) na forma (5.2.2)

$$\begin{aligned}
7x_1 + 9x_2 - x_3 + 2x_4 &= 8, && (5.2.2.a)\\
4x_1 - 5x_2 + 2x_3 - 7x_4 &= 7, && (5.2.2.b)\\
x_1 + 6x_2 - 3x_3 - 4x_4 &= 5, && (5.2.2.c)\\
3x_1 - 2x_2 - x_3 - 5x_4 &= 11. && (5.2.2.d)
\end{aligned}$$

Na resolução do sistema (5.2.2) pelo método da eliminação de Gauss os multiplicadores $m_2 = a_{21} / a_{11}$, $m_3 = a_{31} / a_{11}$ e $m_4 = a_{41} / a_{11}$ são os números 4/7, 1/7 e 3/7, respectivamente, como as representações decimais desses números não são finitas, e precisam ser armazenadas num espaço finito, limitação do computador, comete-se aí os erros de arredondamentos. O uso de números com essas representações infinitas durante a transformação do sistema, pelos métodos diretos, faz com que a solução do sistema (5.2.2) não seja exata, devido o surgimento de erros de arredondamentos e a propagação dos mesmos de uma operação para outra.

Uma maneira de se obter a solução exata a partir de tais métodos é utilizando aritmética racional ou de resíduos. Uma outra maneira de se obter a solução exata de um sistema linear, é utilizar um método em que todas as operações sejam executadas exatamente, inclusive divisão.

Para obter a solução do sistema linear (5.2.2) pelo método proposto por Fox deve-se manter inteiros os elementos da matriz associada ao sistema linear durante o processo de transformação do sistema linear, contanto que durante a transformação não apareça elemento cuja representação supere o limite de precisão disponível no computador.

No primeiro passo da transformação, os coeficientes e termo independente da equação (5.2.2.b) do sistema linear (5.2.2) são dados pelos determinantes

$$\det\begin{pmatrix}7 & 7\\ 4 & 4\end{pmatrix}=0,\ \det\begin{pmatrix}7 & 9\\ 4 & -5\end{pmatrix}=-71,\ \det\begin{pmatrix}7 & -1\\ 4 & 2\end{pmatrix}=18,\ \det\begin{pmatrix}7 & 2\\ 4 & -7\end{pmatrix}=-57 \text{ e } \det\begin{pmatrix}7 & 8\\ 4 & 7\end{pmatrix}=17,$$

eliminando x_1 da equação (5.2.2.b), para a equação (5.2.2.c) são dados por

$$\det\begin{pmatrix}7 & 7\\ 1 & 1\end{pmatrix}=0,\ \det\begin{pmatrix}7 & 9\\ 1 & 6\end{pmatrix}=33,\ \det\begin{pmatrix}7 & -1\\ 1 & -3\end{pmatrix}=-20,\ \det\begin{pmatrix}7 & 2\\ 1 & -4\end{pmatrix}=-30 \text{ e } \det\begin{pmatrix}7 & 8\\ 1 & 5\end{pmatrix}=27,$$

eliminando x_1 da equação (5.2.2.c) e para a equação (5.2.2.d) são dados por

$$\det\begin{pmatrix}7 & 7\\ 3 & 3\end{pmatrix}=0,\det\begin{pmatrix}7 & 9\\ 3 & -2\end{pmatrix}=-41,\ \det\begin{pmatrix}7 & -1\\ 3 & -1\end{pmatrix}=-4,\ \det\begin{pmatrix}7 & 2\\ 3 & -5\end{pmatrix}=-41 \text{ e } \det\begin{pmatrix}7 & 8\\ 3 & 11\end{pmatrix}=53,$$

eliminando x_1 da equação (5.2.2.d).

O sistema linear (5.2.3) obtido após a eliminação de x_1 nas três últimas equações de (5.2.2) fica

$$\begin{aligned} 7x_1+9x_2 - x_3 + 2x_4 &= 2, && \text{(5.2.3.a)}\\ -71x_2 + 18x_3 - 57x_4 &= 17, && \text{(5.2.3.b)}\\ 33x_2 - 20x_3 - 30x_4 &= 27, && \text{(5.2.3.c)}\\ -41x_2 - 4x_3 - 41x_4 &= 53. && \text{(5.2.3.d)} \end{aligned}$$

Analogamente, elimina-se x_2 do sistema formado pela equações (5.2.3.b), (5.2.3.c) e (5.2.3.d), obtendo o sistema linear (5.4)

$$\begin{aligned} 7x_1 + 9x_2 - x_3 + 2x_4 &= 2, && (5.2.4.a)\\ -71x_2 + 18x_3 - 57x_4 &= 17, && (5.2.4.b)\\ 826x_3 + 4011x_4 &= -2478, && (5.2.4.c)\\ 1022x_3 + 574x_4 &= 3066. && (5.2.4.d) \end{aligned}$$

Como os coeficientes e termos independentes das equações (5.2.4.c) e (5.2.4.d) são divisíveis por 7 (pivô da equação (5.2.2.a)), essa divisão será justificada mais adiante, o sistema (5.2.4) após a divisão das equações (5.2.4.c) e (5.2.4.d) por 7 fica representado pelo sistema linear (5.2.5)

$$\begin{aligned} 7x_1 + 9x_2 - x_3 + 2x_4 &= 2, && (5.2.5.a)\\ -71x_2 + 18x_3 - 57x_4 &= 17, && (5.2.5.b)\\ 118x_3 + 573x_4 &= -354, && (5.2.5.c)\\ 146x_3 + 82x_4 &= 438. && (5.2.5.d) \end{aligned}$$

As equações (5.2.5.c) e (5.2.5.d) formam um sistema de duas equações a duas incógnitas, x_3 e x_4. Analogamente, elimina-se x_3 de (5.2.5.d) obtendo o sistema linear (5.2.6)

$$\begin{aligned} 7x_1 + 9x_2 - x_3 + 2x_4 &= 2, && (5.2.6.a)\\ -71x_2 + 18x_3 - 57x_4 &= 17, && (5.2.6.b)\\ 118x_3 + 573x_4 &= -354, && (5.2.6.c)\\ -73982x_4 &= 0. && (5.2.6.d) \end{aligned}$$

O coeficiente e o termo independente da equação (5.2.6.d), são divisíveis por −71 (pivô da equação (5.2.5.b)), finalmente, o sistema linear (5.2.6) após a divisão da equação (5.2.6.d) por −71 fica representado pelo sistema linear (5.2.7)

$$
\begin{aligned}
7x_1 + 9x_2 \;-\; x_3 \;+\; 2x_4 &= 2, && \text{(5.2.7.a)}\\
-71x_2 \;+\; 18x_3 \;-\; 57x_4 &= 17, && \text{(5.2.7.b)}\\
118x_3 \;+\; 573x_4 &= -354, && \text{(5.2.7.c)}\\
1042x_4 &= 0. && \text{(5.2.7.d)}
\end{aligned}
$$

Antes de completar a solução será analisado o sistema (5.2.7), particularmente os coeficientes da diagonal, e justificar a divisibilidade exata das equações (5.2.4.c) e (5.2.4.d) pelo coeficiente de x_1 da equação (5.2.3.a).

Para a análise seja o sistema linear, não singular, com n equações

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1s}x_s + \ldots + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2s}x_s + \ldots + a_{2n}x_n = b_2,\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3s}x_s + \ldots + a_{3n}x_n = b_3,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{r1}x_1 + a_{r2}x_2 + a_{r3}x_3 + \ldots + a_{rs}x_s + \ldots + a_{rn}x_n = b_r, \qquad \text{(5.2.8)}\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{rs}x_s + \ldots + a_{nn}x_n = b_n.
\end{aligned}
$$

Para $a_{11} \neq 0$ (primeiro pivô). Obtém-se as $n - 1$ últimas equações de (5.2.8) pelos determinantes de ordem 2

$$
a'_{rs} = \det\begin{pmatrix} a_{11} & a_{1s} \\ a_{r1} & a_{rs} \end{pmatrix} = a_{11}a_{rs} - a_{1s}a_{r1}
$$

e

$$
b'_r = \det\begin{pmatrix} a_{11} & b_1 \\ a_{r1} & b_r \end{pmatrix} = a_{11}b_r - b_1a_{r1}, \qquad \text{(5.2.9)}
$$

com $r = 2, 3,\ldots,n$ e $s = 1, \ldots, n$.

O sistema linear após a eliminação de x_1 nas $n - 1$ últimas equações de (5.2.8) fica

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1s}x_s + \ldots + a_{1n}x_n = b_1,\\
&\quad a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2s}x_s + \ldots + a'_{2n}x_n = b'_2,\\
&\quad a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3s}x_s + \ldots + a'_{3n}x_n = b'_3,\\
&\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\quad a'_{r2}x_2 + a'_{r3}x_3 + \ldots + \boldsymbol{a'_{rs}x_s} + \ldots + a'_{rn}x_n = \boldsymbol{b'_r},\\
&\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\quad a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{ns}x_s + \ldots + a'_{nn}x_n = b'_n.
\end{aligned}
\qquad (5.2.10)
$$

Analogamente, para $a'_{22} \neq 0$ (segundo pivô), repete-se o procedimento para obter as $n - 2$ últimas equações de (5.2.10)

$$
a''_{rs} = \det\begin{pmatrix} a'_{22} & a'_{2s} \\ a'_{r2} & a'_{rs} \end{pmatrix} = a'_{22}a'_{rs} - a'_{2s}a'_{r2}
$$

e

$$
b''_{r} = \det\begin{pmatrix} a'_{22} & b'_{2} \\ a'_{r2} & b''_{r} \end{pmatrix} = a'_{22}b'_{r} - b'_{2}a'_{r2}, \qquad (5.2.11)
$$

com $r = 3, 4, \ldots, n$ e $s = 2, \ldots, n$.

O sistema linear após a eliminação de x_2 das $n - 2$ últimas equações de (5.2.10) fica

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1s}x_s + \ldots + a_{1n}x_n = b_1,\\
&\quad a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2s}x_s + \ldots + a'_{2n}x_n = b'_2,\\
&\quad a''_{33}x_3 + \ldots + a''_{3s}x_s + \ldots + a''_{3n}x_n = b''_3,\\
&\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\quad a''_{r3}x_3 + \ldots + \boldsymbol{a''_{rs}x_s} + \ldots + a''_{rn}x_n = \boldsymbol{b''_r},\\
&\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\quad a''_{n3}x_3 + \ldots + a''_{ns}x_s + \ldots + a''_{nn}x_n = b''_n,
\end{aligned}
\qquad (5.2.12)
$$

em que $a''_{33} \neq 0$ (terceiro pivô).

Resolvendo o sistema constituído pela equações (5.2.9) e (5.2.11) resulta

$$\begin{aligned} a''_{rs} &= (a_{11}a_{22} - a_{21}a_{12})(a_{11}a_{rs} - a_{r1}a_{1s}) - (a_{11}a_{r2} - a_{r1}a_{12})(a_{11}a_{2s} - a_{21}a_{1s}) = \\ &= (a_{11}^2 a_{22}a_{rs} - a_{11}a_{22}a_{r1}a_{1s} - a_{21}a_{12}a_{11}a_{rs} + a_{21}a_{12}a_{r1}a_{1s}) - (a_{11}^2 a_{r2}a_{2s} - \\ &\quad - a_{11}a_{r2}a_{21}a_{1s} - a_{r1}a_{12}a_{11}a_{2s} + a_{r1}a_{12}a_{21}a_{1s}) = a_{11}(a_{11}a_{22}a_{rs} - a_{22}a_{r1}a_{1s} - \\ &\quad - a_{21}a_{12}a_{rs} - a_{11}a_{r2}a_{2s} + a_{r2}a_{21}a_{1s} + a_{r1}a_{12}a_{2s}), \end{aligned} \tag{5.2.13.a}$$

$$\begin{aligned} b''_{r} &= (a_{11}a_{22} - a_{21}a_{12})(a_{11}b_r - a_{r1}b_1) - (a_{11}a_{r2} - a_{r1}a_{12})(a_{11}b_2 - a_{21}b_1) = \\ &= (a_{11}^2 a_{22}b_r - a_{11}a_{22}a_{r1}b_1 - a_{21}a_{12}a_{11}b_r + a_{21}a_{12}a_{r1}b_1) - (a_{11}^2 a_{r2}b_2 - \\ &\quad - a_{11}a_{r2}a_{21}b_1 - a_{r1}a_{12}a_{11}b_2 + a_{r1}a_{12}a_{21}b_1) = a_{11}(a_{11}a_{22}b_r - a_{22}a_{r1}b_1 - \\ &\quad - a_{21}a_{12}b_r - a_{11}a_{r2}b_2 + a_{r2}a_{21}b_1 + a_{r1}a_{12}b_2). \end{aligned} \tag{5.2.13.b}$$

Os termos a''_{rs} e b''_{r} são divisíveis por a_{11} (primeiro pivô em (5.2.8)), os termos a'''_{rs} e b'''_{r} da próxima redução são divisíveis pelo pivô a'_{22} (segundo pivô em (5.2.10)) e assim sucessivamente até que o sistema linear (5.2.8) fique

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1, \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2, \\ a''_{33}x_3 + \ldots + a''_{3n}x_n &= b''_3, \\ \ldots\ldots\ldots\ldots\ldots & \\ a^{n-1}_{nn}x_{nn} &= a^{n-1}_{n}. \end{aligned}$$

Considere a relação entre a matriz A (matriz dos coeficientes do sistema 5.2.2) e a matriz triangular U (matriz dos coeficientes de 5.2.7), no processo de eliminação dos elementos abaixo da diagonal principal da matriz A, é executada uma seqüência de operações matriciais representada por $JA = U$, em que J é um produto da forma

$$JA = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & u_{22}^{-1} \end{pmatrix}}_{J_5} \times \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & * & u_{33} \end{pmatrix}}_{J_4} \times \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & u_{11}^{-1} & 0 \\ 0 & 0 & 0 & u_{11}^{-1} \end{pmatrix}}_{J_3} \times \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & * & u_{22} & 0 \\ 0 & * & 0 & u_{22} \end{pmatrix}}_{J_2} \times$$

$$\times \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ -a_{21} & a_{11} & 0 & 0 \\ -a_{31} & 0 & a_{11} & 0 \\ -a_{41} & 0 & 0 & a_{11} \end{pmatrix}}_{J_1} \times \underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}}_{A} = \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ & u_{22} & u_{23} & u_{24} \\ & & u_{33} & u_{34} \\ & & & u_{44} \end{pmatrix} = U \qquad (5.2.14)$$

em que:

u_{11}^{-1} é o inverso do pivô a_{11} na equação (5.2.2),

u_{22}^{-1} é o inverso do pivô a'_{22} na equação (5.2.3) e

u_{rr} , $r = 1,..,4$, são os pivôs usados na triangularização da matriz A .

* são os elementos obtidos após a permuta de linhas, caso seja necessária durante a triangularização da matriz A .

Seja

$$\det U = \det JA = \det J_5 \times \det J_4 \times \det J_3 \times \det J_2 \times \det J_1 \times \det A \,, \qquad (5.2.15)$$

como

$$\det J_5 = u_{22}^{-1} \,, \det J_4 = u_{33} \,, \det J_3 = u_{11}^{-2} \,, \det J_2 = u_{22}^{2} \quad \text{e} \quad \det J_1 = a_{11}^{3}, \qquad (5.2.16)$$

de (5.2.15) e (5.2.16) obtém-se

$$u_{22}^{-1} u_{33} u_{11}^{-2} u_{22}^{2} a_{11}^{3} \det A = u_{11} u_{22} u_{33} u_{44} \,. \qquad (5.2.17)$$

Mas $u_{11} = a_{11}$, então,

$$\det A = u_{44}. \tag{5.2.18}$$

O elemento u_{44} da matriz triangular superior U é o determinante da matriz A do sistema linear (5.2.2). Por analogia u_{33} é o determinante da matriz obtida de A eliminando a última linha e a última coluna, e assim por diante.

Aplica-se o processo de substituição regressiva em (5.2.7) para obter a solução do nosso exemplo. Os componentes do vetor solução **x** são calculados a partir de uma relação cujo denominador é o det A. Como os coeficientes do sistema (5.2.7) são inteiros os numeradores da relação também são inteiros. Pode-se então introduzir um vetor auxiliar.

$$\mathbf{y} = \det A \times \mathbf{x}, \tag{5.2.19}$$

cujos componentes são números inteiros.

Multiplicando os termos independentes das equações de (5.2.7) por det A = 1042 obtém-se o sistema linear (5.2.20)

$$\begin{aligned}
7y_1 + 9y_2 - y_3 + 2y_4 &= 2048, && \text{(5.2.20.a)}\\
-71y_2 + 18y_3 - 57y_4 &= 17714, && \text{(5.2.20.b)}\\
118y_3 + 573y_4 &= -368868, && \text{(5.2.20.c)}\\
1042y_4 &= 0. && \text{(5.2.20.d)}
\end{aligned}$$

Resolvendo o sistema linear (5.2.20) por substituição regressiva obtém-se

$$\mathbf{y} = (2084, -1042, -3126, 0)^T \tag{5.2.21}$$

De (5.2.19) e (5.2.21), obtém-se o vetor solução **x** do sistema linear (5.2.2) cujos componentes racionais são dados pela relação

$$x_i = y_i / \det A , \qquad i = 1,...,4. \tag{5.2.22}$$

Assim a solução do sistema linear (5.2.2) na forma racional fica

$$x_1 = 2084/1042,\ x_2 = -1042/1042,\ x_3 = -3126/1042 \text{ e } x_4 = 0/1042. \tag{5.2.23}$$

Simplificando o resultado (5.2.23) obtém-se o vetor solução

$$\mathbf{x} = (2, -1, -3, 0)^T \tag{5.2.24}$$

Durante a fatoração da matriz A na matriz U (5.2.14) se um dos pivôs for nulo deve-se fazer uma permuta de linhas.

Sempre que for feita uma permuta de linhas o sinal do determinante da matriz A do sistema linear (5.2.2) muda , conforme a relação

$$u_{nn} = \begin{cases} \det A \ \text{ para } p \text{ par}, \\ -\det A \ \text{ para } p \text{ impar}, \end{cases} \tag{5.2.25}$$

em que p é o número de permutações.

Exemplo, na seqüência de reduções abaixo, na redução 1.a obtém-se o det A sem a permutação de linhas e na redução 1.b com o permutação de linhas. O sinal do determinante da seqüência da redução 1.b é determinado segundo o número de permutação de linhas, de acordo com (5.2.25). Os elementos das matrizes representadas por colchetes são os valores obtidos antes da divisão pelo pivô, ver (5.2.13).

(A)

$$\begin{pmatrix} \mathbf{9} & 1 & 5 & 6 \\ 1 & 4 & 1 & 3 \\ -8 & 7 & 9 & 2 \\ 2 & 6 & 7 & 4 \end{pmatrix}$$

(A')

$$\begin{pmatrix} \mathbf{35} & 4 & 21 \\ 71 & 121 & 66 \\ 52 & 53 & 24 \end{pmatrix}$$

(A'')

$$\begin{pmatrix} \mathbf{439} & 91 \\ 183 & -28 \end{pmatrix}\begin{bmatrix} 3951 & 819 \\ 1647 & -252 \end{bmatrix}$$

(A''')

$(\mathbf{-827})[-28945]$

redução 1.a

(A)

$$\begin{pmatrix} 9 & 1 & 5 & 6 \\ \mathbf{1} & 4 & 1 & 3 \\ -8 & 7 & 9 & 2 \\ 2 & 6 & 7 & 4 \end{pmatrix}$$

(A')

$$\begin{pmatrix} -35 & -4 & -21 \\ 39 & 17 & 26 \\ \mathbf{-2} & 5 & -2 \end{pmatrix}$$

(A'')

$$\begin{pmatrix} -229 & 26 \\ \mathbf{183} & -28 \end{pmatrix}$$

(A''')

$(\mathbf{827})[-1654]$.

redução 1.b

Na redução 1.b foi feita três permutações de linhas, logo o sinal do det A = 827, redução 1.b, é oposto ao sinal do det A = – 827, redução 1.a

Na redução 1.a os elementos crescem muito em número de dígitos, para minimizar o crescimento desses números foi escolhido em cada passo da redução 1.b o menor elemento, em valor absoluto e diferente de zero, da primeira coluna de cada passo.

Assim, para o sistema linear (5.2.8), a cada passo k da transformação da matriz completa desse sistema linear deve-se escolher um pivô $a_{kk}^{k-1} \neq 0$, ou seja, fazer um pivotamento parcial da linha k com a linha r, tal que

$$|a_{rk}^{k-1}| = \min_{i=k,\ldots,n} |a_{ik}^{k-1}|, \qquad (5.2.26)$$

em que $k \neq r$.

O pivotamento parcial no método proposto por Fox difere do pivotamento no método da eliminação de Gauss (seção 4.6), neste, o pivô escolhido é o maior elemento, em valor absoluto e diferente de zero, da primeira coluna de cada passo.

5.2.1 Algoritmo do método proposto por Fox - Seja a matriz A^*, matriz completa do sistema linear (5.2.8). Este algoritmo computa a transformação da matriz A^*, usando a estratégia de pivotamento parcial, de modo que, a matriz A passe a ser triangular superior, p é obtido conforme a permuta de linhas.

for $r = 1, n\text{-}1$

Determinar $p \in \{r, r+1, \ldots, n\}$ **de maneira que** $|a_{p\,r}| = \min_{r \leq i \leq n} |a_{i\,r}|$ **com** $a_{i\,r} \neq 0$

$r_r := p$

permuta de a_{rj} e a_{pj} $\qquad (j = r, \ldots, n)$

$w_j := a_{rj}$ $\qquad (j = r+1, \ldots, n)$

for $i = r+1, \ldots, n$

$a_{ir} := \eta$

for $j = r+1, \ldots, n+1$

$a_{ij} := a_{ij}\, a_{rr} - a_{rj}\, a_{ir}$

if r != 2

$a_{ij} := a_{ij} / a_{r-1,\, r-1}$

end if

end for

end for

end for

5.2.2 Algoritmo do Vetor Auxiliar e do Vetor Solução do Sistema - Dada uma matriz triangular superior U (algoritmo 4.3) de ordem n, não singular, e o vetor **b** de ordem n, este algoritmo encontra o vetor **y** e o vetor solução racional **x** do sistema linear, de modo que $U\,\mathbf{y} = \mathbf{b}$ e $x_i = y_i / a_{nn}$, $i = 1,...,n$ em que $a_{nn} = \det A$..

begin

$\lambda = n\text{-}1$ (onde n é o número de equações do sistema)

$y_n := a_{n.q} \times a_{nn}$ $(q = n+1)$

$x_n := y_n / a_{nn}$

$i := \lambda$

while $i \geq 1$ **do**

$\rho := i + 1$

$\varepsilon := 0$

for $j = \rho, \ldots, n$

$\varepsilon = \varepsilon + a_{ij}\, y_j$

$y_i := (a_{nn}\, a_{iq} - \varepsilon)/a_{ii}$

$i := i - 1$

end for

end while

end begin

5.2.3 Algoritmo de Euclides - Dados o vetor **y** e $a_{nn} = \det A$.(algoritmo 5.2.2), este algoritmo é usado na simplificação da solução $x_i = y_i / a_{nn}$, $i = 1, ..., n$ do sistema linear.

for $i = 1, \ldots, n$

$\rho := y_i$

$\varepsilon := a_{nn}$

while $\varepsilon \neq 0$ **do**

$\lambda := \mathrm{mod}(\rho, \varepsilon)$

$\rho := \varepsilon$

$\varepsilon := \lambda$

end while

$m = \varepsilon$

end for

Os métodos diretos usados na solução de sistemas lineares são, algumas vezes, comparados quanto a eficiência, servindo de base o número de operações aritméticas requerido. A partir dos algoritmos (5.2.1) e (5.2.2), é possível fazer uma contagem do número de operações. Durante o processo de triangularização do sistema linear essa contagem é feita como segue:

tabela 5.1: N° de operações realizadas pelo algoritmo (5.2.1)

Passo	Divisões	Multiplicações	Adições
1	-	$2n(n-1)$	$n(n-1)$
2	$(n-1)(n-2)$	$2(n-1)(n-2)$	$(n-1)(n-2)$
3	$(n-2)(n-3)$	$2(n-2)(n-3)$	$(n-2)(n-3)$
.	.	.	.
.	.	.	.
.	.	.	.
$n-1$	2×1	$2 \times 2 \times 1$	2×1
Total	$\sum_{k=2}^{n-1} k(k-1)$	$2\sum_{k=2}^{n} k(k-1)$	$\sum_{k=2}^{n} k(k-1)$

Das relações

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \qquad \text{e} \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}, \tag{5.2.27}$$

obtém-se $\frac{1}{3}(n^3 - 3n^2 + 2n)$ divisões,

$\frac{2}{3}(n^3 - n)$ multiplicações e

$$\frac{1}{3}(n^3 - n) \qquad \text{adições.} \qquad (5.2.28)$$

No cálculo do vetor auxiliar por substituições regressivas e do vetor solução por substituições progressivas obtém-se

$$\begin{aligned} & n && \text{divisões,} \\ & 1+2+3+...+n) = \frac{n(n+1)}{2} && \text{multiplicações e} \\ & 1+2+3+...+(n-1) = \frac{n(n-1)}{2} && \text{adições.} \end{aligned} \qquad (5.2.29)$$

Resultando:

$$\begin{aligned} & \frac{n^3}{3} - n^2 + \frac{5n}{3} && \text{divisões,} \\ & \frac{2n^3}{3} + \frac{n^2}{2} - \frac{n}{6} && \text{multiplicações e} \\ & \frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} && \text{adições.} \end{aligned} \qquad (5.2.30)$$

A contagem dessas operações conduzirá, aproximadamente, a um valor igual a $4n^3/3 - 2n/3$. Portanto, para um valor elevado de n, o termo dominante é $4n^3 / 3$.

A tabela 5.2 mostra, para n elevado, o termo dominante do número de divisões, multiplicações e adições realizadas pelos métodos diretos [Patel94, Johnston82, Stainberg74] e proposto por Fox.

tabela 5.2: N° de operações realizadas pelos métodos

Métodos	**N° de operações**
fatoração *LU*	$2n^3/3$
eliminação de Gauss	$2n^3/3$
eliminação de Jordan	n^3
proposto por Fox	$4n^3/3$

Fazendo uma comparação entre os métodos: a eliminação de Jordan requer aproximadamente 50% de operações a mais que os métodos de fatoração *LU* eeliminaçãode Gauss, já o método proposto por Fox requer duas vezes mais operações aritméticas que o método de fatoração *LU* e eliminação de Gauss. Na resolução de um sistema linear, qualquer um dos métodos pode ser utilizado. Nos métodos diretos os erros de arredondamentos surgidos durante a transformação do sistema interferem na obtenção da solução exata do sistema linear, já o método proposto por Fox, se na transformação do sistema os coeficientes não crescerem muito, em número de dígitos, permite obter a solução exata do sistema linear.

5.3 Aritmética em múltipla precisão

Na seção 5.2, mostra como o resultado de uma operação pode ser afetado quando o computador não dispor de precisão suficiente para representar esse resultado. Uma maneira de representar esse resultado é utilizar aritmética em múltipla precisão (simula números inteiros ou em ponto flutuante de vários tamanhos, usando cadeias de caracteres).

Na resolução de sistemas lineares, pelo método proposto por Fox, o uso dessa aritmética é de fundamental importância na obtenção da solução exata. Por exemplo, no sistema linear , não singular

$$\begin{aligned} 2977x_1 + 2971x_2 + 2971x_3 &= 8919, \\ 2971x_1 - 2977x_2 + 2971x_3 &= 8919, \\ 2971x_1 + 2971x_2 + 2977x_3 &= 8919, \end{aligned} \tag{5.3.1}$$

quando resolvido com uma precisão de 8 dígitos a solução obtida não é exata.

Na primeira redução do sistema linear (5.3.1) obtém-se

$$
\begin{aligned}
2977x_1 + 2971x_2 + 2971x_3 &= 8919, \\
35688x_2 + 17826x_3 &= 53514, \\
17826x_2 + 35688x_3 &= 53514.
\end{aligned}
\tag{5.3.2}
$$

A próxima redução não é possível devido a precisão ser insuficiente para representar os resultados

$$
\det\begin{pmatrix} 35688 & 17826 \\ 17826 & 35688 \end{pmatrix} = 955867068 \quad \text{e} \quad \det\begin{pmatrix} 35688 & 53514 \\ 17826 & 53514 \end{pmatrix} = 955867068. \tag{5.3.3}
$$

Se a precisão usada fosse de no mínimo 9 dígitos o sistema linear (5.3.1) seria resolvido sem erros. O exemplo (5.3.1) mostra bem a necessidade de se usar uma precisão maior para se obter a solução exata.

Na aritmética convencional em ponto flutuante o número de dígitos do expoente e da mantissa são limitados, como o conjunto de números reais é infinito, então não é possível representar todos os seus elementos utilizando essa aritmética, ver seção 5.2.

No método proposto por Fox todas as operações são realizadas com números inteiros, como a aritmética utilizada neste trabalho é em ponto flutuante, então, é preciso representar esses números inteiros em ponto flutuante, com o uso da aritmética de ponto flutuante em múltipla precisão é possível representar qualquer número inteiro x em ponto flutuante, para isso é necessário que

$$
\mathrm{n}(x) = e, \tag{5.3.4}
$$

em que: $\mathrm{n}(x)$ é o número de dígitos do número inteiro x e e é o expoente da base do sistema em ponto flutuante (seção 3.4).

Por exemplo, para representar o número inteiro 12310245164 em ponto flutuante é necessário que $\mathrm{n}(f) \geq 11$ e $e = 11$, assim sua representação em decimal flutuante fica

$$.12310245164 \times 10^{11}$$

5.4 Cálculo da Precisão

Na seção 5.2, os coeficientes das $n-1$ últimas equações do sistema linear (5.2.8) após a eliminação de x_1 , foram obtidos pelos determinantes de (5.2.10)

$$a'_{rs} = a_{11}\, a_{rs} - a_{r1}\, a_{1s} \quad \text{e} \quad b'_r = a_{11}\, b_r - a_{r1}\, b_1 , \quad r = 2, 3,...,n \ \text{e}\ s = 1, 2,...,n \tag{5.4.1}$$

A precisão requerida para solução exata do sistema linear (5.2.8) está relacionada com o número de dígitos do maior dos produtos, em valor absoluto, dos termos

$$a_{11}\, a_{rs} , \quad a_{r1}\, a_1 , \quad a_{11}\, b_r \quad \text{ou} \quad a_{r1}\, b_1 \tag{5.4.2}$$

Seja pq o maior produto de (5.4.2), em valor absoluto, e m o número de dígitos desse produto, dado por

$$m = \text{n}(pq) \tag{5.4.3}$$

Como visto na seção 5.3, pode-se obter resultados errados num cálculo por falta de precisão. Uma maneira de evitar esses erros no método proposto por Fox é informar a precisão quando o programa for executado ou utilizar uma precisão superior.

Na resolução de sistema linear

$$A\mathbf{x} = \mathbf{b}, \tag{5.4.4}$$

com estrutura simétrica do tipo

$$\begin{pmatrix} 1 & k & k & \dots & k \\ k & 1 & n & \dots & k \\ k & k & 1 & \dots & k \\ \vdots & \vdots & \vdots & \dots & \vdots \\ k & k & k & \dots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{5.4.5}$$

em que k = 3, 9, 31, 99, 316, 999, 3162, etc., k foi escolhido dessa forma devido o produto pq, conforme (5.4.3), ser máximo para 1, 2, 3, 4, 5, 6, 7, etc. dígitos, respectivamente, pelo método proposto por Fox, utilizando o pacote FP de aritmética em múltipla precisão [Smith91], verificou-se uma estimativa da precisão requerida (tabela 5.3) para obter a solução exata do sistema linear (5.4.5), a partir de m e n (número de equações do sistema linear).

tabela 5.3: Estimativa da precisão requerida para resolver o sistema (5.4.5)

N° de eq. do sist. linear	m (número de dígitos do maior produto dos termos de 5.4.2)						
	1	2	3	4	5	6	...
2	2	3	4	5	6	7	...
3	3	5	7	9	11	13	...
4	4	7	10	13	16	19	...
5	5	9	13	17	21	25	...
6	6	11	16	21	26	31	...
7	7	13	19	25	31	37	...
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	
n	n	$2n$-1	$3n$-2	$4n$-3	$5n$-4	$6n$-5	...

Para obter a solução exata do sistema linear (5.4.5) com n equações a precisão requerida é dada por

$$mn - m + 1. \tag{5.4.6}$$

Por exemplo, no sistema linear (5.4.5) para $n = 4$ e $k = 999$ o número de dígitos do maior produto (pq) em valor absoluto é $m = \mathrm{n}(pq) = 6$ dígitos. Da relação (5.4.6) a precisão requerida para solução exata desse sistema é de 19 dígitos.

Uma outra maneira usada para justificar os dados da tabela 5.3 foi observar na matriz transformada U (5.2.14) o número de dígitos do produto

$$u_{nn} \times u_{n-2\ n-2}. \tag{5.4.7}$$

Como visto na seção anterior qualquer número inteiro pode ser representado em ponto flutuante, então, resolvendo o sistema linear (5.4.5) utilizando o pacote FP, chega-se as precisões da tabela 5.3 a partir do produto (5.4.7). No exemplo, anterior $n = 4$ e $k = 999$ após a transformação do sistema linear a matriz U fica

$$U = \begin{pmatrix} 1 & 999 & 999 & 999 \\ 0 & -997002 & -998000 & -997002 \\ 0 & 0 & -995007996 & 995007996 \\ 0 & 0 & 0 & -2980047928064 \end{pmatrix} \tag{5.4.8}$$

como $u_{nn} = u_{44} = -2980047928064$ e $u_{n-2\ n-2} = u_{22} = -997002$, de (5.4.7) obtém-se o resultado em ponto flutuante

$$.2971113744375664128 \times 10^{19} \tag{5.4.9}$$

para representá-lo na forma inteira é necessário 19 dígitos, como o resultado desse produto é o maior número inteiro surgido durante a transformação do sistema linear, então, a precisão requerida é dada por e (expoente da base do sistema em ponto flutuante), seção (3.4) e (5.3).

Resolver sistemas lineares com um grande número de equações pelo método proposto por Fox, mesmo utilizando aritmética em múltipla precisão, não é uma boa opção devido a precisão requerida ser muito elevada. Por exemplo, no sistema linear (5.4.5) para $n = 500$ e

m = 3 pela relação (5.4.6) a precisão requerida para obter a solução exata é de 1499 dígitos, tornando-se inviável o uso desse método.

O pacote FP de aritmética em múltipla precisão, citado nesta seção, foi concebido por David M. Smith, ele é composto por um conjunto de subrotinas em linguagem Fortran que desempenham operações aritméticas em múltipla precisão [Smith91].

Além do pacote FP, existem outros pacotes em múltipla precisão como o MPFUN de David B.Bailey [Bailey93] e o BMP de Richard P. Brent [Brent78], esses não foram utilizados devido os recursos disponíveis durante a fase de estudo e implementação deste trabalho.

Capítulo VI

Testes e Resultados

6.1 Introdução

Este capítulo apresenta a descrição das matrizes dos sistemas lineares usados em teste, observações sobre a implementação do método proposto por Fox, os testes, os resultados obtidos nos testes realizados e uma análise desses resultados tomando como base o tempo de execução e a qualidade da solução obtida.

6.2 Descrição das matrizes

No sistema linear não singular,

$$A\mathbf{x} = \mathbf{b}, \tag{6.2.1}$$

usado nos testes, A é uma matriz simétrica de ordem n e os componentes do vetor **b** são iguais a 1.

$$\begin{pmatrix} n+1 & n & n & \dots & n \\ n & n+1 & n & \dots & n \\ n & n & n+1 & \dots & n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & n & n & \dots & n+1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{6.2.2}$$

Outras estruturas de sistemas lineares usados nos testes:

$$a_{ij} = \begin{cases} 1 & \text{se} \quad i = j, \\ i + j & \text{se} \quad i > j, \\ i - j & \text{se} \quad i < j, \end{cases} \quad e \quad b_i = \sum_{j=1}^{n} a_{ij}, \tag{6.2.3}$$

$i, j = 1, \dots, n.$

$$a_{ij} = \begin{cases} i + j & \text{se} \quad i \geq j, \\ 1 & \text{se} \quad i < j, \end{cases} \quad e \quad b_i = \sum_{j=1}^{n} a_{ij}, \tag{6.2.4}$$

$i, j = 1, \dots, n.$

Para os sistemas lineares (6.2.2) os testes foram realizados para 80, 120, 160, 200 e 240 equações.

Para fazer as comparações práticas entre os métodos diretos e proposto por Fox, foram escolhido os sistemas lineares (5.2.2) (pag.37), (6.2.2), (6.2.3) e (6.2.4). O sistema linear(6.2.2) foi o mais explorado nos testes por minimizar o crescimento dos números, em número de dígitos, no processo de triangularização da matriz completa do sistema linear pelo método proposto por Fox.

6.3 Observações sobre implementação

Na implementação do método proposto por Fox decidiu-se:.

- utilizar sistemas lineares em que os coeficientes e termos independentes sejam números inteiros;

- adotar critérios de interrupções no caso de sistemas lineares singulares;

- utilizar, sempre que possível, uma estimativa de precisão inicial ou uma precisão fixa na obtenção da solução exata;

- permitir o uso da aritmética convencional desde que durante a transformação da matriz completa do sistema linear não surjam elementos com o número de dígitos superior ao limite disponível nessa aritmética, caso contrário, deve-se usar o pacote FP de aritmética em múltipla precisão de [Smith91];

- não adotar nenhum método para economizar espaço de memória para armazenar a matriz completa do sistema linear e a matriz triangularizada, essas matrizes são inteiramente armazenadas; e

- usar o algoritmo de Euclides para simplificação da solução exata racional dos sistemas lineares.

6.4 Resultados dos testes

Nesta seção será mostrado os resultados dos testes das rotinas que implementam os métodos diretos e proposto por Fox, quanto ao tempo de execução e a qualidade da solução obtida

Utilizando a aritmética convencional, os testes realizados utilizando o AIX XL Fortran Compiler numa máquina IBM PowerPC da UFPB Campus II, para os sistemas lineares (5.2.2) (pag. 37), (6.2.2), (6.2.3) e (6.2.4) os métodos diretos e proposto por Fox apresentaram os resultados, listados a seguir:

A tabela 6.3 apresenta a solução do sistema linear (5.2.2) (pag. 37) pelos métodos diretos e proposto por Fox.

A tabela 6.4 apresenta os erros cometidos na obtenção da solução do sistema linear (5.2.2) (pag. 37) pelos métodos diretos e proposto por Fox.

A tabela 6.5 apresenta as soluções do sistema linear (6.2.2) com 80, 120, 160, 200 e 240 equações. No método proposto por Fox todas as soluções obtidas foram iguais e exatas, nos racionais, já nos métodos diretos como as soluções obtidas não foram iguais, optou-se pela menor e maior raiz.

As tabelas 6.6 apresenta os erros absolutos cometidos na obtenção das soluções da tabela 6.5, pelos métodos diretos e proposto por Fox.

A tabela 6.7 apresenta o resultado do tempo de execução, em segundos, na obtenção das soluções da tabela 6.5.

Utilizando o pacote FP de aritmética em múltipla precisão de [Smith91] o método proposto por Fox obteve as soluções exatas dos sistemas lineares (6.2.3) e (6.2.4) com até 48 equações.

tabela 6.3: Solução do sistema linear (5.2.2)

Raízes / Métodos	x_1	x_2	x_3	x_4
fatoração *LU*	2.0000000000000000	−1.0000000000000000	−3.0000000000000000	0.0000000000000001
elim. de Gauss	2.0000000000000000	−1.0000000000000000	−3.0000000000000000	0.0000000000000001
elim. de Jordan	2.0000001901841316	−1.0000001348331682	−3.0000002216992243	0.0000001129747224
proposto por Fox	2.0000000000000000	−1.0000000000000000	−3.0000000000000000	0.0000000000000000

tabela 6.4: Erros absolutos cometidos na obtenção da solução do sistema (5.2.2)

Raízes / Métodos	x_1	x_2	x_3	x_4
fatoração *LU*	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.1×10^{-15}
elim. de Gauss	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.1×10^{-15}
elim. de Jordan	$0.19018413160\times10^{-6}$	$0.1348331682\times10^{-6}$	$0.2216992243\times10^{-6}$	$0.1129747224\times10^{-6}$
proposto por Fox	0.0000000000000000	0.0000000000000000	0.0000000000000000	0.0000000000000000

tabela 6.5: Soluções de sistemas lineares com estrutura (6.2.2)

Métodos / Nº de equações	fatoração *LU*		eliminação de Gauss		eliminação de Jordan		proposto em por Fox
	menor raiz ($\times 10^{-4}$)	maior raiz ($\times 10^{-4}$)	menor raiz ($\times 10^{-4}$)	maior raiz ($\times 10^{-4}$)	menor raiz ($\times 10^{-4}$)	maior raiz ($\times 10^{-4}$)	raizes exatas
80	1.562255897516	1.562255897516	1.562255897516	1.562255897516	1.562255823620	1.562255900265	$\frac{1}{6401}$
120	0.694396222484	0.694396222485	0.694396222484	0.694396222485	0.694396220222	0.694396255530	$\frac{1}{14401}$
160	0.390609741807	0.390609741807	0.390609741807	0.390609741807	0.390609729452	0.390609746049	$\frac{1}{25601}$
200	0.249993750156	0.249993750156	0.249993750156	0.249993750156	0.249993741871	0.249993757561	$\frac{1}{40001}$
240	0.173608097081	0.173608097082	0.173608097081	0.173608097082	0.173608095075	0.173608110717	$\frac{1}{57601}$

tabela 6.6: Erros absolutos* cometidos na obtenção das soluções da tabela 6.5

Métodos / Nº de equações	fatoração *LU*		eliminação de Gauss		eliminação de Jordan		proposto por Fox
	menor raiz ($\times 10^{-20}$)	maior raiz ($\times 10^{-20}$)	menor raiz ($\times 10^{-20}$)	maior raiz ($\times 10^{-20}$)	menor raiz ($\times 10^{-20}$)	maior raiz ($\times 10^{-20}$)	raiz exata ($\times 10^{-20}$)
80	130	130	130	130	7389601303	274898697	0
120	5496	4504	5496	4504	22625496	33045450	0
160	393	393	393	393	123549607	424203932	0
200	2461	3521	2461	3521	82852461	74047539	0
240	6479	3521	6479	3521	20066479	13635352	0

* Devido os erros serem muito pequenos os cálculos na tolerância de (10^{-20}) foram feitos utilizado o pacote FP em múltipla precisão de [Smith91]

tabela 6.7: Tempo de execução, em segundos, em sistemas lineares com a estrutura (6.2.2)

Métodos / N° de equações	fatoração *LU*	elim. de Gauss	elim. de Jordan	proposto por Fox
80	0.06	0.06	0.19	0.15
120	0.14	0.13	0.56	0.43
160	0.26	0.27	1.27	0.96
200	0.49	0.48	2.45	1.82
240	0.79	0.80	5.27	3.15

Utilizando o pacote FP em múltipla precisão de [Smith91], o método proposto por Fox obteve as soluções exatas dos sistemas da tabela 6.7 em 8.44, 27.97, 65.75, 127.90 e 220.09 segundos, respectivamente.

Capítulo VII

Conclusões

Baseados nos estudos realizados verificou-se que o bom desempenho do método proposto por Fox depende da estrutura da matriz do sistema linear e do número de equações.

Na resolução do sistema linear não singular

$$A\mathbf{x} = \mathbf{b}, \tag{7.1}$$

com $2 \leq n \leq 400$ equações, em que a matriz A tem estrutura simétrica do tipo

(7.2)

e **b** de componentes iguais a 1, a solução obtida utilizando esse método foi exata, nos racionais. Em sistemas lineares com outras estruturas e com um reduzido número de equações os resultados também foram exatos. Se esse número de equações for elevado a resolução por esse método torna-se inviável, devido ao surgimento de elementos com muitos dígitos durante o processo de triangularização da matriz completa do sistema. Por exemplo, para

obter a solução exata de um sistema linear qualquer, não singular, com $n = 800$ e $m = 5$ dígitos é necessário 3996 dígitos de precisão, seção 5.4.

Dependendo do número de equações e da estrutura do sistema linear, pode-se usar a aritmética convencional ou em múltipla precisão. A dificuldade no uso da aritmética convencional é quando o número de dígitos de um elemento obtido durante a transformação do sistema superar o limite da precisão disponível no computador, essa dificuldade pode ser minimizada usando uma aritmética em múltipla precisão.

A tabela 7.1 mostra que a partir de uma precisão inicial, do número de dígitos m e da relação 5.4.6 é possível calcular o número máximo de equações que um sistema linear simétrico deve ter para ser resolvido nesta precisão.

tabela 7.1: N° máximo de equações (n) de um sistema linear simétrico a ser resolvido numa precisão de 50 dígitos segundo o número m

m	1	2	3	4	5	6	7	8	9	10 à 12	13 à 16	17 à 24	25 à 49
n	50	25	17	13	10	9	8	7	6	5	4	3	2

Por exemplo, para se resolver um sistema linear simétrico com $m = 4$ dígitos usando uma precisão de 50 dígitos o número de equações deverá ser $2 \leq n \leq 13$.

A tabela 7.1 mostra n inversamente proporcional a m numa precisão fixa. Seguindo essa proporcionalidade, os resultados obtidos para sistemas lineares simétricos com até 50 equações foram exatos.

Para se resolver um sistema linear qualquer não singular por esse método:

- os elementos da matriz completa do sistema linear devem ser números inteiros

- a precisão, tabela 5.3, para obtenção da solução exata deverá ser informada ao programa que implementa o método, caso a precisão seja insuficiente ele falha, ou

- usar a precisão disponível no pacote FP em múltipla precisão de [Smith91].

A vantagem do método proposto por Fox em relação aos métodos diretos é a ausência de divisões com resultados não exatos implicando na eliminação dos erros de arredondamentos, isso favorece o método quando na obtenção da solução exata do sistema linear.

As desvantagens do método proposto por Fox em relação aos métodos diretos são a utilização da aritmética em múltipla precisão e o crescimento dos números, em número de dígitos, a medida que o número de operações realizadas crescem, se esse crescimento for muito elevado inviabiliza o uso desse método.

Considerando vantagens e desvantagens, o método proposto é freqüentemente de importância suficiente a justificar seu uso quando na obtenção da solução exata racional de sistemas lineares em que o número de equações não seja muito elevado.

Trabalhos futuros

A finalidade deste trabalho foi o estudo e implementação do método proposto por Fox para obtenção da solução exata, nos racionais, de sistemas lineares utilizando o pacote de aritmética em múltipla precisão de [Smith91].

Para trabalhos futuros sugere-se:

- estudo do cálculo da precisão requerida para resolver sistemas lineares em geral e com outras estruturas específicas.

- implementar o método proposto em outros pacotes de múltipla precisão como o BMP [Brent78], MPFUN[Bailey93], etc.

- implementar um pacote específico de aritmética em múltipla precisão para ser usado no método proposto.

Referências Bibliográficas

[Albrecht73] Albrecht, Peter, *Análise Numérica*, Livros Técnicos e Científicos Editora S.A, Rio de Janeiro, 1973.

[Bailey93] Bailey, D. B., *"Algorithm 719 - Multiprecision Translation and Execution of FORTRAN Programs"*, ACM Transactions on Mathematical Software, Vol. 9, No. 3, pp. 288-319, 1993.

[Brent78] Brent, R. P., *A FORTRAN multiple-precision arthmetic package*, ACM Transactions on Mathematical Software, Vol. 4, No. 1, pp. 57-70, 1978.

[Cabay77] Cabay, S. and Lam, T. P. L., *"Congruence Techniques for the Exact Solution of Integer Systems of Linear Equations"*, ACM Transactions on Mathematical Software, Vol. 3, No. 4, pp. 386-397, 1977.

[Dongarra79] Dongarra, J. J., Moler, C. B., Bunch, J. R. and Stewart, G. W., *Linpack Users' Guide*, Siam, Philadelphia, 1979.

[Dorn72] Dorn, W. S. and McCracken, Daniel D., *Numerical Methods with Fortran IV Case Studies*, John Wiley & Sons,Inc., New York, 1972.

[Figueiredo89] Figueiredo, M. A. B., *Um Pacote de Aritmetica de Múltipla Precisão*, Dissertação de Mestrado, Departamento de Sistemas e Computação, Universidade Federal da Paraiba, Campina Grande, 1989.

[Fox64] Fox, L., *An Introduction to Numerical Linear Algebra*, , Oxford University Press, Oxford, 1964.

[Golub96] Golub, G. H. & Loan, C. F. V., *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[Hattori94] Hattori, Mario T. and Queiroz, B. C. N., *Métodos e Software Numéricos*, Manuscrito, Departamento de Sistemas e Computação, Universidade Federal da Paraiba, Campina Grande - PB, 1994.

[Hopkins88] Hopkins Tim and Phillips Chris, *Numerical Methosd in Practice: using a NAG Library*, Addison-Wesley Publishing Company, Inc., New York, 1988.

[Hostetter91] Hostetter Gene H., Santina M. S. and D'Carpio-Montalvo P., *Analitical, Numerical, and Computational Methods for Science and engineering*, Prentice-Hall International Editions, London,1991.

[Johnston82] Johnston, Robert L., *Numerical Methods*, John Wiley & Sons, New York, 1982 .

[Ketter69] Ketter, Robert L., *Modern Methods of Engineering Computation*, McGraw-Hill Book Company, New York, 1969.

[Knuth69] Knuth, Donald E., *The Art of Computer Programming • Seminumerical Algorithms*, vol. 2, Reading, Addison-Wesley, Massachussets, 1969.

[Lipson81] Lipson, John D., *Elements of Algebra and Algebraic Computing*, Addison • Wesley Publishing Company, U.S.A., 1981.

[Ludovice98] Ludovice, D., *Subrotina para Resouçãor de Sistemas Lineares pelo Método de Jordan*, Georgia Institute of Technology, Atlanta, 1998. http://www.chemse.getech.edu/~che2210/projs95/hwsol4.html.

[MaClellan77]MaClellan, M. T., *"A Comparison of Algorithms for the Exact Solution of Linear Equations"*, ACM Transactions on Mathematical Software, Vol. 3, No. 2, pp. 147-158, 1977.

[MaClellan73] MaClellan, M. T., *" The Exact Solution of Systems of Linear Equations with Polynomial Coefficients"*, ACM Transactions on Mathematical Software, Vol. 20, No. 4, pp. 563-588, 1973.

[MaClellan77]MaClellan, M. T., *"The Exact Solution of Linear Equations with Rational Function Coefficients"*, ACM Transactions on Mathematical Software, Vol. 3, No. 1, pp. 1-25, 1977.

[Patel94] Patel, Vital A., *Numerical Analysis for Werth*, Sauders College, pp 175-193, 1994

[Ramos96] Ramos, Carlos V. da Costa., *Aceleração de Métodos Iterativos para solução de Sistemas Lineares - avaliação crítica*, Dissertação de Mestrado, Departamento de Sistemas e Computação, Universidade Federal da Paraiba, Campina Grande, 1996.

[Schreiner92] Schreiner, W. and Stahi, V., *The Exact Solution of Linear Equations Systems on Shared Memory Multiprocessor*, Technical Report, Johannes Kepler University, Austria, 1992.

[Soares83] Soares, M. N. A., *Estudo e Implementação de Métodos Iterativos para Solução de Sistemas de Equações Lineares Integrados a uma Biblioteca*

Numérica, Dissertação de Mestrado, Departamento de Sistemas e Computação, Universidade Federal da Paraiba, Campina Grande, 1983.

[Smith91] Smith, David M., *"Algorithm 693: A FORTRAN package for Floating Point Multiple-precision Arithmetic"*, ACM Transactions on Mathematical Software,Vol. 17, No. 2, pp. 273283, 1991.

[Steinberg74] Steinberg, David I., *Computational Matrix Algebra,* McGraw-Hill, Inc., U.S.A, 1974.

[Young71] Young, D. M., *Interative Solution of Large Linear Systems*, Academic Press, New York, 1971.