JOÃO BOSCO LUGNANI
Licenciado em Matemática

O PROBLEMA DOS SISTEMAS DE EQUAÇÕES LINEARES MAL CONDICIONADOS E SUAS IMPLICAÇÕES EM GEODÉSIA

Tese de Grau de
"MESTRE EM CIÊNCIAS"

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE TECNOLOGIA
DEPARTAMENTO DE GEOCIÊNCIAS
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIAS GEODÉSICAS

Curitiba, setembro de 1975

JOÃO BOSCO LUGNANI

LIC. MATEMÁTICA

O PROBLEMA DOS SISTEMAS DE EQUAÇÕES LINEARES MAL CONDICIONADOS E SUAS IMPLICAÇÕES EM GEODÉSIA

TESE DE GRAU DE MESTRE EM CIÊNCIAS APRESENTADA AO CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIAS GEODÉSICAS DO DEPARTAMENTO DE GEOCIÊNCIAS, SETOR DE TECNOLOGIA DA UNIVERSIDADE FEDERAL DO PARANÁ

CURITIBA - PARANÁ 1975

A MEUS PAIS MARIA BERNARDINO E JOSÉ LUGNANI

QUERO EXTERNAR AGRADECIMENTOS

ao Doutor Camil Gemael, coordenador do Curso de Pós-Graduação em Ciências Geodésicas pela orientação e estímulo dispensados na elaboração deste trabalho;

ao Doutor Urho A. Uotila, Chefe do Departamento de Ciências Geodésicas da Ohio State University pelo aconselhamento do tema, por ocasião de sua estada, como professor visitante do Curso de Pós-Graduação em Ciências Geodésicas e pela posterior remessa de valiosas fontes de consultas;

ao Doutor Jacob Pallis que prontamente nos orientou na obtenção de literatura para pesquisa quando recorremos ao IMPA;

à CAPES pela bolsa de estudos concedida;

ao Banco Nacional de Desenvolvimento Econômico, cujo suporte financeiro permitiu nossa vinculação provisória ao Curso de Pós-Graduação;

às demais pessoas que apresentaram sugestões ou que de alguma forma contribuiram na elaboração deste trabalho.

SINOPSE

Este trabalho consiste de uma análise da condição dos sistemas de equações lineares e suas consequências na solução numérica; estuda as diferentes formas de identificar o mau condicionamento e os procedimentos recomendáveis para evitar ou minimizar as perturbações da solução, oriundas dos erros de arredondamento, da inadequação de escala ou da má escolha de pivô. Estuda ainda o limite das perturbações da solução, causadas pela imprecisão dos coeficientes obtidos experimentalmente, caso que ocorre com grande freqüência nos problemas de Geodésia e Fotogrametria.

SYNOPSIS

This paper consists of the analysis of the conditions of the systems of linear equations and their implications in the numerical solutions. It studies the different forms in order to identify the ill-conditioning and the advisable procedures to avoid or minimise the perturbations of the solution deriving from rounding errors, from no scaling or no pivoting.

It also studies the limit of the solution's perturbations deriving from inaccuracy of the experimental coefficients. This case usually happens in geodetic and photogrametric problems.

CONTEÚDO

APÍTULO IV

CAPÍTULO I

INTRODUÇÃO

Ao colocarmos em prática conhecimentos teóricos, não raras vezes nos deparamos com dificuldades nunca antes imaginadas. Estas dificuldades existentes em todo o campo do conhecimento humano são freqüentemente mais difundidas nas Ciências Físicas e Tecnologia.

A utilização dos exuberantes recursos que a tecnologia nos coloca às mãos, tais como calculadoras, computadores, bibliotecas de programas e subprogramas ou mais genéricamente o "hardware" e o "software", nos dão a falsa imagem de que mesmo na prática a ciência matemática não sofre modificações.

Por sua característica de ciência exata, as dificuldades matemáticas de ordem prática não apenas causam impacto ao pesquisador principiante como também vislumbram desafiante campo para investigação.

Freqüentemente o usuário da Matemática que utiliza uma calculadora eletrônica ou um computador para resolver um sistema de equações, uma vez averiguadas as correções do modelo matemático, do programa e dos dados e não havendo discrepâncias perspec-

tíveis dentro de certa expectativa, tende a ter como correta a solução encontrada. Outros mais precavidos às vezes ainda verificam os resíduos através de uma substituição da solução encontrada. No caso de um sistema mal condicionado nem mesmo um pequeno resíduo pode assegurar uma boa solução.

O tema a que nos propusemos analisar neste trabalho, sistemas mal condicionados, é um desses problemas com que não raramente se depara o usuário.

Estando conscientes:

a) Das dificuldades de que se reveste o problema, como bem ressalta o professor Ralston |1| p. 398, "How to calculate - solutions of Ax = b, as accurate as the data warrant, when the system is ill-conditioned is probably the single most difficult problem encountered in the solution of simultaneous linear equations";

b) Das limitações a que estamos sujeitos sob o aspecto experimental quando comparado nosso trabalho isolado a trabalhos de equipes com elementos altamente especializados nos diferentes campos de conhecimentos correlatos;

c) Da razoavelmente escassa fonte de consulta e de outras dificuldades de cunho prático; não poderíamos estabelecer como nosso objetivo a solução do problema.

Na ciência em geral e particularmente em Geodésia e Fotogrametria, o ajustamento desempenha papel fundamental. Há quem diga que a história da geodésia se confunde com a do método dos mínimos quadrados. O problema do ajustamento de redes de triangulação, de nivelamento e de qualquer forma de observação super abundante recai sempre na solução de sistemas de equações lineares (ou linearizadas) simultâneas.

Com o progresso que se verifica no campo da computação

eletrônica e nos instrumentos de observação e registro de dados, o volume de informações cresce dia a dia, semelhantemente se avolumam os cálculos sem entretanto constituirem barreira intransponível. Em conseqüência cresce o número de aplicações dos sistemas e as dimensões destes sistemas que o geodesista ou fotogrametrista deve "manipular". Sobrecarregados que estão com a tarefa de acompanhar o desenvolvimento da Geodésia e da Fotogrametria estes pesquisadores não encontram disponibilidade de tempo para buscar, na escassa e "diluida" literatura, um melhor conhecimento acerca dos sistemas mal condicionados.

Visando contribuir neste setor estabelecemos como objetivos de nossa pesquisa:

- evidenciar o problema e suas conseqüências;
- coletar os estudos já realizados neste campo de conhecimento e oferecê-los ao usuário numa linguagem clara e simples;
- elaborar um conjunto de procedimentos e cuidados recomendáveis.

CAPÍTULO II

O PROBLEMA E SUAS IMPLICAÇÕES EM GEODÉSIA E FOTOGRAMETRIA

1. CONSIDERAÇÕES GERAIS

Sabe-se que a resolução de pequenos sistemas de equações lineares simultâneas é bem conhecida desde a época de Gauss e Legendre. Entretanto o desenvolvimento dos instrumentos de observação, dos instrumentos de registro de dados e de cálculo têm tornado atual o problema que parecia definitivamente solucionado. As necessidades de precisão cada vez mais elevada, de resolução de sistemas cada vez maiores e a crescente aplicação deste ramo da ciência matemática, têm ensejado que matemáticos de renome voltem suas atenções para este aspecto da Matemática Aplicada. Isto tem ocorrido de maneira crescente nas últimas décadas. O primeiro nome, que temos conhecimento, a preocupar-se com o mau condicionamento foi F.R. Moulton |2|, em 1913. Desde então é crescente o número de trabalhos relacionados com o problema.

2. CONCEITOS E EXEMPLOS

Um sistema de equações lineares simultâneas é dito mal condicionado se sua solução é muito "sensível" a "pequenas"

variações nos coeficientes das incógnitas ou nos termos independentes. Vê-se claramente que este é um conceito deficiente sob o aspecto matemático. As expressões "muito sensível" e "pequenas" são vagas. Os sistemas mal condicionados têm a propriedade de ampliar os erros, quer sejam erros de observação, de arredondamento para representação decimal, arredondamento da conversão decimal-binário, arredondamento no processo computacional; erro devido à perda de dígitos significativos em subtrações de números aproximadamente iguais ou erro de linearização. (*)

Mesmo que o cálculo se processe isento de erro se introduzirmos uma pequena variação em um dos coeficientes de um sistema mal condicionado a solução sofrerá acentuada variação. O sistema de equações lineares simultâneas abaixo:

$$\begin{aligned} x_1 + x_2 &= 5 \\ x_1 + 1{,}001\, x_2 &= 5{,}003 \end{aligned} \qquad (2.1)$$

cuja solução exata é $x_1 = 2$ e $x_2 = 3$ é exemplo de um sistema mal condicionado. Se variarmos o termo independente de 5,003 para 4,99 obtemos o sistema:

$$\begin{aligned} x_1 + x_2 &= 5 \\ x_1 + 1{,}001\, x_2 &= 4{,}99 \end{aligned} \qquad (2.2)$$

cuja solução é $x_1 = 15$ e $x_2 = -10$. Nota-se que uma variação da ordem do centésimo num dos termos independentes deu origem a solução completamente diferente.

Se os erros são enormemente ampliados num sistema mal condicionado, torna-se indispensável uma especial atenção a toda forma de erro possível, até mesmo aqueles que pareceriam

---

(*) É usual linearizar uma função usando os dois primeiros termos da série de Taylor.

absolutamente despresíveis em um sistema bem condicionado, onde pequenas variações nos coeficientes produzem pequenas variações na solução. Parece-nos útil o exemplo seguinte. Seja o sistema:

$$\begin{aligned} x_1 + x_2 &= 7 \\ x_1 - x_2 &= 1 \end{aligned} \qquad (2.3)$$

com a solução $x_1 = 4$ e $x_2 = 3$. Fazendo variar um de seus termos independentes de 1 para 1,01 teremos

$$\begin{aligned} x_1 + x_2 &= 7 \\ x_1 - x_2 &= 1{,}01 \end{aligned} \qquad (2.4)$$

cuja solução é $x_1 = 4{,}0025$ e $x_2 = 2{,}9975$. Variação da ordem do centésimo em um dos coeficientes produz variações menores na solução. Justifica-se dizer que pequenas variações nos coeficientes produzem pequenas variações na solução e que o sistema é bem condicionado.

Se pretendemos resolver um sistema de equações lineares mal condicionado em um computador, mesmo que os coeficientes do sistema sejam exatos, dada a conversão decimal-binária não ser exata e ainda devido à limitação de "bits" por palavra e de palavras para o armazenamento de cada dado, teremos erros que ampliados poderão afetar a solução.

3. FATORES QUE AGRAVAM O PROBLEMA EM GEODÉSIA E FOTOGRAMETRIA

Se a solução de um sistema mal condicionado com coeficientes exatos, ao ser computada, pode sofrer variações, podemos esperar que, no caso de coeficientes experimentais, oriundos

de observações com limitado número de casas decimais de precisão, a solução sofra variações consideravelmente maiores. Na realidade isto ocorre nos problemas cujos coeficientes são de origem física e ocorre particularmente em Geodésia e Fotogrametria com bastante freqüência. Vários fatores concorrem para que os sistemas mal condicionados mereçam cuidados especiais nestes setores:

a) Serem os coeficientes experimentais e, muitas vezes, com poucas casas decimais incluindo nestas uma incerteza;

b) As dimensões dos sistemas que se originam no ajustamento de redes de triangulação, de nivelamento, de aerotriangulação e outros são, em geral, enormes. Como frizamos anteriormente, há uma tendência a crescer com o aperfeiçoamento dos computadores e desenvolvimento tecnológico em geral. O número de operações aritméticas necessárias para resolver um sistema de equações por um determinado método, é geralmente função da ordem n da matriz dos coeficientes, função esta que quase sempre envolve a potência cúbica da ordem. Por exemplo, se o método adotado for o da eliminação de Gauss teremos:

$$n \quad \text{divisões}$$

$$\frac{1}{3}n^3 + n^2 - \frac{1}{3}n \quad \text{multiplicações e}$$

$$\frac{1}{3}n^3 + n^2 - \frac{5}{6}n \quad \text{adições} \quad (*) \ .$$

Em geral nos métodos diretos ou exatos, como também são chamados, o número total de operações não varia muito. Vemos portanto que o número de operações necessárias para resolver um

---

(*) Para obter o número de operações para diferentes métodos - ver |3| p. 100.

sistema de equações cresce com o cubo da ordem do mesmo.

Por outro lado ao introduzirmos um trabalho de considerâveis dimensões num computador, temos sempre de nos preocupar - com a economia de memória, para evitar "estouro". Devemos então optar entre precisão simples ou precisão estendida, lembrando - que para o primeiro caso temos um menor número de dígitos binários ou decimais, enquanto no segundo temos um gasto de memória' consideravelmente maior.

Se pretendemos, por exemplo, efetuar o seguinte produto interno

$$S = \sum_{i=1}^{N} a_i \, b_i \qquad (2.5)$$

os cálculos sendo conduzidos em "ponto fixo" e o computador utilizado tendo acumulador que efetua o arredondamento de cada produto, o valor obtido para a (2.5) será:

$$\bar{S} = \sum_{i=1}^{N} a_i \, b_i + \varepsilon$$

$$|\varepsilon| \leq \frac{1}{2} N \, 2^{-t} \qquad (2.6)$$

onde t é o número de dígitos binários e ε erro de arredondamento. Se o computador dispõe de recursos para efetuar o somatório em precisão estendida com um só arredondamento final, então o limite superior de erros ε cai para:

$$|\varepsilon| \leq \frac{1}{2} \, 2^{-t} \qquad (2.7)$$

Este é outro fator que deve ser considerado na opção, uma vez conhecidas as possibilidades do computador a ser utilizado e que, às vezes, pode vir a exigir a precisão estendida. Felizmente a maioria dos computadores modernos permite o arredondamento

único com o limite (2.7) para o erro de arredondamento. Entre - tanto o limite de erros de arredondamento pode assumir a forma (2.6) onde o fator N, num grande sistema, é um número enorme. Isto pode ocorrer nos problemas de ajustamento e afetar sensi - velmente a solução, principalmente se o sistema é mal condicio- nado. (*)

c) Freqüentemente ocorrem nos problemas de Ajustamento, modelos matemáticos como $F(x_a) = \ell_a$, que não são lineares. Tais modelos são linearizados pela série de Taylor, no caso do modelo acima, por exemplo, resultando $Ax + \ell = v$ onde A é u'a matriz retangular m x n de coeficientes dada por:

$$A = \left.\frac{\partial F}{\partial x_a}\right|_{x_a = x_o} = \begin{bmatrix} \frac{\partial F_1}{\partial x_{ai}} \\ \vdots \\ \frac{\partial F_m}{\partial x_{ai}} \end{bmatrix}$$

sendo F_i a i-ésima equação não linear; x o vetor das correções aos parâmetros aproximados dados pelo vetor x_o; o vetor ℓ é função dos valores observados e de x_o, e, v é o vetor das correções aos valores observados. A equação $Ax + \ell = v$ é resolvida - pelo método dos mínimos quadrados |5|. A aproximação realizada utilizou apenas os dois primeiros termos da série, o que acarretou erros nos elementos da matriz A.

d) No sistema m x n, $Ax = b$, inconsistente, o vetor

(*) Semelhantes considerações para computação em ponto flutuantes são feitas por Wilkinson |4|.

x que minimiza a soma dos quadrados dos resíduos r = b - Ax é dado por:

$$x = (A^T A)^{-1} A^T b \qquad (2.8)$$

desde que $A^T A$ seja não singular. Ocorrem aqui dificuldades práticas de mau condicionamento. Consideremos um caso particular onde A seja quadrada e numa escala adequada (*). Neste caso a condição da matriz A, como veremos no cap. IV, pode ser indicada por det(A) (**) e a condição de $A^T A$ é, semelhantemente indicada por $\det(A^T A) = (\det A)^2$. Ocorre que o determinante de u'a matriz normalizada será sempre menor ou igual à unidade. Conseqüentemente:

$$\det(A) > \det(A^T A) \qquad (2.9)$$

o que indica pior condição para a matriz simétrica.

No caso em que a matriz A é retangular m x n, m > n, temos:

$$\det(A^T A) = \sum_p (\det A_p)^2 \qquad (2.10)$$

onde A_p são todas as submatrizes quadradas de ordem n que podemos obter de A. Se o maior det A_p for pequeno comparado com a unidade, o $\det(A^T A)$, dado pela (2.10) será em geral menor ainda indicando que novamente a matriz simétrica é pior condicionada. Taussky |6|, demonstra que a condição de $A^T A$ é pior que a condição de A para A quadrada, usando os números de condição (***) de Neumann-Goldstine e os números de Turing.

(*) Veremos no capítulo IV adequação de escala.

(**) Determinante de A pequeno indica mau condicionamento.

(***) No Cap. IV serão estudados os números de condição.

Os sistemas com os quais trabalhamos em Geodésia e disciplinas afins são, na esmagadora maioria, sistemas de equações normais, portanto com matrizes simétricas e definidas positivas |13|.

Os fatores indicados vêm reforçar a necessidade que têm geodesistas e fotogrametristas de algum conhecimento da má condição e das perturbações da solução que ela não produz, mas amplia.

4. A QUALIDADE DA SOLUÇÃO DE UM SISTEMA

Frequentemente somos encorajados empiricamente a testar a qualidade da solução de um sistema de equações lineares pela grandeza dos elementos do vetor dos resíduos que se origina da substituição das incógnitas por seus valores calculados. Por exemplo, o sistema Ax = b,

$$\begin{aligned} x_1 + 10\, x_2 &= 11 \\ 10x_1 + 101\, x_2 &= 111 \end{aligned} \qquad (2.11)$$

cuja solução exata é $x_1 = x_2 = 1$, normalizado (*) se torna:

$$\begin{aligned} 0{,}0995\, x_1 + 0{,}995\, x_2 &= 1{,}0945 \\ 0{,}0985\, x_1 + 0{,}995\, x_2 &= 1{,}0937 \end{aligned} \qquad (2.12)$$

e o determinante deste sistema é ~ 0,001, o que indica mau condicionamento. Se tomarmos o vetor de componentes $x_1 = 1{,}001$;

(*) No Cap. IV veremos alguns métodos de normalização.

$x_2 = 1{,}01$ como solução aproximada teremos o vetor de resíduos : $r_1 = 0{,}01$; $r_2 = 0{,}01$ e tomando $x_1' = 11{,}1$ e $x_2' = 0$, solução completamente absurda, o vetor de resíduos será: $r_1 = 0{,}01$ e $r_2' = 0$. Vemos que neste exemplo a solução aproximada tem para soma dos quadrados dos resíduos o valor 0,0002, enquanto que uma solução completamente absurda produz como soma dos quadrados dos resíduos o valor 0,0001.

É fácil entender analíticamente o fenômeno. Para isto consideremos o sistema n x n, $Ax = b$ e sejam x e $\hat{x}$ soluções exata e aproximada respectivamente. Teremos então:

$$r = b - A\hat{x}$$

como $b = Ax$, podemos escrever

$$r = Ax - A\hat{x}$$

e

$$x - \hat{x} = A^{-1}r. \qquad (2.13)$$

Desta última igualdade vemos que pequeno resíduo não implica em pequena discrepância entre soluções exata e aproximada. Portanto a análise da qualidade de uma solução não pode prescindir de algum estudo da condição. Segundo o doutor Eisemann |7| o mau condicionamento é um dos maiores obstáculos para a precisão das soluções dos sistemas de equações lineares. Tal problema vem preocupando pensadores de elite das grandes industrias da computação e de outros ramos da Matemática Aplicada. Podemos enfocá-lo por diferentes "pontos de vista", tais como: identificação - do mau condicionamento; estimativa de quais as implicações na solução dos erros dos coeficientes; análise dos efeitos dos erros de arredondamento oriundos do processamento em diferentes precisões e modos computacionais e finalmente técnicas para

melhorar a condição do sistema mal condicionado.

Apesar dos espetaculares softwares providos pelas indústrias da computação, que facilitam enormemente a resolução de problemas e particularmente dos sistemas de equações lineares, por diferentes processos, o usuário deve fazer uma cuidadosa análise dos resultados e quando estes sistemas assumem grandes dimensões, fato não raro nos problemas de ajustamento, deve assumir o controle geral do processamento, desde a escala dos coeficientes; escolha de métodos para solução; opção do modo e precisão computacional até o tipo de erro a ser analisado.

CAPÍTULO III

REQUISITOS BÁSICOS

1. DERIVAÇÃO MATRICIAL

Parece útil tecermos considerações sobre alguns ítens da Matemática bem como algum comentário sobre erro. Entre os ítens que consideraremos, existem alguns que não são facilmente encontrados em nossas bibliotecas, outros, apesar da abundante literatura que existe a respeito, não são de uso muito corrente para nossos usuários da Matemática Aplicada.

1.1. Conceitos

Facilmente encontramos o seguinte conceito de derivação matricial, nos textos de Cálculo Matricial: Seja Y uma matriz cujos elementos y_{ij} são funções de uma variável x. Chama-se derivada de primeira ordem de Y em relação a x a matriz dY/dx , das dimensões de Y, cujos elementos são as derivadas de primeira ordem em relação a x dos correspondentes elementos de Y, |8|, ou seja:

$$Y = \left[y_{ij}\ (x) \right]$$

então

$$\frac{dY}{dx} = \left[\frac{d}{dx}\, y_{ij}(x)\right] \qquad (3.1)$$

Entretanto estamos interessados numa derivação mais generalizada, onde a variável independente é também u'a matriz , e onde a matriz Y pode ser: soma de diversas matrizes; produto envolvendo a matriz variável X e sua transposta X^T e ainda matrizes constantes A, B, C, ...; matriz inversa de X, etc.

Dois tipos de derivação matricial são de particular interesse:

a) Primeiro tipo: derivada da matriz Y retangular k x ℓ em relação a um elemento pré-fixado x_{mn} da matriz X também retangular r x s. Este primeiro tipo é uma generalização, sob certo aspecto, para a derivação inicialmente definida (3.1). Para representá-lo usaremos a seguinte notação:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left[y_{ij}\right] = \left[\frac{\partial y_{ij}}{\partial x_{mn}}\right] \qquad (3.2)$$

com i = 1, 2, ..., k e j = 1, 2, ..., ℓ.

É fácil notar que quando a matriz X degenera em um escalar a (3.2) se particulariza em (3.1).

b) Segundo tipo: derivada de um elemento pré-fixado y_{pq} da matriz Y em relação à matriz X, i.e., derivada de um escalar em relação a u'a matriz, representada por:

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}}\right] y_{pq} = \left[\frac{\partial y_{pq}}{\partial x_{ij}}\right] \qquad (3.3)$$

com i = 1, 2, ..., r e j = 1, 2, ..., s.

Observe-se que em (3.1) e (3.2) a derivada tem as mesmas dimensões de Y e em (3.3) as mesmas dimensões de X.

Vamos exemplificar o primeiro tipo, quando particularizado para (3.1) com funções transcendentes, entretanto a aplicação da (3.2) e (3.3) restringir-se-á à matriz Y, envolvendo multiplicações, adições etc. como especificado acima. Temos interesse numa forma sistemática da derivada de tais produtos.

Exemplo 1. Seja:

$$Y = \begin{bmatrix} x & 3x^3 & 2x^{-5} \\ \operatorname{sen} x & \log_e x & e^x \end{bmatrix}$$

$$\frac{\partial Y}{\partial x_{mn}} = \frac{dY}{dx} = \begin{bmatrix} 1 & 9x^2 & -10x^{-6} \\ \cos x & x^{-1} & e^x \end{bmatrix}$$

Exemplo 2. Sejam:

$y_{pq} = x_{11}x_{32} - x_{31}x_{12}$ um elemento prefixado de Y e

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

a matriz variável. Então o segundo tipo de derivação nos dará:

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}}\right] y_{pq} = \begin{bmatrix} x_{32} & -x_{31} \\ 0 & 0 \\ -x_{12} & x_{11} \end{bmatrix}$$

Exemplo 3. Sejam A, X e Y tais que:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \; ; \; X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

$$Y = \begin{bmatrix} a_{11}x_{11} + a_{12}x_{21} + a_{13}x_{31} & a_{11}x_{12} + a_{12}x_{22} + a_{13}x_{32} \\ a_{21}x_{11} + a_{22}x_{21} + a_{23}x_{31} & a_{21}x_{12} + a_{22}x_{22} + a_{23}x_{32} \end{bmatrix} \quad (3.4)$$

Efetuando o primeiro tipo de derivação de Y dado pela (3.4) temos:

$$\frac{\partial Y}{\partial x_{11}} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & 0 \end{bmatrix} \; ; \; \frac{\partial Y}{\partial x_{12}} = \begin{bmatrix} 0 & a_{11} \\ 0 & a_{21} \end{bmatrix} \; ; \; \text{etc}$$

onde o número total de matrizes resultantes da derivação será r x s. (Lembrar que r e s são número de linhas e colunas de X , no exemplo acima 6). Estas seis matrizes podem ser representa - das, genericamente por:

$$\frac{\partial Y}{\partial x_{mn}} = A \, J_{mn} \quad (3.5)$$

sendo J_{mn} u'a matriz de mesmas dimensões de X, portanto r x s , com todos elementos nulos, exceto o (m, n) - ésimo elemento que é igual à unidade. Cada uma das r x s matrizes tem as mesmas dimensões, k X ℓ, de Y.

Efetuando o segundo tipo de derivação de Y dado pela (3.4) teremos:

$$\frac{\partial y_{11}}{\partial X} = \begin{bmatrix} a_{11} & 0 \\ a_{12} & 0 \\ a_{13} & 0 \end{bmatrix} \; ; \quad \frac{\partial y_{12}}{\partial X} = \begin{bmatrix} 0 & a_{11} \\ 0 & a_{12} \\ 0 & a_{13} \end{bmatrix} \; ; \text{ etc}$$

o número total de matrizes resultantes da derivação será k x ℓ , lembrando que k e ℓ são dimensões de Y, no exemplo acima 4. Também estas podem ser genericamente representadas por:

$$\frac{\partial y_{pq}}{\partial X} = A^T K_{pq} \tag{3.6}$$

sendo K_{pq} u'a matriz de mesmas dimensões de Y, portanto k x ℓ , com todos elementos nulos, exceto o (p. q) - ésimo elemento que é igual à unidade. Cada uma destas k x ℓ matrizes tem dimensões r x s, neste exemplo 3 x 2, portanto as mesmas dimensões de X.

Usaremos letras maiúsculas para representar matrizes, minúsculas indexadas para escalares e não indexadas para vetores. Y, U, V, ... representarão matrizes cujos elementos são funções de x_{ij} de X e A, B, C, ... são matrizes cujos elementos independem de x_{ij} de X.

Além das matrizes J_{mn} e K_{pq} definidas, necessitaremos, na seqüência dos assuntos, de suas transpostas, J^T_{nm} e K^T_{qp}.

1.2. Derivada de u'a Matriz Constante

Seja:

$$Y = A = \left[a_{ij}\right]$$

para qualquer a_{ij} de A e qualquer x_{mn} de X temos:

$$\frac{\partial y_{ij}}{\partial x_{mn}} = \frac{\partial a_{ij}}{\partial x_{mn}} = 0$$

logo:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}} \left[y_{ij} \right] = 0 \qquad (*) \qquad (3.7)$$

e

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}} \right] y_{pq} = 0 \qquad (3.8)$$

estas duas últimas relações, (3.7) e (3.8) são respectivamente - primeiro e segundo tipo de derivação anteriormente referidos.

1.3. Derivada de u'a Matriz em Relação a si Mesma

Seja:

$$Y = X = \left[x_{ij} \right]$$

para qualquer y_{ij} de Y e qualquer x_{mn} de X teremos:

$$\frac{\partial y_{ij}}{\partial x_{mn}} = \frac{\partial x_{ij}}{\partial x_{mn}} = \begin{cases} 1 \text{ se } i = m \text{ e } j = n \\ 0 \text{ nos demais casos} \end{cases}$$

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}} \left[y_{ij} \right] = \frac{\partial}{\partial x_{mn}} \left[x_{ij} \right] = J_{mn} \qquad (3.9)$$

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}} \right] y_{pq} = \left[\frac{\partial}{\partial x_{ij}} \right] x_{pq} = K_{pq} \qquad (3.10)$$

(*) Representamos por 0 tanto a matriz como o elemento nulo.

onde as igualdades (3.9) e (3.10) dão as derivadas respectivamente do primeiro e segundo tipo.

1.4. Derivada da Matriz Transposta X^T em Relação a Matriz X.

Seja $Y = X^T$, para qualquer y_{ij} de Y temos:

$$Y = \left[y_{ij}\right] = X^T = \left[x_{ji}\right]$$

e para todo y_{ij} de Y e todo x_{mn} de X podemos escrever:

$$\frac{\partial y_{ij}}{\partial x_{mn}} = \frac{\partial x_{ji}}{\partial x_{mn}} = \begin{cases} 1 \text{ se } j = m \text{ e } i = n \\ 0 \text{ nos demais casos} \end{cases}$$

logo:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial X^T}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left[x_{ji}\right] = J^T_{nm} \qquad (3.11)$$

e

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}}\right] y_{pq} = \left[\frac{\partial}{\partial x_{ij}}\right] x_{qp} = K^T_{qp} \qquad (3.12)$$

são as derivadas, primeiro e segundo tipo, da matriz transposta de X.

1.5. Derivada da Soma e da Diferença de Matrizes.

Seja:

$$Y = U + V - W = \left[u_{ij} + v_{ij} - w_{ij}\right]$$

para qualquer y_{ij} de Y e qualquer x_{mn} de X, lembrando o referido no ítem 1.1. anterior de que os elementos de U, V, ... são funções de x, termos:

$$\frac{\partial y_{ij}}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left[u_{ij} + v_{ij} - w_{ij}\right] = \frac{\partial u_{ij}}{\partial x_{mn}} + \frac{\partial v_{ij}}{\partial x_{mn}} - \frac{\partial w_{ij}}{\partial x_{mn}}$$

logo:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left[y_{ij}\right] = \frac{\partial}{\partial x_{mn}}\left[u_{ij} + v_{ij} - w_{ij}\right]$$

com $i = 1, \ldots, k$ e $j = 1, \ldots, \ell$ e qualquer x_{mn} de X, temos portanto:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial U}{\partial x_{mn}} + \frac{\partial V}{\partial x_{mn}} - \frac{\partial W}{\partial x_{mn}} \qquad (3.13)$$

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}}\right](u_{pq} + v_{pq} - w_{pq}) = \left[\frac{\partial}{\partial x_{ij}}\right]u_{pq} + \left[\frac{\partial}{\partial x_{ij}}\right]v_{pq} -$$

$$- \left[\frac{\partial}{\partial x_{ij}}\right]w_{pq}$$

com $i = 1, \ldots, r$ e $j = 1, \ldots, s$ e qualquer y_{pq} de Y. Temos:

$$\frac{\partial y_{pq}}{\partial X} = \frac{\partial u_{pq}}{\partial X} + \frac{\partial v_{pq}}{\partial X} - \frac{\partial w_{pq}}{\partial X} \qquad (3.14)$$

1.6. Derivada do Produto de duas Matrizes.

Sejam U e V de dimensões c x d e d x e respectivamente e Y tal que:

$$Y = UV = \left[y_{ij}\right] = \left[\sum_{s=1}^{d} u_{is}\, v_{sj}\right]$$

com $i = 1, \ldots, c$ e $j = 1, \ldots, e$. Assim a derivada de um escalar y_{ij} em relação a um escalar x_{mn} será dada por:

$$\frac{\partial y_{ij}}{\partial x_{mn}} = \frac{\partial}{\partial x_{mn}}\left(\sum_{s=1}^{d} u_{is}v_{sj}\right) = \sum_{s=1}^{d} \frac{\partial u_{is}}{\partial x_{mn}} v_{sj} + \sum_{s=1}^{d} u_{is} \frac{\partial v_{sj}}{\partial x_{mn}}$$

qualquer que seja y_{ij} de Y e x_{mn} de X. Logo

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial U}{\partial x_{mn}} V + U \frac{\partial V}{\partial x_{mn}} \qquad (3.15)$$

é o resultado para o primeiro tipo de derivação e:

$$\frac{\partial y_{pq}}{\partial X} = \left[\frac{\partial}{\partial x_{ij}}\right] y_{pq} = \sum_{s=1}^{d} \frac{\partial u_{ps}}{\partial X} v_{sq} + \sum_{s=1}^{d} u_{ps} \frac{\partial v_{sq}}{\partial X} \qquad (3.16)$$

o resultado para o segundo tipo. Da (3.15) e (3.16) vemos que a derivada do produto de duas matrizes variáveis é igual à soma dos resultados das derivadas nas quais se considere primeiro uma, depois outra variável como constante.

1.7. Aplicações e Exemplos Particulares.

Consideremos inicialmente o produto entre duas matrizes, uma constante, outra variável. Seja A de dimensões c x d, X, d x e e Y = AX. Da (3.15) temos:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial A}{\partial x_{mn}} X + A \frac{\partial X}{\partial x_{mn}} = 0 + AJ_{mn}$$

$$\frac{\partial Y}{\partial x_{mn}} = AJ_{mn} \qquad (3.17)$$

da (3.16) temos:

$$\frac{\partial y_{pq}}{\partial X} = \sum_{s=1}^{d} \frac{\partial a_{ps}}{\partial X} x_{sq} + \sum_{s=1}^{d} a_{ps} \frac{\partial x_{sq}}{\partial X} = 0 + \sum_{s=1}^{d} a_{ps} \frac{\partial x_{sq}}{\partial X}$$

donde:

$$\frac{\partial y_{pq}}{\partial X} = a_{p1}K_{1q} + a_{p2}K_{2q} + \ldots + a_{pd}K_{dq} = A^T K_{pq}$$

portanto:

$$\frac{\partial y_{pq}}{\partial X} = A^T K_{pq} \tag{3.18}$$

Consideremos ainda a derivada do produto entre duas matrizes uma variável outra constante, na ordem inversa da anterior. Seja $Y = XB$. Da (3.15) temos:

$$\frac{\partial Y}{\partial x_{mn}} = \frac{\partial X}{\partial x_{mn}} B + X \frac{\partial B}{\partial x_{mn}} = J_{mn} B$$

portanto

$$\frac{\partial Y}{\partial x_{mn}} = J_{mn} B \tag{3.19}$$

da (3.16) temos:

$$\frac{\partial y_{pq}}{\partial X} = \sum_{s=1}^{d} \frac{\partial x_{ps}}{\partial X} b_{sq} + \sum_{s=1}^{d} x_{ps} \frac{\partial b_{sq}}{\partial X} = \sum_{s=1}^{d} \frac{\partial x_{ps}}{\partial X} b_{sq} + 0$$

e

$$\frac{\partial y_{pq}}{\partial X} = K_{pq} B^T \tag{3.20}$$

De modo análogo, se $Y = AX^T$ obteríamos, respectivamente para os dois tipos de derivação:

$$\frac{\partial Y}{\partial x_{mn}} = A J_{nm}^T \tag{3.21}$$

e

$$\frac{\partial y_{pq}}{\partial X} = K_{qp}^T A \tag{3.22}$$

e para $Y = X^T B$ teríamos:

$$\frac{\partial Y}{\partial x_{mn}} = J_{nm}^T B \tag{3.23}$$

e

$$\frac{\partial y_{pq}}{\partial X} = B K_{qp}^T \tag{3.24}$$

1.8. Regra Prática para Derivação de Produto Envolvendo as Matrizes X, X^T e A, B, C, ...

São úteis as seguintes regras para obter a derivada do produto de matrizes.

a) a derivada do produto de matrizes terá tantas parcelas quantos fatores variáveis tiver o produto;

b) cada parcela é obtida considerando todos os fatores como constantes, exceto um variável.

Para o primeiro tipo de derivação; em cada parcela:

c) substitui-se a matriz variável X ou X^T por J ou J^T respectivamente.

Para o segundo tipo de derivação; em cada parcela:

c') substitui-se a matriz variável (única) X ou X^T por K ou K^T respectivamente; ao substituirmos X por K tomamos a transposta do produto que precede e do que segue a variável substituída; ao substituirmos X^T por K^T aplicamos uma inversão da ordem do produto entre o conjunto de fatores que precedem a variável e esta, e novamente inversão da ordem do produto deste resultado e o conjunto de fatores que segue a variável.

Usando as regras a) b) e c) podemos obter a derivada do primeiro tipo, enquanto que usando as regras a) b) e c') obtemos a derivada do segundo tipo. Omitimos ao formular estas regras e omitiremos a seguir os índices das matrizes J, J^T, K e K^T que são sempre indicados pelas mesmas letras. Também o elemento pré-fixado x_{mn} de X ou y_{pq} de Y, passarão a ser representados por <X> e <Y> respectivamente.

As deduções referentes ao primeiro tipo de derivação são obtidas de modo bastante simples. Já as referentes ao segundo tipo não são tão facilmente obtidas. Algumas destas deduções podem ser encontradas na ref. |9|.

Podemos ainda obter as derivadas do segundo tipo a partir das de primeiro. Isto é útil uma vez que as regras a), b) e c) referidas são aplicáveis não apenas ao produto referido , mas também ao produto envolvendo a matriz inversa X^{-1} e função - de função. Assim a regra dada abaixo juntamente com aquelas três primeiras tornam-se mais abrangentes.

Dada a derivada de primeiro tipo podemos obter a de' segundo procedendo do seguinte modo:

d) substituimos J e J^T por K e K^T respectivamente;

e) tomamos a transposta do pré e pós-multiplicador de J;

f) o pré e pós-multiplicadores de J^T tornam-se pós e pré-multiplicadores de K^T.

A suscinta exposição dos ítens precedentes nos permite efetuar derivações matriciais usuais em Ajustamento de Observações, Estatística e Regressão.

1.9. Tabela da Derivada Matricial

Com bases nas regras práticas de derivação podemos facilmente obter os resultados constantes da tabela que segue. As fórmulas assinaladas com * são obtidas conforme exposições dos ítens 1.10 a 1.13.

TABELA DE DERIVAÇÃO MATRICIAL DE PRODUTOS E FUNÇÃO USUAIS

	Y	$\frac{\partial Y}{\partial \langle X \rangle}$ (1º TIPO)	$\frac{\partial \langle Y \rangle}{\partial X}$ (2º TIPO)
	AB	0	0
	AX	AJ	A^TK
	XX	$XJ + JX$	$X^TK + KX^T$
	XA	JA	KA^T
	AX^T	AJ^T	K^TA
	X^TA	J^TA	AK^T
	XX^T	$JX^T + XJ^T$	$KX + K^TX$
	X^TX	$J^TX + X^TJ$	$XK^T + XK$
	X^TX^T	$J^TX^T + X^TJ^T$	$X^TK^T + K^TX^T$
	ABC	0	0
	ABX	ABJ	B^TA^TK
	AXX	$AJX + AXJ$	$A^TKX^T + X^TA^TK$
	XXX	$JXX + XJX + XXJ$	$KX^TX^T + X^TKX^T + X^TX^TK$
	AXB	AJB	A^TKB^T
	XAB	JAB	KB^TA^T
	XXA	$JXA + XJA$	$KA^TX^T + X^TKA^T$
	ABX^T	ABJ^T	K^TAB
	AX^TB	AJ^TB	BK^TA
	X^TAB	J^TAB	ABK^T

Y	$\frac{\partial Y}{\partial <X>}$ (1º TIPO)	$\frac{\partial <Y>}{\partial X}$ (2º TIPO)
AXX^T	$AJX^T + AXJ^T$	$A^TKX + K^TAX$
XX^TA	$JX^TA + XJ^TA$	$KA^TX + AK^TX$
X^TXA	$J^TXA + X^TJA$	$XAK^T + XKA^T$
x^TAx		$2Ax$ *
X^{-1}	$- X^{-1}JX^{-1}$	$- (X^{-1})^TK(X^{-1})^T$ *
Z^TZ e $Z = AX$	$J^TA^TZ + Z^TAJ$	$A^TZK^T + A^TZK$ *
X^n	$\sum_{s=0}^{n-1} X^s JX^{n-s-1}$	$\sum_{s=0}^{n-1} (X^T)^s K(X)^{n-s-1}$ *

1.10. Derivada da Forma Quadrática

Se a matriz X for particularizada para um vetor coluna x e A for u'a matriz simétrica, $y = x^TAx$ constitui uma forma quadrática, de grande aplicação em Ajustamento. y é um escalar. Efetuando o segundo tipo de derivação e usando a notação simplificada temos:

$$\frac{\partial <Y>}{\partial X} = \frac{\partial <Y>}{\partial x} = AxK^T + A^TxK$$

como vimos, K tem as mesmas dimensões de Y e K^T as mesmas dimensões de Y^T, logo 1 x 1, ambos iguais à unidade e podemos escrever:

$$\frac{\partial <Y>}{\partial x} = Ax + A^Tx = (A + A^T)\,x$$

como $A = A^T$ temos:

$$\frac{\partial <Y>}{\partial x} = 2Ax \qquad (3.25)$$

1.11. Derivada da Inversa de X.

Se $Y = X^{-1}$, as derivadas do primeiro e segundo tipo de Y serão obtidas através das regras de derivação vistas e da identidade $I = XX^{-1}$. Como I é matriz constante temos:

$$\frac{\partial I}{\partial \langle X \rangle} = 0 = \frac{\partial}{\partial \langle X \rangle}(XX^{-1}) = JX^{-1} + X\frac{\partial X^{-1}}{\partial \langle X \rangle}$$

logo

$$\frac{\partial X^{-1}}{\partial \langle X \rangle} = - X^{-1}JX^{-1} \qquad (3.26)$$

e usando as regras d) e e) do ítem 1.8 temos:

$$\frac{\partial \langle X^{-1} \rangle}{\partial X} = - (X^{-1})^{T}K(X^{-1})^{T} \qquad (3.27)$$

1.12. Derivada da Função de Função

Seja $Y = Z^{T}Z$ e $Z = AX$ temos:

$$\frac{\partial Y}{\partial \langle X \rangle} = \frac{\partial Z^{T}}{\partial \langle X \rangle} Z + Z^{T}\frac{\partial Z}{\partial \langle X \rangle}$$

como $Z^{T} = X^{T}A^{T}$ e

$$\frac{\partial Z^{T}}{\partial \langle X \rangle} = J^{T}A^{T} \qquad \text{e} \qquad \frac{\partial Z}{\partial \langle X \rangle} = AJ$$

podemos escrever:

$$\frac{\partial Y}{\partial \langle X \rangle} = J^{T}A^{T}Z + Z^{T}AJ$$

e aplicando as regras d), e) e f) temos:

$$\frac{\partial \langle Y \rangle}{\partial X} = A^{T}ZK^{T} + A^{T}ZK \qquad (3.28)$$

1.13. Derivada da Potência n, Inteira de X.

Usando as regras de derivação do produto, podemos, facilmente, obter a generalização para n fatores. Convenciona - mos $X^{o} = I$. Genericamente temos:

$$\frac{\partial X^{n}}{\partial \langle X \rangle} = \sum_{s=0}^{n-1} X^{s} \, J \, X^{n-s-1} \qquad (3.29)$$

e

$$\frac{\partial \langle X^{n} \rangle}{\partial X} = \sum_{s=0}^{n-1} (X^{T})^{s} \, K \, (X)^{n-s-1} \qquad (3.30)$$

2. NORMAS VETORIAIS E MATRICIAIS

Ao estudarmos a condição dos sistemas de equações e os erros absolutos ou relativos decorrentes da solução, calculada para estes sistemas, teremos necessidade de representar por um único número a grandeza de um vetor ou de u'a matriz. - Com este objetivo introduzimos a seguir alguns conceitos e propriedades das principais normas vetoriais e matriciais.

2.1. Normas Vetoriais

Em geral a norma de um vetor x é um número não negativo $||x||$, que satisfaz às condições:

$$||x|| > 0 \qquad (3.31)$$

para qualquer $x \neq 0$ e $||0|| = 0$

$$||kx|| = |k| \; ||x|| \qquad (3.32)$$

para qualquer k numérico

$$||x + y|| \leq ||x|| + ||y|| \qquad (3.33)$$

esta última comumente chamada desigualdade triangular. Das condições (3.32) e (3.33) deduz-se facilmente que:

$$||x-y|| \geq ||x'|-||y|| \quad \text{e} \quad ||y-x|| \geq ||x||-||y|| \tag{3.34}$$

Consideremos os seguintes casos particulares:

a) o vetor x degenerado em uma só componente. Um número não negativo adequado para representar sua grandeza seria o seu módulo $|x|$.

b) o vetor x com duas componentes. Neste caso alguns dos estimadores de sua grandeza seriam:

$$||x|| = |x_1| + |x_2| \,;$$

$$||x|| = \left[|x_1|^2 + |x_2|^2\right]^{1/2};$$

$$||x|| = \max_{i=1,2} |x_i| \,;$$

c) o vetor x com três componentes. Sua grandeza poderia ser estimada por:

$$||x|| = |x_1| + |x_2| + |x_3| \,;$$

$$||x|| = \left[|x_1|^2 + |x_2|^2 + |x_3|^2\right]^{1/2};$$

$$||x|| = \left[|x_1|^3 + |x_2|^3 + |x_3|^3\right]^{1/3};$$

$$||x|| = \max_{i=1,2,3} |x_i|$$

Dos exemplos depreende-se que é possível uma definição geral de norma de vetor de dimensão genérica n como:

$$||x||_k = \left[\sum_{i=1}^{n} |x_i|^k\right]^{1/k} \qquad (3.35)$$

e das normas definidas pela (3.35) são de maior interesse para o nosso trabalho as definidas para k = 1,2 e ∞, onde $||x||_\infty$ é interpretada como o máximo $|x_i|$; a $||x||_2$ é o comprimento do vetor x ou norma euclideana e $||x||_1$ é a soma dos módulos das componentes do vetor. Estas normas satisfazem as condições (3.31) , (3.32) e (3.33) e entre estas normas se verificam as seguintes desigualdades:

$$||x||_\infty \leq ||x||_1 \leq n\,||x||_\infty$$

$$||x||_\infty \leq ||x||_2 \leq \sqrt{n}\,||x||_\infty \qquad (3.36)$$

$$\frac{1}{\sqrt{n}}\,||x||_1 \leq ||x||_2 \leq ||x||_1$$

2.2. Normas Matriciais

Define-se norma de uma matriz quadrada A como um número não negativo $||A||$ que satisfaz as condições:

$$||A|| \geq 0$$

para qualquer $A \neq 0$ e $||0|| = 0$ (*)

$$||cA|| = |c|\,||A||$$

(*) Lembramos que 0 representa matriz nula ou elemento zero , conforme o contexto o exija.

para qualquer c numérico

$$||A + B|| \leq ||A|| + ||B|| \tag{3.37}$$

chamada desigualdade triangular e

$$||AB|| \leq ||A||\ ||B||$$

desigualdade de Schwarz.

Uma norma vetorial pode ser associada à correspondente norma matricial. Isto é de grande utilidade uma vez que vetores e matrizes ocorrem relacionados num mesmo problema. Este relacionamento fica assegurado e as normas, vetorial e matricial, são ditas compatíveis se esta última satisfizer a relação:

$$||Ax|| \leq ||A||.||x|| \tag{3.38}$$

e a norma matricial é dita subordinada se:

$$||A|| = \max ||Ax||/||x|| \qquad (x \neq 0) \tag{3.39}$$

As normas matriciais subordinadas às três normas vetoriais acima são:

$$||A||_1 = \max_j \left(\sum_i |a_{ij}|\right) \tag{3.40}$$

chamada norma de coluna;

$$||A||_s = \max_{1<i<n} (\lambda_i) \tag{3.41}$$

chamada norma espectral e

$$||A||_\infty = \max_i \left(\sum_j |a_{ij}|\right) \tag{3.42}$$

que é chamada norma de linha. Na (3.41) λ_i é o autovalor de $A^H A$, sendo A^H a transposta hermitiana.

São ainda de uso comum, as seguintes normas:

$$||A||_E = (\sum_{ij} |a_{ij}|^2)^{1/2} = \left|tr\ (A^H A)\right|^{1/2} \qquad (3.43)$$

onde A^H é a transposta conjugada de A. Esta norma é chamada norma euclidiana;

$$||A||_A = \max |a_{ij}| \qquad (3.44)$$

norma absoluta;

$$M\ (A) = n \max |a_{ij}| \quad (*) \qquad (3.45)$$

a norma euclidiana é freqüentemente representada, para A real por:

$$N\ (A) = \left|\Sigma |a_{ij}|^2\right|^{1/2} = \left|tr\ A^T A\right|^{1/2} \qquad (3.46)$$

Todas as normas acima satisfazem as condições (3.37), com exceção da norma absoluta que não satisfaz a quarta daquelas condições, satisfazendo em seu lugar a condição:

$$||AB||_A \leq n||A||_A\ ||B||_A \qquad (3.47)$$

A norma matricial M(A) é compatível com as normas vetoriais definidas pela (3.35) para k = 1, 2 e ∞, e N(A) é compatível com a norma vetorial euclidiana.

2.3. Relação Entre Normas

Entre as diferentes normas definidas acima se verificam desigualdades, algumas das quais relacionamos abaixo |11|.

(*) Alguns textos omitem o fator n na definição de M(A).

$$\frac{1}{n} M(A) \leq ||A||_{\infty} \leq M(A)$$

$$\frac{1}{n} M(A) \leq ||A||_{1} \leq M(A)$$

$$\frac{1}{n} M(A) \leq N(A) \leq M(A)$$

$$\frac{1}{\sqrt{n}} N(A) \leq ||A||_{\infty} \leq \sqrt{n}\, N(A)$$

$$\frac{1}{\sqrt{n}} N(A) \leq ||A||_{1} \leq \sqrt{n}\, N(A)$$

$$\frac{1}{n} ||A||_{\infty} \leq ||A||_{1} \leq n||A||_{\infty}$$

$$||A||_{S} \leq ||A||_{E} \leq \sqrt{n}||A||_{S}$$

$$||A||_{S} \leq \left[||A||_{1} \cdot ||A||_{\infty}\right]^{1/2}$$

$$\frac{1}{n} M(A) \leq ||A||_{S} \leq M(A)$$

$$\frac{1}{n} N(A) \leq ||A||_{S} \leq N(A)$$

3. ERRO DE ARREDONDAMENTO

A todo pesquisador que manipula valores experimentais é comum o conhecimento fundamental da Teoria do Erro. Assim seria desnecessário determo-nos em conceitos como erros absolutos ou relativos; erros grosseiros, sistemáticos ou acidentais e outros conceitos básicos ou ainda no método dos mínimos quadrados. Parece-nos útil breve consideração sobre os erros de arredondamento. Este erro não constituí geralmente problema sério para trabalhos cujo volume de cálculo é pequeno ou médio. Portan

to, só mais recentemente, tem merecido maior atenção por parte dos pesquisadores, uma vez que só modernamente tem sido possível e freqüente a resolução de problemas com grandes volumes de cálculos.

Conforme frisamos anteriormente, em Geodésia e Fotogrametria o problema de Ajustamento de uma rede de triangulação, de nivelamento, de gravimetria ou ainda de uma faixa ou de um bloco aerotriangulado é enorme. É grande portanto, o volume de cálculo empregado na resolução de tais sistemas e conseqüentemente se justifica que dediquemos alguma atenção aos erros de arredondamento, que particularmente num sistema mal condicionado assumem aspecto muito significativo.

Tendo o mau condicionamento a propriedade de ampliar os erros, estes, ainda que pequenos devem receber especial atenção desde os sistemas de dimensões médias.

3.1. Modos Computacionais

De um modo geral, poderíamos dizer que os problemas cujos erros de arredondamento justificam uma análise mais cuidadosa, seriam problemas cujo volume de cálculo justifica também a utilização de um computador digital em sua solução.

São dois os principais modos computacionais: ponto fixo e ponto flutuante.

Em ponto fixo, o valor calculado x, deve estar sempre no intervalo [-1, 1] (*). Em geral cada número terá um número fixo t, de dígitos binários. Se for necessário trabalhar

---

(*) A rigor o intervalo é aberto ou semi-aberto. Considerá-lo fechado simplifica o raciocínio.

com precisão superior que uma parte em 2^t, pode-se empregar precisão múltipla, i. e., com múltiplos de t dígitos binários.

Em ponto flutuante, cada número x é representado por um par ordenado (a, b) tal que:

$$x = 2^b\ (a) \tag{3.48}$$

onde b é um inteiro positivo ou negativo e a é tal que:

$$-\frac{1}{2} \geqslant a \geqslant -1 \quad \text{ou} \quad \frac{1}{2} \leqslant a \leqslant 1 \tag{3.49}$$

aqui novamente os intervalos, a rigor, não são fechados, mas esta consideração evita minudências desnecessárias numa visão geral do assunto.

3.2. Formas de Análise de Erros

Quando calculamos um valor x a partir de valores a_i dados, cálculo este que consideramos envolver apenas as operações fundamentais, podemos representar este passo, pela equação:

$$x = g\ (a_i) \tag{3.50}$$

com i = 1, ..., n. Como no processamento ocorrem erros de arredondamento, o valor calculado que obtemos é $\bar{x}$ e não x.

A maior parte dos textos estabelece como princípio da análise dos erros, a comparação entre x e $\bar{x}$, forma de análise que Wilkinson [4] denomina FORWARD, na qual tenta-se obter limite para $|\bar{x} - x|$, para cada passo.

A forma de análise que parece mais eficiente em nosso trabalho é aquela na qual o princípio básico consiste em admitir que, num dado passo, o valor calculado $\bar{x}$ é o valor exa

to x para a equação:

$$x \equiv g\ (a_i + \varepsilon_i) \qquad (*) \qquad (3.51)$$

i = 1, ..., n, e então procurar estabelecer limites para ε_i. Esta é a forma de análise, por Wilkinson denominada BACKWARD.

A análise pode ser aplicada a qualquer dos modos computacionais referidos: ponto-fixo ou ponto flutuante e para distingüi-los usaremos:

$$d = fi\ (ab + c)$$

ou (3.52)

$$d = f\ell\ (ab + c)$$

respectivamente.

3.3. Erros de Arredondamento na Computação em Ponto Fixo.

Na adição e subtração, se matematicamente, c = a $\pm$ b, então o valor calculado será c ≡ a $\pm$ b, i. e., não existe erro de arredondamento na computação, nas operações de adição e subtração. É fácil compreendê-lo pelo exposto no ítem 3.1 deste capítulo.

Na multiplicação de fatores de t dígitos situados no intervalo permitido para ponto-fixo, o produto, exatamente representado, terá 2t dígitos. Sua representação em t dígitos implicará numa aproximação de até 5 unidade da ordem t + 1 ou 0,5 unidade no algarismo de ordem t. Portanto o arredondamento é equivalente a até $1/2\ \ 2^{-t}$ (**). A notação posicional, para uma base genérica pode ser representada por:

(*) O sinal de identidade é empregado para indicar x calculado.

(**) Se se tratar de t dígitos no sistema decimal e arredondamento é $1/2\ \ 10^{-t}$.

$$(\ldots C_3 C_2 C_1 C_o, C_{-1} C_{-2} C_{-3} \ldots)_b = \ldots + C_3 b^3 + C_2 b^2 + C_1 b + C_o b^o + \tag{3.53}$$
$$+ C_{-1} b^{-1} + C_{-2} b^{-2} + C_{-3} b^{-3} + \ldots$$

onde b é a base do sistema e C_k um inteiro pertencente ao intervalo $[0, b)$. Entretanto estaremos nos referindo ao sistema binário, b = 2 usual em computação e às vezes exemplificando no sistema decimal, b = 10 para maior clareza.

Se matematicamente tivermos:

$$C = ab$$

o valor computado será:

$$C \equiv ab + \varepsilon \tag{3.54}$$

e

$$|\varepsilon| \leq \frac{1}{2} 2^{-t} \tag{3.55}$$

Por exemplo, no sistema decimal e para t = 4 ; seja: a = 0,1245 e b = 0,3289. Temos ab = 0,04094805 e c ≡0,0409, sendo ε = 0,00004805, portanto

$$|\varepsilon| \leq \frac{1}{2} \cdot 10^{-4}$$

Se o arredondamento for feito simplesmente abandonando as t últimas casas, o limite do erro de arredondamento cresce para: $|\varepsilon| \leq 2^{-t}$ ou $|\varepsilon| \leq 10^{-t}$, respectivamente no sistema binário ou decimal.

O quociente de a por b só estará no intervalo permitido para ponto fixo se |a| < |b| e em geral exige infinitos algarismos para sua representação exata. A representação em t dígitos é arredondada.

Exemplificando no sistema decimal, seja a=0,0119, b = 0,2117 e t = 4. Temos a/b = 0,056211 ... . O valor calculado será 0,0562 e

$$C \equiv \frac{a}{b} + \varepsilon$$

com (3.56)

$$|\varepsilon| = 0{,}00001 \ldots \text{ e } |\varepsilon| \leq \frac{1}{2}\, 10^{-4}$$

Da (3.56) multiplicando ambos os membros por b, e lembrando que b é um número em ponto fixo temos

$$\left| bC - a \right| \leq \frac{1}{2}\, 2^{-t} \qquad (3.57)$$

Se considerarmos o produto escalar, ainda em ponto fixo:

$$S = \sum_{i=1}^{n} a_i b_i \qquad (3.58)$$

com arredondamento de cada produto separadamente teremos S calculado dado por:

$$S \equiv \sum_{i=1}^{n} a_i b_i + \varepsilon$$

e (3.59)

$$|\varepsilon| \leq \frac{I}{2}\, n\, 2^{-t}$$

Muitos computadores e calculadoras desenvolvem as multiplicações e o somatório em 2t dígitos efetuando um só arredondamento final. Neste caso o limite do erro de S calculado na (3.59) cai para:

$$|\varepsilon| \leq \frac{1}{2}\, 2^{-t} \qquad (3.60)$$

Nestes computadores e calculadoras, cujos acumuladores comportam números em 2t dígitos e executam um único arre

dondamento, a expressão matemática:

$$d = \sum_{i=1}^{n} a_i b_i / c$$

computada será:

$$d \equiv (\sum_{i=1}^{n} a_i b_i / c) + \varepsilon$$

com

$$|\varepsilon| \leq \frac{1}{2} 2^{-t}$$

Por exemplo: seja $a_1 = 0,6325$; $a_2 = -0,3127$; $b_1 = 0,4126$; $b_2 = 0,8313$ e $c = 0,0013$. Suponhamos inicialmente que dispomos de computador cujo acumulador preserva 2t dígitos no somatório. Temos então: $d = 0,78614 \ldots$; $d \equiv 0,7861 + \varepsilon$ e $|\varepsilon| = 0,00004 \ldots$. Suponhamos agora um acumulador que execute arredondamento para cada parcela. Teremos: $d = 0,84615 \ldots$; $d \equiv 0,8462 + \varepsilon$ e $|\varepsilon| = 0,06005 \ldots$.

É fácil perceber que a diferença entre os limites dos erros $\varepsilon\varepsilon$, nos diferentes tipos de acumuladores, cresce com o número de produtos ou quocientes existentes na expressão processada.

3.4. Erros de Arredondamento na Computação em Ponto Flutuante.

Sejam x_1 e x_2 números representados em ponto flutuante, conforme visto no ítem 3.1.

$$x_1 = 2^{b_1} a_1 \quad \text{e} \quad x_2 = 2^{b_2} a_2$$

Admitamos ainda que o computador dispõe de acumulador que opere com 2t dígitos.

Na adição as parcelas só serão somadas se a diferença entre os expoentes dos números de maior e de menor módulos for menor ou igual a t. Em caso contrário, i. e., se a diferença referida é maior que t, a menor parcela não alterará nenhum dígito da maior. Seja, por exemplo, $b_1 - b_2 < t$. Então - as mantissas deverão ser compatibilizadas em termos de ordem de grandezas de seus algarismos. Para isto divide-se a mantissa da parcela de menor módulo por $2^{b_1-b_2}$ e soma-se $b_1 - b_2$ ao seu expoente. O número de dígitos necessários para representar exatamente o total será inferior a 2t + 1. Para que este total fique no intervalo permitido para a mantissa, poderá ser necessário multiplicá-lo por potências de 2 e compensar o expoente. A seguir procede-se o arredondamento, para t dígitos.

Os exemplos abaixo em base decimal facilitam a compreensão. Consideremos as adições em ponto flutuante e com t = 5.

a) $x_1 = 10^{-3}(0,43225)$ e $x_2 = 10^{6}(0,45932)$;

$|x_2| > |x_1|$ a diferença entre expoentes $b_2 - b_1 = 9 > t$. Logo

$$x_1 + x_2 \equiv x_2 \equiv 10^6\ (0,45932)$$

b) $x_1 = 10^{2}(0,75326)$ e $x_2 = 10^{5}(0,29478)$;

$|x_2| > |x_1|$ a diferença entre expoentes é $b_2 - b_1 = 3 < t$. Logo

$$\begin{array}{r} 0,2947800000 \cdot 10^5 \\ +\ 0,0007532600 \cdot 10^5 \\ \hline 0,2955332600 \cdot 10^5 \end{array}.$$

esta é a soma exata em 2t dígitos. Arredondando para t dígitos teriamos: $10^5 . (0,29553)$

c) $x_1 = 10^5(0,87528)$ e $x_2 = 10^5(0,84738)$;

$|x_1| > |x_2|$ com $b_1 - b_2 = 0 < t$.

$$\begin{array}{r} 0,8752800000 \cdot 10^5 \\ + \ 0,8473800000 \cdot 10^5 \\ \hline 1,5226600000 \cdot 10^5 \end{array}$$

sendo esta a soma exata em 2t dígitos. Arredondando para t dígitos temos: $10^6(0,15227)$.

d) $x_1 = 10^{-5}(0,57427)$ e $x_2 = 10^{-5}(0,57453$;

$|x_2| > |x_1|$ com $b_2 - b_1 = 0 < t$

$$\begin{array}{r} 0,5745300000 \cdot 10^{-5} \\ - \ 0,5742700000 \cdot 10^{-5} \\ \hline 0,0002600000 \cdot 10^{-5} \end{array}$$

soma exata que é representada em t dígitos por $10^{-8}(0,26000)$.

e) $x_1 = 10^{-5}(0,10783)$ e $x_2 = 10^{-6}(0,99652)$ teremos então:

$$\begin{array}{r} 0,1078300000 \cdot 10^{-5} \\ - \ 0,0996520000 \cdot 10^{-5} \\ \hline 0,0081780000 \cdot 10^{-5} \end{array}$$

resultado em 2t dígitos. Arredondando para t dígitos teremos: $10^{-7}(0,81780)$.

Vemos que os resultados computados nas condições acima serão equivalentes aos resultados exatos, normalizados para atender as exigências (3.49) e arredondados para t dígitos. Representando o total por $2^{b_3}(a_3)$, para os exemplos, base 10, $10^{b_3}(a_3)$, vemos que na mantissa o limite do erro é ainda:

$$\frac{1}{2}\,2^{-t} \quad \text{ou} \quad \frac{1}{2}\,10^{-t}$$

o expoente do total desloca este limite para

$$2^{b_3}.\ \frac{1}{2}\ .\ 2^{-t} \text{ ou } 10^{b_3}\ .\ \frac{1}{2}\ .\ 10^{-t} \qquad (3.61)$$

Obviamente este é o limite do erro absoluto útil em alguns problemas. Entretanto em outros, necessário se faz considerar o erro relativo. Neste caso o limite será:

$$|\varepsilon| \leq 2^{-t} \quad \text{ou} \quad |\varepsilon| \leq \frac{1}{2}\,10^{1-t} \qquad (3.62)$$

e podemos escrever:

$$fl(x_1 + x_2) \equiv (x_1 + x_2)(1 + \varepsilon) \qquad (3.63)$$

Na multiplicação de dois números em ponto flutuante, $x_1 = 2^{b_1}a_1$ e $x_2 = 2^{b_2}a_2$, representando o resultado por $2^{b_3}a_3$ teremos:

$$b_3 = b_1 + b_2 \quad \text{e} \quad \frac{1}{4} \leq |a_3| \leq 1 \qquad (*)$$

este a_3 pode ser normalizado conforme (3.49) através deslocamento à esquerda (**).

Por exemplo: $x_1 = 10^{-3}(0{,}15421)$ e $x_2 = 10^{-2}(0{,}13251)$. Matematicamente temos:

$$x_1\ x_2 = 10^{-5}(0{,}0204343671)$$

(*) No sistema decimal teríamos $\frac{1}{100} \leq |a_3| \leq 1$.

(**) Um deslocamento para a esquerda equivale a multiplicar e dividir respectivamente a mantissa e a potência por 2 ou 10.

cujo valor computado é:

$$fl\ (x_1 .\ x_2) \equiv 10^{-6}(0,20434)$$

ou seja:

$$fl\ (x_1 .\ x_2) \equiv x_1 .\ x_2(1 + \varepsilon) \tag{3.64}$$

onde

$$|\varepsilon| \leq 2^{-t} \quad \text{ou} \quad |\varepsilon| \leq \frac{1}{2}\, 10^{1-t} \tag{3.65}$$

Semelhantemente, para a divisão obtemos:

$$fl\ (x_1/x_2) \equiv (x_1/x_2)(1 + \varepsilon) \tag{3.66}$$

com

$$|\varepsilon| \leq 2^{-t} \quad \text{ou} \quad |\varepsilon| \leq \frac{1}{2}\, 10^{1-t} \tag{3.67}$$

Através dos exemplos de a) até e) retro, podemos observar que quando utilizamos computador de acumulador simples, apenas t dígitos, os resultados podem ser piores. Isto é ilustrado pelo exemplo e); em alguns casos os últimos t dígitos são usados mas não alteram o resultado, exemplo b) e em outros não são usados os últimos t dígitos.

Para este tipo de acumulador o limite do erro na adição cresce para:

$$|\varepsilon| \leq (1\ \tfrac{1}{2}) \cdot 2^{-t} \tag{3.68}$$

e a multiplicação e divisão terão como limite de erros:

$$|\varepsilon| \leq 2^{1-t} \tag{3.69}$$

em

$$fl\ (x_1 .\ x_2) \equiv x_1 .\ x_2\ (1 + \varepsilon)$$

e

$$f\ell\ (x_1/x_2) \equiv x_1/x_2\ (1 + \varepsilon)$$

O acumulador simples eleva o limite de erros , entretanto, qualquer que seja a técnica de arredondamento usada estes limites devem ser inferior a 4.2^{-t}.

3.5. Limites de Erros em Expressões Usuais.

Damos a seguir alguns limites de erros de arredondamento sem demonstrações. Estas poderão ser encontradas em Wilkinson |4|. Sejam as expressões matemáticas:

a) $$p = \prod_{i=1}^{n} x_i$$

temos:

$$p \equiv x_1x_2 \ldots x_n(1+\varepsilon_2)(1+\varepsilon_3) \ldots (1+\varepsilon_n) \quad (3.70)$$

supondo acumulador com 2t dígitos:

$$|\varepsilon_r| \leq 2^{-t}$$

e

$$f\ell\ (x_1x_2 \ldots x_n) \equiv x_1 \ldots x_n\ (1+E) \quad (3.71)$$

sendo

$$(1-2^{-t})^{n-1} \leq 1 + E \leq (1+2^{-t})^{n-1} \quad (3.72)$$

desde que $n-1 << 2^t$, o erro relativo será baixo.

b) $$q = \frac{\left(\prod_{i=1}^{m} x_i\right)}{\left(\prod_{j=1}^{n} y_i\right)}$$

temos:

$$f\ell(x_1x_2 \ldots x_m/y_1y_2 \ldots y_n) \equiv (x_1x_2 \ldots x_m/y_1y_2 \ldots y_n)(1+E)$$

e

$$(1-2^{-t})^{m+n-1} \leq 1+E \leq (1+2^{-t})^{m+n-1} \quad (3.73)$$

c) $$S = \sum_{i=1}^{n} x_i$$

A esta expressão não são extensíveis as leis de limites do erro relativo aplicável à adição de duas parcelas. Neste caso o limite depende da ordem n do somatório

$$S_n = f\ell(x_1 + x_2 + \ldots + x_n) \equiv x_1' + x_2' + \ldots + x_n'$$

sendo

$$x' = x_i\ (1+\eta)$$

e

$$1 + \eta_r = (1+\varepsilon_r)(1+\varepsilon_{r+1}) \ldots (1+\varepsilon_n)$$

com r = 2, ... , n. Os limites serão:

$$(1-2^{-t})^{n+1-r} \leq 1 + \eta_r \leq (1+2^{-t})^{n+1-r} \qquad (3.74)$$

O limite superior do erro é mínimo se a adição for efetuada em ordem crescente dos módulos das parcelas.

d) $$S_n = \sum_{i=1}^{n} a_i b_i$$

O valor computado será dado por:

$$S_n \equiv a_1b_1(1+\varepsilon_1) + a_2b_2(1+\varepsilon_2) + \ldots + a_nb_n(1+\varepsilon_n)$$

e os limites ε_r dados por:

$$(1-2^{-t})^{n-r+2} \leq (1+\varepsilon_r) \leq (1+2^{-t})^{n-r+2} \qquad (3.75)$$

Os limites crescem em grau de complexidade com as respectivas expressões, entretanto seus limites são estabele

cidos de modo análogo.

Os limites vistos são muito rigorosos. De um modo geral, podemos esperar que o erro de arredondamento de cada operação seja randomicamente distribuido no intervalo:

$$(-\frac{1}{2}\, 2^{-t},\ \frac{1}{2}\, 2^{-t})$$

e seria mais razoável considerarmos, por exemplo, para o produtório da expressão a) retro o limite

$$|\varepsilon| \leq n^{1/2}\, 2^{-t} \qquad (3.76)$$

quando n é grande.

Semelhantemente para outras expressões os limites vistos só são aproximados em condições muito particulares e é oportuno adotar valores que melhor representem estatísticamente as grandezas destes limites de erros.

3.6. Limites dos Erros em Expressões Matriciais

Sendo a álgebra matricial uma linguagem concisa e eficiente para a solução de sistemas de equações através da computação eletrônica, frisamos, sem pormenores, limites de erros de algumas operações com matrizes (*).

Consideremos inicialmente a multiplicação de uma matriz A por um escalar k. Matemáticamente temos:

$$B = k\, A$$

(*) O leitor interessado em deduções e análise minudente deve consultar Wilkinson [4] p. 79.

que computado em ponto flutuante proporciona:

$$b_{ij} \equiv k a_{ij} (1+\varepsilon_{ij})$$

com

$$|\varepsilon_{ij}| \leq 2^{-t}$$

$$B - kA \equiv k\, a_{ij}\varepsilon_{ij}$$

para expressar esta diferença com um único valor utilizaremos a norma euclidiana referida neste capítulo. Teremos então

$$||B - kA||_E \leq |k|\ 2^{-t}\ ||A||_E \qquad (3.77)$$

em norma espectral:

$$||B - kA||_S \leq |k|\ 2^{-t}\ n^{1/2}\ ||A||_S \qquad (3.78)$$

ambas são expressões do erro absoluto; a (3.77) pode ser escrita na forma:

$$\frac{||B - kA||_E}{|k|\ \ ||A||_E} \leq 2^{-t} \qquad (3.79)$$

que expressa o erro relativo.

A mesma multiplicação computada em ponto fixo , teria como limite:

$$||B - kA||_E \leq n\ 2^{-t-1} \qquad (3.80)$$

Consideremos agora o produto de u'a matriz quadrada A por um vetor x:

$$y = A\ x$$

em ponto flutuante:

$$y \equiv A\,x + e$$

sendo a norma do vetor e, dada por

$$||e||_2 = |||e||| \leq 2^{-t_1} n\, ||A||_E ||x||_2 \qquad (3.81)$$

e t_1 é aproximadamente igual a t - 0,08406.

Se tivermos o produto entre duas matrizes, por exemplo:

$$C = AB$$

então

$$C \equiv fl(AB) \equiv AB+E$$

$$||E||_E \leq 2^{-t_1} n\, ||A||_E ||B||_E \qquad (3.82)$$

No exemplo acima, se $B = A^T$ e n não é muito grande temos:

$$\frac{||C - AA^T||_E}{||AA^T||_E} \leq 2^{-t} \qquad (3.83)$$

que é a expressão do erro relativo, e, para o erro absoluto temos:

$$||C - AA^T||_E \leq 2^{-t} ||AA^T||_E (1+ \frac{3}{2} 2^{-t_2} n^{3/2}) \qquad (3.84)$$

onde t_2 difere bem pouco de t.

CAPÍTULO IV

MAU CONDICIONAMENTO, IDENTIFICAÇÃO E SOLUÇÃO

1. CONSIDERAÇÕES GERAIS

Parece-nos oportuno lembrar que o problema do mau condicionamento que nos propomos a analisar, é o do mau condicionamento dos sistemas de equações lineares, embora o problema ocorra também na determinação das raízes de um polinômio. Podemos distingüir duas etapas no estudo e tentativa de solução do mesmo. A primeira é a de identificação da ocorrência de mau condicionamento no sistema. A segunda é a solução e análise do resultado. Para ambas as etapas, apesar de grandes autoridades no assunto terem voltado suas atenções ao problema, não dispomos de um procedimento prático único, capaz de identificar o mau condicionamento bem como não dispomos atualmente de um método que solucione de modo satisfatório qualquer sistema mal condicionado.

O que existe são diferentes enfoques na busca de superar ambas as etapas, ora tratadas como um todo, ora separadamente. Passaremos a analisá-los.

2. IDENTIFICAÇÃO DO MAU CONDICIONAMENTO

Ao solucionarmos sistemas de equações lineares , mesmo usando métodos exatos, ocorrem erros devidos: ao arredondamento; ao desaparecimento de dígitos na adição de números aproximadamente simétricos e insuficiência de dígitos em coeficientes oriundos de observações. Estes erros assumem maiores ou menores proporções na solução dependendo do condicionamento do sistema.

2.1. O Determinante Pequeno e a Condição

Consideremos o sistema de n equações lineares simultâneas a n incógnitas:

$$Ax = b \tag{4.1}$$

com a solução exata:

$$x = A^{-1}b \tag{4.2}$$

desde que $|A| \neq 0$.

Surge então uma primeira dificuldade de caráter numérico. Enquanto matemáticamente existe uma dicotomia: $|A| = 0$ ou $|A| \neq 0$, praticamente quando calculamos o determinante de u'a matriz, só eventualmente podemos obter $|A| = 0$. Em geral teremos algum resíduo proveniente das diferentes fontes de erros. Perguntaríamos então: que valores de $|A|$ devem ser considerados despresíveis? É fácil entender que se, por exemplo, a variação dos coeficientes de A, dentro dos limites de precisão com que foram observados, permitir anular o determinante, a solução do sistema não merece nenhuma confiança.

Pela equação (4.2) vemos que as variações na solução poderão ter origem em variações nos elementos α_{ij} da inversa A^{-1} de A ou variações nos elementos b_i do vetor b dos termos

independentes.

A matriz inversa A^{-1} é dita instável se "pequenas" variações nos elementos a_{ij} de A produzem "grandes" variações nos elementos α_{ij} da inversa e estável se "pequenas" variações nos a_{ij} produzem "pequenas" variações nos α_{ij} (*).

A variação de A^{-1} está estreitamente relacionada com a variação do determinante $|A|$. Estudemos então a variação do determinante em função da variação de um elemento qualquer a_{ij} de A. Expressemos o $|A|$ segundo os cofatores A_{ij} de a_{ij}:

$$|A| = \sum_{j=1}^{n} a_{ij} A_{ij} \qquad (4.3)$$

derivando temos:

$$\frac{\partial |A|}{\partial a_{ij}} = A_{ij} \qquad (4.4)$$

A igualdade (4.4) mostra que a variação do determinante devida a uma variação no elemento a_{ij} é igual ao cofator deste elemento. Exemplifiquemos utilizando a matriz de Wilson:

$$W = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \qquad (4.5)$$

cujo determinante $|W| = 1$, e

(*) O conceito de estabilidade é deficiente e a inclusão dos termos "pequeno" e "grande" torna-o quantitativamente vago.

$$W^{-1} = \begin{bmatrix} 68 & -41 & -17 & 10 \\ -41 & 25 & 10 & -6 \\ -17 & 10 & 5 & -3 \\ 10 & -6 & -3 & 2 \end{bmatrix} \qquad (4.6)$$

Entre os cofatores de W o maior é $W_{11} = 68$ o que indica que as variações mais acentuadas no determinante serão produzidas por variações no elemento $w_{11} = 5$. De fato, adicionando um ε ao elemento w_{11} temos:

$$W_{(\varepsilon)} = \begin{bmatrix} 5+\varepsilon & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \qquad (4.7)$$

se $\varepsilon = 0{,}0002$ temos W_1, cujo determinante é 1,0136 e

$$W_1^{-1} = \begin{bmatrix} 67{,}088 & -40{,}450 & -16{,}772 & 9{,}866 \\ -40{,}450 & 24{,}664 & 9{,}862 & -5{,}916 \\ -16{,}772 & 9{,}862 & 4{,}943 & -2{,}966 \\ 9{,}866 & -5{,}916 & -2{,}966 & 1{,}980 \end{bmatrix} \qquad (4.8)$$

se $\varepsilon = -0{,}01$ temos W_2 onde $|W_2| = 0{,}320$ e

$$W_2^{-1} = \begin{bmatrix} 204{,}82 & -128{,}12 & -53{,}12 & 31{,}25 \\ -128{,}12 & 77{,}53 & 31{,}78 & -18{,}81 \\ -53{,}12 & 31{,}78 & 14{,}03 & -8{,}31 \\ 31{,}25 & -18{,}81 & -8{,}31 & 5{,}12 \end{bmatrix} \qquad (4.9)$$

Verificamos dos exemplos, que pequenas variações no elemento w_{11} produzem acentuadas variações no determinante e nos elementos da inversa.

O determinante de W(ε), igualdade (4.7), é:

$$|W(\varepsilon)| = 1 + 68\varepsilon \tag{4.10}$$

e se $\varepsilon = -1/68$, teremos $|W(\varepsilon)| = 0$. Isto nos permite concluir que se os elementos da matriz W fossem obtidos de observações - com precisão conhecida até 0,02 a matriz W deveria ser considerada singular.

2.2. Instabilidade da Inversa

Podemos considerar a variação da inversa A^{-1} em relação a um elemento a_{ij} genérico de A. Usando a (3.26) do capítulo anterior temos:

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1}J_{ij}A^{-1} \tag{4.11}$$

como $J_{ij} = J_{i1}\,J_{1j}$ podemos escrever:

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1}J_{i1}J_{1j}A^{-1} \tag{4.12}$$

e

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -\begin{bmatrix} \alpha_{1i}\alpha_{j1} & \alpha_{1i}\alpha_{j2} \cdots \alpha_{1i}\alpha_{jn} \\ \alpha_{2i}\alpha_{j1} & \alpha_{2i}\alpha_{j2} \cdots \alpha_{2i}\alpha_{jn} \\ \cdots\cdots\cdots\cdots\cdots \\ \alpha_{ni}\alpha_{j1} & \alpha_{ni}\alpha_{j2} \cdots \alpha_{ni}\alpha_{jn} \end{bmatrix} \tag{4.13}$$

donde:

$$\frac{\partial \alpha_{k\ell}}{\partial a_{ij}} = - \alpha_{ki}\alpha_{j\ell} \qquad (4.14)$$

relação que dá a variação sofrida por um elemento fixo $\alpha_{k\ell}$ de A^{-1}, devida às pequenas variações de um único elemento a_{ij} de A. A variação que o elemento $\alpha_{k\ell}$ sofreria por influência de pequenas variações em qualquer a_{ij} de A seria:

$$d\alpha_{k\ell} = -\Sigma\alpha_{ki}\alpha_{j\ell}\, da_{ij} \qquad (4.15)$$

Podemos ainda expressar a variação de um elemento $\alpha_{k\ell}$ da inversa em função dos cofatores:

$$d\alpha_{k\ell} = - \frac{1}{|A|^2} \cdot \Sigma\, A_{ik}\, A_{\ell j}\, da_{ij} \qquad (4.16)$$

As equações de (4.14) a (4.16) dão uma boa estimativa das variações dos elementos da inversa para ε infinitésimo e matrizes estáveis. À medida que ε cresce e, a matriz se torna instável, a qualidade da estimativa se degenera. Isto pode ser observado em exemplos numéricos.

2.3. Sensibilidade da Solução às Variações dos Coeficientes.

Considerando b constante no sistema (4.2) e derivando em relação a a_{ij} de A, temos:

$$\frac{\partial x}{\partial a_{ij}} = \frac{\partial A^{-1}}{\partial a_{ij}} \cdot b$$

como b = Ax e usando novamente a (3.26) do capítulo anterior - vem:

$$\frac{\partial x}{\partial a_{ij}} = - A^{-1} J_{ij}\, x = - \begin{bmatrix} \alpha_{1i}\, x_j \\ \alpha_{2i}\, x_j \\ \vdots \\ \alpha_{ni}\, x_j \end{bmatrix} \qquad (4.17)$$

donde a variação de uma raiz x_k em relação à variação de um elemento a_{ij} de A será:

$$\frac{\partial x_k}{\partial a_{ij}} = - \alpha_{ki} x_j \qquad (4.18)$$

Se considerarmos A constante e b variável, podemos derivar a (4.2) em relação a b_i de b. Teremos então:

$$\frac{\partial x}{\partial b_i} = A^{-1} J = \begin{bmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \vdots \\ \alpha_{ni} \end{bmatrix} \qquad (4.19)$$

Das igualdades (4.18) e (4.19) obtemos:

$$dx_k = - \sum_{i,j=1}^{n} \alpha_{ki} x_j\, da_{ij} \qquad (4.20)$$

e

$$dx_k = \sum_{i=1}^{n} \alpha_{ki}\, db_i \qquad (4.21)$$

que dão as variações em uma das raízes produzidas por variações em a_{ij} ou b_i.

Exemplo: consideremos o sistema: $Wx = b$, onde W é a matriz (4.5) retro e b o vetor:

$$b^T = \begin{bmatrix} 23 & 32 & 33 & 31 \end{bmatrix}$$

a solução exata é:

$$x^T = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

Consideremos agora o sistema

$$Wx' = b' \qquad (4.22)$$

onde $b' = b + \delta b$, tal que:

$$b'^T = \begin{bmatrix} 23{,}01 & 31{,}99 & 32{,}99 & 31{,}01 \end{bmatrix}$$

Aplicando a (4.21) para, por exemplo, $k = 1$ temos $dx_1 = 1{,}36$ e resolvendo o sistema (4.22) teremos $x_1' = 2{,}36 = x_1 + dx_1$.

A matriz cuja inversa é instável, é chamada matríz mal condicionada para a resolução de sistemas de equações lineares. Se num sistema a matriz dos coeficientes é instável o sistema é mal condicionado.

2.4. Restrições

É necessário frisar que embora o determinante dê uma indicação da condição da matriz, ele não constitui condição característica de condicionamento. Num sistema de equações lineares, matrizes que difiram entre si por um fator constante, poderiam ser consideradas igualmente condicionadas, enquanto

que seus determinantes diferem da n-ésima potência desse fator (*). Por outro lado, matrizes com o mesmo determinante podem ser diferentemente condicionadas. Por exemplo:

$$\begin{bmatrix} 50 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0{,}02 \end{bmatrix} \quad e \quad \begin{bmatrix} 50 & 0 & 0 \\ 0 & 0{,}1 & 0 \\ 0 & 0 & 0{,}2 \end{bmatrix}$$

têm, ambas, determinante igual a 1. Suas inversas respectivas são:

$$\begin{bmatrix} 0{,}02 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 50 \end{bmatrix} \quad e \quad \begin{bmatrix} 0{,}02 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

indicando, pelas dimensões dos elementos das inversas, que a segunda matriz é melhor condicionada para inversão. Esta variação ocorre mesmo que as matrizes tenham elementos da mesma ordem de grandeza.

Os sistemas:

$$\begin{aligned} x_1 + x_2 &= 3 \\ x_1 - x_2 &= 1 \end{aligned} \qquad (4.23)$$

e

$$\begin{aligned} 10^{-10}x_1 + 10^{-10}x_2 &= 3.10^{-10} \\ 10^{-10}x_1 - 10^{-10}x_2 &= 10^{-10} \end{aligned} \qquad (4.24)$$

(*) Sendo n a ordem da matriz dos coeficientes

que são equivalentes e bem condicionados, possuem determinantes diferentes para as matrizes dos coeficientes e o segundo sistema possuí determinante muito próximo de zero, -2×10^{-10} e elemen - tos da inversa da ordem de 10^{10}.

Outros dispositivos para investigar a condição de um sistema de equações lineares simultâneas são necessários ante a insuficiência do determinante.

É útil lembrar que o mau condicionamento não constitui por si só um problema para a solução de sistemas. Não fossem os erros das diferentes fontes, tal problema não existiria, razão pela qual para qualquer estudo do condicionamento partimos da premissa de que erros ocorrem, quer sejam nos coeficientes - por serem de origem experimental, quer sejam na mudança de bases, decimal-binária, ou ainda no arredondamento ou perda de dígitos nas subtrações. Esta premissa é indiscutivelmente verdadeira no caso prático.

2.5. Números de Condição para Variações Absolutas ou Relativas.

Em cada trabalho é necessária uma opção entre a consideração do erro absoluto ou erro relativo uma vez que a conveniência de adotar um ou outro tipo varia de um trabalho para outro.

Define-se número de condição para variações absolutas causadas por variações absolutas dos coeficientes, o número C tal que:

$$|\delta x| = C\,|\delta a| \qquad (4.25)$$

onde $|\delta a|$ representa o modulo de uma pequena variação de um parâmetro a, que produz a variação $|\delta x|$ numa quantidade x calculada. Semelhantemente, se para pequenos δa:

$$\left|\frac{\delta x}{x}\right| = K \left|\frac{\delta a}{a}\right| \qquad (4.26)$$

K é chamado número de condição para variações relativas em x causadas por variações relativas em a.

As relações (4.25) e (4.26) podem ser definidas fazendo δa tender a zero e então obtemos:

$$C = \left|\frac{dx}{da}\right| \qquad (4.27)$$

e

$$K = \left|\frac{a}{x}\,\frac{dx}{da}\right| \qquad (4.28)$$

Assim considerando o sistema A'x' = b' correspondente ao (4.1), onde A' = A + δA, b' = b + δb e x' = x + δx, sendo A e A' não singulares temos:

$$(A + \delta A)(x + \delta x) = b + \delta b \qquad (4.29)$$

subtraindo desta a equação (4.1) vem:

$$\delta x = A^{-1} (\delta b - \delta A x') \qquad (4.30)$$

Supondo existirem variações apenas em um elemento b_k do vetor b, a (4.30) daria:

$$\delta x_j = \alpha_{jk}\delta b_k = \frac{A_{kj}}{\Delta}\,\delta b_k \quad (*) \qquad (4.31)$$

que nos permite escrever comparando com a definição (4.25):

(*) Usamos Δ em lugar de $|A|$ para representar o determinante, uma vez que esta notação passou a ser usada para representar módulo. No contexto parece claro.

$$C_{jk} = |\alpha_{jk}| = \frac{|A_{kj}|}{|\Delta|} \qquad (4.32)$$

Para o número de condição de variações relativas - obteríamos:

$$K_{jk} = \left| \frac{\alpha_{jk}\ b_k}{x_j} \right| = \frac{|A_{kj}\ b_k|}{|x_j\ \Delta|} \qquad (4.33)$$

que mostra serem os elementos da inversa números de condição de variações absolutas e estarem estreitamente relacionados ao número de condição de variações relativas. Por exemplo, seja o sistema:

$$\begin{aligned} 3\ x_1 + \quad & 3\ x_2 = 6 \\ 3\ x_1 + 3{,}0003\ & x_2 = 6{,}0003 \end{aligned} \qquad (4.34)$$

cuja solução é $x_1 = x_2 = 1$ e $\Delta = 0{,}0009$; cofator $A_{22} = 3$ e $b_2 = 6{,}0003$ temos então, para este sistema, aplicando a (4.32 e (4.33) respectivamente os valores:

$$C_{22} = \frac{1}{3}10^4 \text{ e } K_{22} \simeq 2.10^4$$

grandes números de condição, comparados à unidade indicando mau condicionamento do sistema. Realmente a solução do sistema é muito sensível a pequenas variações no termo b_2. Assim a solução do sistema:

$$\begin{aligned} 3\ x_1' + \quad & 3\ x_2' = 6 \\ 3\ x_1' + 3{,}0003\ & x_2' = 5{,}997 \end{aligned}$$

será: $x_1' = 12$ e $x_2' = -10$, extremamente sensível a pequenas variações em b_2.

Para o sistema (4.23) retro, cuja solução é $x_1 = 2$ e $x_2 = 1$ e $\Delta = -2$, $A_{22} = 1$ e $b_2 = 1$ temos: $C_{22} = -0,5$ e $K_{22} = -0,5$ que são os mesmos para o sistema (4.24).

Os pequenos números de condição comparados à unidade indicando que o sistema (4.23) equivalente ao (4.24) é bem condicionado, o que realmente se verifica. Observamos que os números de condição conforme definidos acima, indicam melhor a condição, que o determinante.

Os números de condição exemplificados acima constituem uma particularização da (4.30), para maior clareza. Considerações mais gerais serão feitas na seqüência dos assuntos.

2.6. Arbitrariedade da Condição.

Da mesma forma que, num problema específico, devemos decidir qual tipo de erro deve ser analisado, se relativo ou absoluto, poderíamos também fixar P e p respectivamente como números que constituiriam limites para os números de condição C e K. Estabelecendo um P bem maior que a unidade e um p mais ou menos próximo dela, digamos: $P = 10^4$ e $p = 10$. O estudo de tais limites para os principais tipos de problemas de Geodésia e Fotogrametria deve ser feito, para facilitar o trabalho do usuário. Talvez seja possível estabelecer alguns padrões, tomando-se por base a precisão dos coeficientes; o coeficiente de maior cofator; a ordem do sistema; a mínima precisão aceitável para a solução e outros fatores (*).

Supondo que dispomos de limites P e p como referi

(*) Este assunto dá margem a investigação. Não poderíamos tentar desenvolvê-lo num capítulo de nosso trabalho.

dos acima, para um determinado tipo de problema, poderíamos então reformular nosso conceito de mau condicionamento, de modo a tornar-se matemáticamente mais correto.

Um sistema de equações lineares simultâneas é dito mal condicionado se, para variações δa (*) dos parâmetros a da matriz aumentada do sistema, os números de condição C e K definidos pelas relações (4.25) e (4.26) são superiores a P. O sistema será bem condicionado se C e K permanecerem inferiores a p.

Vemos portanto que a decisão de se, um sistema de equações é ou não mal condicionado é específica para um problema ou um tipo de problema, uma vez que os limites P e p devem ser fixados conforme a peculiaridade. Por outro lado não existe um limite nítido entre sistema mal condicionado e bem condicionado, i. e., o limite é um intervalo entre p e P, no qual a solução - provavelmente não proporcionaria a confiança necessária.

Suponhamos agora outro caso particular da (4.30) onde ocorra variação apenas num elemento a_{pq} de Λ. Estudemos então a variação produzida em x_j e chamemos de $C_{j,pq}$ e $K_{j,pq}$ os números de condição correspondentes.

A igualdade (4.17) aplicada para o elemento A_{pq} dá:

$$\frac{\partial x}{\partial a_{pq}} = - \Lambda^{-1} J_{pq} x = - \begin{bmatrix} \alpha_{1p}\, x_q \\ \alpha_{2p}\, x_q \\ \cdots \\ \alpha_{jp}\, x_q \\ \cdots \\ \alpha_{np}\, x_q \end{bmatrix} \qquad (4.35)$$

(*) Usamos δa significando infinitésimo.

Particularizando, a variação no elemento x_j será a j -ésima componente do vetor da (4.35). Teremos então:

$$C_{j,pq} = |\alpha_{jp}\, x_q| = \frac{|A_{pj}\, x_q|}{|\Delta|} \qquad (4.36)$$

e para o número de condição das variações relativas:

$$K_{j,pq} = \left|\frac{a_{pq}}{x_j} \cdot \frac{dx}{da}\right| = \left|\frac{\alpha_{jp}\, a_{pq}\, x_q}{x_j}\right| \qquad (4.37)$$

$$K_{j,pq} = \left|\frac{\alpha_{jp} a_{pq} x_q}{x_j}\right| = \frac{|A_{pj} a_{pj}|}{|\Delta|} \frac{|a_{pq}\, x_q|}{|a_{pj}\, x_j|}$$

Pela (4.36) e (4.37) poderíamos verificar que o sistema dado pela (4.34) é mal condicionado enquanto que o dado pela (4.24) é bem condicionado. Observamos ainda que ambos os números de condição decrescem à medida que o determinante cresce. Um número aproximadamente igual de parcelas aproximadamente simétricas no desenvolvimento do determinante, deverá proporcionar grandes números de condição e conseqüentemente sistemas mal condicionados. Uma das raízes muito pequena, na (4.37), poderá proporcionar um número de condição de variações relativas elevado, sem que possamos considerar o sistema mal condicionado, e sim, que estamos usando inadequadamente a análise de erro relativo.

2.7. Outros Números de Condição

Os números de condição acima não são práticos. O volume de trabalho necessário para identificar a condição é enorme. Outros números de condição, considerados clássicos, serão vistos a seguir:

a) Números de Turing |10|, definidos como:

$$M = \frac{1}{n} \cdot M(A) \cdot M(A^{-1}) \tag{4.38}$$

e

$$N = \frac{1}{n} N(A) \cdot N(A^{-1}) \tag{4.39}$$

sendo M(A) e N(A) as normas definidas pelas igualdades (3.45) e (3.46) do capítulo III. (*)

b) Números de Todd, definidos como:

$$P = \frac{\max |\lambda_i|}{\min |\lambda_i|} \tag{4.40}$$

onde λ_i são autovalores da matriz A.

c) Número de condição H, definido por:

$$H = \sqrt{\frac{\max |\lambda_i|}{\min |\lambda_i|}} \tag{4.41}$$

sendo λ_i autovalores de $A^T A$.

Estes números de condição estão relacionados pelas seguintes inequações; p. 126 ref. |11|.

$$\begin{aligned} N \leqslant M \leqslant n^2 N \\ N \leqslant H \leqslant n \ N \\ P \leqslant H \end{aligned} \tag{4.42}$$

estas relações são verificáveis a partir da definição do número

---

(*) A definição original do número de Turing M é: $M = nM(A) . M(A^{-1})$. Esta diferença tem origem na definição da norma M como $M(A) = \max |A_{ij}|$.

de condição e das relações entre normas dadas na tabela do ítem 2.3 do capítulo III.

Para matrizes ortogonais os números N, H e P são iguais à unidade. Demonstra-se (referência |11|) que todo número clássico de condição é maior ou igual à unidade. O sistema é tanto melhor condicionado quanto menores forem os números de condição. Infelizmente existe limite inferior mas não existe o limite superior. Este terá que ser pesquisado por experimentação possivelmente usando matrizes de testes. E aqui surge novamente uma parcela de arbitrariedade do mau condicionamento peculiar a cada problema ou tipo de problema, a exemplo do que frisamos para os números p e P. Aqui temos também campo para investigação na busca de problemas padrões e respectivos limites.

Se considerarmos o sistema Ax = b com b exato e a_{ij} de A, valores randômicos, independentes, tais que a_{ij} são valores médios com a dispersão $\sigma^2 << a_{ij}$, o número de condição - N, sob o enfoque de probabilidade, tem o seguinte significado :

$$N = \frac{Q_1}{Q_2}$$

onde

$$Q_1 = \frac{m_{vx}}{m_x} \quad e \quad Q_2 = \frac{m_{va}}{m_a} \tag{4.43}$$

sendo m_{vx}, m_x, m_{va} e m_a a média quadrática respectivamente: do erro da incógnita, da incógnita, da variação dos coeficientes e dos coeficientes. (Ver igualdade (4.47) a seguir).

O número de condição H dá a razão entre os semi - eixos maior e menor de um elipsóide de dispersão do vetor cujas

componentes são os erros das incógnitas (*).

2.8. Variação Relativa Total.

Consideremos agora um aspecto geral da condição do sistema (4.30) de equações lineares simultâneas:

$$(A + \delta A)(x + \delta x) = b + \delta b$$

$$||\delta A||_\beta ||A^{-1}||_\beta < 1$$

então

$$\frac{||\delta x||}{||x||} < C||A||.||A^{-1}||\left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||}\right) \tag{4.44}$$

onde

$$C = (1 - ||\delta A||.||A^{-1}||)^{-1} \tag{4.45}$$

Estas relações são demonstradas por Noble |12| p. 433. A (4.44) é verdadeira para normas matriciais e vetoriais quaisquer desde que sejam consistentes. Particularizando, para δA nula teríamos, da (4.45), $C = 1$ e

$$\frac{||\delta x||}{||x||} < ||A||\ ||A^{-1}||\ \frac{||\delta b||}{||b||} \tag{4.46}$$

e para o vetor δb nulo teremos:

(*) Para melhor interpretação do hiperelipsóide de dispersão ver |1| p. 439.

$$\frac{||\delta x||}{||x||} < C\ ||A||\ ||A^{-1}||\frac{||\delta A||}{||A||} \qquad (4.47)$$

Se a perturbação δA em A é pequena, C dado pela (4.45) é aproximadamente igual à unidade, e a condição do sistema será dada por:

$$\bar{M}\ (A) = ||A||\ ||A^{-1}|| \qquad (4.48)$$

igualdade que representa um número de condição genérico que pode ser particularizada, por exemplo, para os números de Turing desde que apliquemos à igualdade (4.48) as normas (3.45) e (3.46).

Um pequeno número de condição $\bar{M}$ implica que o sistema é bem condicionado; a recíproca não é verdadeira, pois $\bar{M}$ é um limite superior das variações relativas em X. $\bar{M}$ grande pode ser considerado como uma indicação de mau condicionamento. Por exemplo:

$$\begin{aligned} x_1 + kx_2 &= 1 \\ x_2 &= -1 \end{aligned} \qquad (4.49)$$

é um sistema bem condicionado para $k > 0$. Aplicando a (4.46) com a norma (3.40) e (3.42) temos:

$$\bar{M} = ||A||\ ||A^{-1}|| = (1 + k)^2$$

que é grande para grandes valores de k, embora a (4.49) seja bem condicionada. Isto dá origem ao estudo da escala.

3. ADEQUAÇÃO DE ESCALA E PIVÔ

A escala de um sistema de equações ou da matriz dos coeficientes freqüentemente assume importância fundamental -

na identificação da condição do sistema.

Se pretendemos resolver o sistema Ax = b pelo método da eliminação de Gauss ou um de seus derivados, a escolha do pivô é importante e exerce, em muitos casos, acentuada influência nos resultados. Esta afirmação é trivial para aqueles que estão "familiarizados" com a Álgebra Linear Aplicada e o Cálculo Numérico e o assunto é sobejamente analisado pelos bons textos de análise numérica; entretanto, é util, neste ponto de nosso trabalho, uma exemplificação para evidenciar a importância da escolha do pivô na solução.

Seja o sistema:

$$\begin{aligned}
0{,}20000\ x_1 + 0{,}16667\ x_2 + 0{,}14286\ x_3 + 0{,}12500\ x_4 &= 1 \\
0{,}16667\ x_1 + 0{,}14286\ x_2 + 0{,}12500\ x_3 + 0{,}11111\ x_4 &= 1 \\
0{,}14286\ x_1 + 0{,}12500\ x_2 + 0{,}11111\ x_3 + 0{,}10000\ x_4 &= 1 \\
0{,}12500\ x_1 + 0{,}11111\ x_2 + 0{,}10000\ x_3 + 0{,}09091\ x_4 &= 1
\end{aligned} \qquad (4.50)$$

Resolvido pela eliminação de Gauss, usando posicionamento parcial por tamanho e com arredondamento de cada valor calculado para cinco decimais temos:

$$\begin{array}{lllll}
0{,}20000 & 0{,}16667 & 0{,}14286 & 0{,}12500 & 1{,}00000 \\
 & 0{,}00694 & 0{,}01071 & 0{,}01278 & 0{,}37500 \\
 & & -0{,}00018 & -0{,}00037 & -0{,}04787 \\
 & & & -0{,}00002 & -0{,}00656
\end{array} \qquad (4.51)$$

O decréscimo progressivo do pivô indica mau condicionamento do sistema. Se os coeficientes da (4.50) são experimentais deve existir uma incerteza de até 5 unidades na ordem $-t + 1$, no exemplo ordem 6. Logo o pivô -0,00002, além dos erros de arredondamento, possui incerteza de cerca de $\pm$ 0,5 unidades

da sua quinta casa decimal. Obviamente a divisão por este pivô amplia enormemente os erros e apesar de efetuarmos os cálculos a cinco decimais, não teremos nenhuma casa na solução com algum grau de confiança.

Através da adequação de escala (*) podemos facilmente mudar de pivô, se o método adotado for o de posicionamento total ou parcial, ou alterar o seu valor, o que implica em alterações dos resultados da solução.

Por adequação de escala de um sistema de equações entende-se: a) multiplicação de cada equação por uma constante não nula e b) substituição de cada incógnita por outra, multipla da anterior.

Por adequação de escala de uma matriz entende-se: multiplicação de suas linhas e colunas por constantes não nulas.

A importância da escala na condição é lembrada desde as definições de condição dadas por vários autores; por exemplo:

a) Se a matriz normalizada (**) A for tal que A^{-1} contenha elementos muito grandes, dizemos que a matriz, e portanto o sistema, é mal condicionado |1|.

b) Se o determinante da matriz normalizada é pequeno comparado com a unidade, o sistema é mal condicionado |3|.

Outros autores que definem a condição baseados na escala, poderiam ser citados. De fato a escala desempenha papel importante na condição da matriz ou do sistema.

(*) Traduzimos como "Adequação de escala" o termo "SCALING".

(**) Diferentes técnicas de normalização serão vistas no ítem 3.4 deste capítulo. Normalização é uma forma de adequação de escala.

3.1. A Escala na Matriz Simétrica

O teorema que segue foi demonstrado por Trencov - |13|: "Os elementos da diagonal das matrizes simétricas definidas positivas estão entre o maior e o menor autovalor da matriz".

$$\lambda_{min} < a_{ii} < \lambda_{máx} \tag{4.52}$$

onde λ são autovalores de A e a_{ij} seus elementos.

Da (4.52) podemos escrever:

$$\lambda_{min} < (a_{ii})_{min} < (a_{ii})_{max} < \lambda_{máx} \tag{4.53}$$

que permite escrever:

$$\frac{(a_{ii})_{max}}{(a_{ii})_{min}} < \frac{\lambda_{max}}{\lambda_{min}} = P \tag{4.54}$$

onde P é o número de Todd, número de condição definido em (4.40). Esta igualdade é importante. Dela depreendemos que altas discrepâncias entre os elementos da diagonal de matrizes simétricas, definidas positivas implicam em grandes números de condição e conseqüentemente em mau condicionamento da matriz e do sistema. Estas discrepâncias podem ser reduzidas pela adequação de escala da matriz.

3.2. Conseqüências nas Equações Normais

Vimos no capítulo II que a matriz A^TA é pior condicionada que a A. A desigualdade (4.54) deduzida para matrizes simétricas definidas positivas, indica que o problema da condição se agrava com grandes diferenças entre a_{ii} máximo e mínimo. Isto é de grande interesse aos problemas da Geodésia e Fotogrametria, pois as equações normais resultantes da aplicação do

método dos mínimos quadrados são simétricas e definidas positivas. A matriz variância-covariância:

$$\Sigma = \begin{bmatrix} \frac{1}{P_1} & \frac{\rho_{12}}{\sqrt{P_1 P_2}} & \cdots & \frac{\rho_{1n}}{\sqrt{P_1 P_n}} \\ \frac{\rho_{12}}{\sqrt{P_1 P_2}} & \frac{1}{P_2} & \cdots & \frac{\rho_{2n}}{\sqrt{P_2 P_n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\rho_{1n}}{\sqrt{P_1 P_n}} & \frac{\rho_{2n}}{\sqrt{P_2 P_n}} & \cdots & \frac{1}{P_n} \end{bmatrix} \qquad (4.55)$$

onde P são os pesos das observações e ρ_{ij} os coeficientes de correlação, é um fator presente em todo ajustamento. Aplicando a (4.53) e (4.54) a esta matriz temos:

$$\frac{P_{max}}{P_{min}} \leqslant P \qquad (4.56)$$

sendo P o número de Todd e $P_{máx}$ e P_{min} os pesos máximos e mínimos.

A (4.56) sugere que na atribuição de pesos, que é vastamente usada para valorizar as observações de melhores qualidades, à medida que as diferenças entre os pesos crescem, a condição do sistema fica prejudicada.

3.3. Dificuldades da Adequação de Escala

As definições da condição de um sistema de equações lineares normalizadas pelo determinante, pelos elementos da inversa ou pelas normas, ainda não constitui uma condição

característica. Deve-se acrescentar que a inclusão da normalização significa um aprimoramento na identificação. Por exemplo, o sistema Ax = b, dado por Noble |12| evidencia este aprimoramento:

$$\begin{aligned} 2x_1 + x_2 + x_3 &= 1 \\ x_1 + \varepsilon x_2 + \varepsilon x_3 &= 2\varepsilon \\ x_1 + \varepsilon x_2 + \varepsilon x_3 &= \varepsilon \end{aligned} \qquad (4.57)$$

tem $|A| = 2\varepsilon(1 - 2\varepsilon)$ e

$$A^{-1} = \begin{bmatrix} -\frac{\varepsilon}{1-2\varepsilon} & \frac{1}{1-2\varepsilon} & 0 \\ \frac{1}{1-2\varepsilon} & -\frac{1+2\varepsilon}{2\varepsilon(1-2\varepsilon)} & \frac{1}{2\varepsilon} \\ 0 & \frac{1}{2\varepsilon} & -\frac{1}{2\varepsilon} \end{bmatrix}$$

Para ε muito pequeno comparado com a unidade temos $|A|$ pequeno e alguns α_{ij} de A^{-1} muito grandes comparados também à unidade. Entretanto o sistema (4.57) cujos elementos da matriz aumentada tinham ordem de grandeza unitária, pode, dentro da mesma ordem de grandeza, mas com escala mais adequada, ser escrito como: $A'x' = b'$ matricialmente, onde

$$A' = PAQ \qquad x' = Q^{-1}x \qquad b' = Pb \qquad (4.58)$$

sendo P e Q matrizes diagonais com $p_{11} = 1$; $p_{22} = 1/\varepsilon$; $p_{33} = 1/\varepsilon$; $q_{11} = \varepsilon$ e $q_{22} = q_{33} = 1$. O novo sistema será:

$$
\begin{aligned}
2\varepsilon x_1' + x_2' + x_3 &= 1 \\
x_1' + x_2' + x_3 &= 2 \\
x_1' + x_2' - x_3 &= 1
\end{aligned}
\tag{4.59}
$$

com $|A'| = 2(1 - 2\varepsilon)$ e

$$
(A')^{-1} = \begin{bmatrix}
-1 & 1 & 0 \\
1 & -\frac{1}{2}(1+2\varepsilon) & \frac{1}{2}(1-2\varepsilon) \\
0 & \frac{1}{2}(1-2\varepsilon) & -\frac{1}{2}(1-2\varepsilon)
\end{bmatrix}
$$

Agora o determinante não é pequeno, bem como os elementos da inversa não são grandes comparados à unidade, quando ε é pequeno. Isto indica que o sistema é bem condicionado e mostra ainda que a preocupação de introduzir escala no conceito de condição é um aprimoramento.

Entretanto, a rigor, não basta introduzir escala no conceito de número de condição. Os exemplos vistos mostram que dentro de uma mesma ordem de grandeza existem escalas mais recomendáveis que outras. O número de condição $\bar{M}(A)$ dado pela igualdade (4.48) varia para diferentes escalas A' de A, onde $A' = PAQ$. Sendo este número de condição um limite superior para variações relativas das incógnitas, a melhor escala é aquela que minimize o $\bar{M}(A')$. Assim o problema se resume em determinar P e Q tais que $\bar{M}(A')$ seja mínimo:

$$
\bar{M}(A') = \min ||PAQ||\ ||Q^{-1}A^{-1}P^{-1}|| \tag{4.60}
$$

Este problema não está completamente resolvido. O seguinte teorema foi demonstrado por Noble |12|:

Seja $\left[|A|\right]=\left[|a_{ij}|\right]$, $\left[|A^{-1}|\right]=\left[|\alpha_{ij}|\right]$, $\left[|A|\right].\left[|A^{-1}|\right] > 0$ e

$$\left||A'^{-1}|\right|.\left||A|\right| > 0$$

então o número $\bar{M}(A)$ de condição tal que:

$$\bar{M}(A) = \min ||PAQ||_{\infty}||Q^{-1}A^{-1}P^{-1}||_{\infty} \qquad (4.61)$$

é igual ao máximo autovalor μ de $\left||A|\right|\left||A^{-1}|\right|$ ou $\left||A^{-1}|\right|\left||A|\right|$. Os elementos p_i e q_i das diagonais de P e Q que satisfazem a (4.61) são dados por:

$$\bar{p} = \left[\frac{1}{p_i}\right] \text{ e } \bar{q} = \left[q_i\right] \qquad (4.62)$$

sendo $\bar{p}$ e $\bar{q}$ autovetores correspondentes a μ de $\left[|A|\right]\left[|A^{-1}|\right]$ e $\left[|A^{-1}|\right]\left[|A|\right]$.

A escala assim definida torna constante a soma dos valores absolutos das linhas de PAQ. Demonstra-se ainda que se s e S são a menor e a maior entre as somas de linhas de $\left[|A|\right]\left[|A^{-1}|\right]$ então $s \leqslant \bar{M}(A) \leqslant S$.

É fácil notar que o valor prático do teorema acima é mínimo, uma vez que o volume de cálculo cresce enormemente e em conseqüência perde em eficiência sob o aspecto tempo de computação; erros de arredondamento acumulados e dificuldades para o usuário.

3.4. Procedimentos Práticos para Adequação de Escala.

Embora não possamos, de um modo geral, propiciar uma escala ótima, devemos nos preocupar com este importante fator que pode alterar sensivelmente a solução e a identificação do condicionamento.

Os seguintes procedimentos podem ser utilizados e são recomendáveis para adequação de escala, embora não sejam ótimos, i. e., não satisfaçam a condição (4.61):

a) Dividimos cada elemento da coluna pelo compri - mento da mesma, inclusive coluna b. Sejam

$$t_j = \frac{1}{\left(\sum_{i=1} a_{ij}^2\right)^{1/2}} \quad e \quad t_{n+1} = \frac{1}{\left(\sum_{i=1} b_i^2\right)^{1/2}} \tag{4.63}$$

e seja D a matriz diagonal cujos elementos não nulos são: t_1 , t_2, ..., t_n. O sistema que obteríamos seria:

$$ADy = t_{n+1}\, b \tag{4.64}$$

com

$$y = (t_{n+1}\, D^{-1})x \qquad e \qquad x = \frac{D}{t_{n+1}}\, y \tag{4.65}$$

então teremos qualquer a_{ij} de A normalizada pelo procedimento a cima menor ou igual a unidade.

b) Aplicando o procedimento acima para a matriz simétrica A*'A, qualquer elemento da matriz será, em módulo , menor ou igual a unidade e os elementos do vetor A*'b normaliza- do serão também menores que a unidade.

c) Se a matriz A é simétrica e definida positiva (*) podemos efetuar adequação de escala multiplicando linha e coluna i, por:

$$\frac{1}{|a_{ii}|^{1/2}}$$

(*) As equações normais obtidas pela aplicação do m.m.q. são simétricas e definidas positivas |13|.

que resulta todos os elementos da diagonal unitários. O vetor b é normalizado ainda dividindo cada elemento pelo comprimento do vetor.

d) Estabelecendo A' diagonalmente equivalente a A. A matriz A' é dita diagonalmente equivalente à matriz A se existem matrizes diagonais D_1^{-1} e D_2 não singulares tais que:

$$A' = D_1^{-1} A D_2 \qquad (4.66)$$

Os elementos de D_1 e D_2, em geral, não são escolhidos com rigor. É recomendável que os elementos destas matrizes sejam definidos como potências inteiras da base β do sistema computacional, afim de que não se introduzam mais operações e conseqüentemente perda de tempo de computação e principalmente de precisão, pelo arredondamento. A adequação de escala utilizando potências inteiras de β só altera a característica b do número arquivado em ponto flutuante, mantendo inalterada sua mantissa a (ver relação (3.48) Cap. III). Se $A' = AD_2$ diz-se que A' é equivalente em escala por coluna; se $A' = D_1^{-1} A$ é equivalente em escala por linha. Assim para resolver o sistema Ax = b, com esta normalização resolveríamos o sistema:

$$D_1^{-1} A D_2 x' = b'$$

onde

$$D_2 x' = x \qquad \text{e} \qquad D_1 b' = b$$

e) Tornando-se a matriz A equilibrada. A matriz A é equilibrada por linha, em relação à norma infinita, se:

$$\beta^{-1} \leq \max_{1 \leq j \leq n} |a_{ij}| \leq 1$$

e equilibrada por coluna em relação à mesma norma, se:

$$\beta^{-1} \leqslant \max_{1 \leqslant i \leqslant n} |a_{ij}| \leqslant 1$$

onde β é como acima. A é dita equilibrada se for equilibrada por linha e por coluna.

Exemplificando, sejam $\beta = 10$ e a matriz A, dada por:

$$A = \begin{bmatrix} 1 & 1 & 2.10^{9} \\ 2 & -1 & 10^{9} \\ 1 & 2 & 0 \end{bmatrix}$$

a matriz B, obtida por adequação de escala de A, conforme o ítem e) será: $B = D_1^{-1} A'$, onde $d_{11} = d_{22} = 10$ e $d_{33} = 10^{10}$.

$$B = \begin{bmatrix} 0,1 & 0,1 & 0,2 \\ 0,2 & -0,1 & 0,1 \\ 0,1 & 0,2 & 0 \end{bmatrix}$$

e B é portanto equilibrada. Semelhantemente a matriz C tal que $C = AD_2$, onde $d_{11} = d_{22} = 10^{-10}$ e $d_{33} = 10^{-1}$,

$$C = \begin{bmatrix} 10^{-10} & 10^{-10} & 0,2 \\ 2,10^{-10} & -10^{-10} & 0,1 \\ 0,1 & 0,2 & 0 \end{bmatrix}$$

é equilibrada. Embora ambas as matrizes sejam equilibradas, possuem diferentes condições, diferentes escalas e teriam diferentes pivôs na solução de um sistema pelo método da eliminação. A matriz equilibrada B, acima, constitui uma escala mais adequada

para A.

4. RESOLUÇÃO DOS SISTEMAS

4.1. Considerações Gerais

A eficiência dos métodos de solução numérica dos sistemas de equações lineares se caracteriza, para grandes sistemas, por diversos aspectos: precisão da solução; tempo de processamento; espaço de memória exigido e aplicabilidade. Dependendo do problema específico, certos aspectos se tornan relevantes. Assim conforme as dimensões de um sistema e a capacidade de memória do computador a terceira característica deve receber maior ou menor atenção. Conforme o volume de cálculos , o tempo de computação é ou não fundamental. Assim a precisão do método pode prescindir ou não de cuidados especiais. Ao trabalharmos com coeficientes exatos e sistemas bem condicionados as precauções não são as mesmas que necessitamos quando temos coeficientes experimentais, com limitadas casas decimais conhecidas e principalmente se a condição do sistema não é boa. Neste caso empreendem-se esforços no sentido de obter a melhor solução que a precisão dos coeficientes, com suas limitações físicas, o permita. Como vimos no artigo anterior, o problema da determinação da condição de um sistema é complexo, além de possuir certo grau de arbitrariedade para cada caso em particular ou grupo que possa constituir um padrão. Em conseqüência desses fatores, e uma vez que a maior parte dos problemas de ajustamento tem coeficientes aproximados e oriundos de observações, sem ignorar os demais aspectos da eficiência de um método de solução de sistema, passaremos a dar tratamento especial ao fator precisão.

É enorme a literatura que trata da resolução numérica de sistemas de equações, como também é enorme o número

de métodos de resolução quer sejam diretos, iterativos ou combinados. Não é nosso objetivo, neste trabalho, analisar métodos de solução de sistemas. Isto é cuidadosamente feito por Faddeev - |11|, Ralston |1|, Forsythe |14| e muitos outros autores. Analisaremos apenas aspectos recomendáveis para reduzir o mau condicionamento, minimizar os erros de arredondamento e perda de dígitos com algumas técnicas para solucionar certos tipos de sistemas mal condicionados e estimativa da qualidade da solução.

4.2. Escala e Pivô na Solução de Sistemas

Há certa unanimidade entre os autores ao recomendarem a adequação de escala dos sistemas antes de qualquer tentativa de solução do mesmo. Não existe entretanto indicação de um critério prático ideal, apenas se reconhece a utilidade da escala. O desconhecimento de uma adequação de escala ideal agrava , em parte, o problema da escolha do pivô nos métodos de eliminação que selecionam pivôs. Em conseqüência do problema da escala não está ainda estabelecido um procedimento único e ótimo para escolha do pivô. Noble |12| fez estudo analítico dessa escolha para o sistema de duas equações a duas incógnitas. O critério - por ele estabelecido para tais sistemas é excessivamente trabalhoso, apesar das mínimas dimensões do sistema (2x2).

Existe nítida concordância no sentido de que o posicionamento parcial ou total é recomendável quando se aplica a eliminação, principalmente nos sistemas mal condicionados.

4.3. Condição Prévia

A condição prévia é um dispositivo através do qual se procura substituir o sistema mal condicionado por outro, matemáticamente equivalente, mas bem condicionado. Por exemplo , o sistema:

$$\begin{aligned} x + \quad y &= 1 \\ x + 1{,}01y &= 2 \end{aligned} \qquad (4.67)$$

cuja solução é x = -99 e y = 100 é sensível, pois:

$$\begin{aligned} x' + \ y' \quad &= 1 \\ x' + 0{,}99y &= 2 \end{aligned}$$

tem para solução x = 101 e y = -100.

Podemos através de operações elementares com as linhas da matriz aumentada obter:

$$\begin{aligned} 3x + 4y &= \ 103 \\ 4x - 3y &= -696 \end{aligned} \qquad (4.68)$$

que é equivalente ao sistema (4.67) e que é bem condicionado. Esta substituição do sistema (4.67) pelo (4.68) constitui o dispositivo denominado"condição prévia".

É importante evidenciar que a condição prévia jamais elimina, nem mesmo reduz, os erros da solução provocados por incertezas dos coeficientes. Se os coeficientes são obtidos por observações, seus valores são incertos num intervalo de dispersão e estas incertezas são inerentes ao sistema. Com respeito a estas incertezas a solução do sistema, por meio da condição prévia, não pode ser melhorada, só seria aprimorada se as observações o fossem através de instrumentos mais precisos e observações mais cuidadosas.

A condição prévia é útil para reduzir os efeitos dos erros que se originam no decorrer do processamento. Parece mais indicada para resolução de sistemas onde os coeficientes - sejam conhecidos exatamente, ou com alta precisão, ou ainda

quando as dimensões do sistema sejam grandes, de modo que, os erros de arredondamento prejudiquem a qualidade da solução. Não existe ainda um algoritmo que permita obter o sistema bem condicionado equivalente a um dado sistema mal condicionado (*). Este é mais um ponto que dá margem à investigação. É oportuno lembrar que transformações que envolvam grande volume de cálculos perdem em termos de eficiência sob todos os aspectos. Portanto seriam de grande utilidade e eficiência, algoritmos que propiciassem o sistema bem condicionado equivalente ao sistema inicial operando o mínimo possível com as mantissas a, dos elementos arquivados - na memória em ponto - flutuante e efetuando as transformações , tanto quanto possível, apenas com as características b (ver - (3.48) cap. III) de tais elementos.

4.4. Matriz Inversa à Esquerda ou à Direita.

Mendelson |15| demonstra o seguinte teorema: Dados ε e K arbitrários, reais, positivos, para toda ordem n existem A e A^{-1} não singulares tais que: se $C = A^{-1}A$ então $|C_{ij} - i_{ij}| < \varepsilon$ e se $D = AA^{-1}$ então $|d_{ij}| > K$, sendo i_{ij} elemento da matriz identidade.

Este teorema mostra a importância de se distinguir matrizes inversas aproximadas à direita e à esquerda. Esta distinção é evidentemente mais importante nos sistemas mal condicionados, onde a inversa inadequada tem discrepâncias acentuadas, como mostra o exemplo Ax = b abaixo, mal condicionado (**).

(*) Pelo menos não constatamos na literatura consultada qualquer referência a tal algoritmo.

(**) Exemplo de Mendelson |15|.

$$\begin{aligned} x_1 + \quad x_2 + \quad x_3 &= 6{,}00 \\ x_1 + \quad x_2 + 1{,}01x_3 &= 6{,}03 \\ x_1 + 0{,}99x_2 + \quad x_3 &= 5{,}98 \end{aligned} \tag{4.69}$$

cuja solução exata é $x_1 = 1$, $x_2 = 2$ e $x_3 = 3$. A inversa à esquerda aproximada e arredondada para duas decimais será:

$$A_E^{-1} = \begin{bmatrix} 1{,}01 & -100 & 100 \\ 99{,}99 & 2 & -102 \\ -102{,}01 & 102 & 0 \end{bmatrix}$$

e o produto $A_E^{-1} A$ dará:

$$A_E^{-1}A = \begin{bmatrix} 1{,}01 & 0{,}01 & 0{,}01 \\ -0{,}01 & 1{,}01 & 0{,}01 \\ -0{,}01 & -0{,}01 & 1{,}01 \end{bmatrix}$$

enquanto que o produto AA_E^{-1} dará:

$$AA_E^{-1} = \begin{bmatrix} -2{,}01 & 4 & -2 \\ -2{,}03 & 5{,}02 & -2 \\ -2{,}01 & 3{,}98 & -0{,}98 \end{bmatrix}$$

e a solução obtida através de A_E^{-1}, $x = A_E^{-1} b$ é $x_1 = 1{,}06$; $x_2 = 2{,}04$ e $x_3 = 3{,}00$ que é uma boa aproximação. Enquanto que a inversa aproximada à direita de A é:

$$A_D^{-1} = \begin{bmatrix} 1{,}01 & -101{,}99 & 100{,}01 \\ 102 & 2 & -100 \\ 102 & -100 & 0 \end{bmatrix}$$

que pós-multiplicada a A dá:

$$AA_D^{-1} = \begin{bmatrix} 1,01 & 0,01 & 0,01 \\ -0,01 & 1,01 & 0,01 \\ -0,01 & -0,01 & 1,01 \end{bmatrix}$$

e pré-multiplicada a A produz:

$$A_D^{-1}A = \begin{bmatrix} -0,97 & -1,97 & -1,99 \\ 4 & 5 & 4,02 \\ -2 & -2 & -1 \end{bmatrix}$$

A solução obtida para o sistema, usando A_D^{-1} pré-multiplicada , $\dot{x} = A_D^{-1}b$, seria $x_1 = -10,88$; $x_2 = 26,06$ e $x_3 = -9$, que é uma péssima solução. Isto nos adverte quanto a necessidade de conhecer que tipo de inversa nos fornece determinado método de inversão matricial aproximada. No caso de desconhecimento do tipo da inversa seria recomendável a resolução e uma verificação com pré e pós - multiplicação. Para matrizes simétricas esta distinção não é importante.

4.5. Teoria da Perturbação.

Como os indicadores de condição dão apenas o limite superior das variações da solução, e uma vez que as incertezas dos coeficientes não podem ser eliminadas matemáticamente ou por processos computacionais, surgem as seguintes perguntas : quais os efeitos destes erros na solução? Qual a mínima precisão dos coeficientes para se obter solução aceitável? Esta foi uma diferente maneira de enfocar o problema, proposta por Wilkinson [4]. A técnica se baseia na computação exata de um sistema per - turbado.

Suponhamos que no sistema $Ax = b$, a matriz A e o vetor b sofram perturbações δA e δb e que estas acarretam variações δx na solução x. Wilkinson p. 91 |4| estabelece limites para estas variações e Morgan |16| introduz algumas modificações aos mesmos, tornando-os mais práticos, embora o volume de cálculos seja ainda enorme. Para perturbações em A e b temos a variação absoluta:

$$||\delta x|| \leq \frac{||A^{-1}||\ ||\delta A||\ ||x|| + ||A^{-1}||\ ||\delta b||}{1 - ||A^{-1}||\ ||\delta A||} \qquad (4.70)$$

Para perturbações em A temos a variação absoluta:

$$||\delta x|| \leq ||\ [(I + C)^{-1} - I]\ ||\ ||x|| \qquad (4.71)$$

onde $C = A^{-1}\ \delta A$. Para variações relativas:

$$\frac{||\delta x||}{||x||} \leq \frac{||A^{-1}||\ ||\delta A||}{1 - ||A^{-1}||\ ||\delta A||} \qquad (4.72)$$

esta última pode ser escrita como

$$\frac{||\delta x||}{||x||} \leq \frac{||A^{-1}||\ ||A||\ \dfrac{||\delta A||}{||A||}}{1 - ||A^{-1}||\ ||A||\ \dfrac{||\delta A||}{||A||}} \qquad (4.73)$$

onde a expressão $||A^{-1}||\ ||A||$ constitui, conforme a norma considerada um dos números de condição já definidos. Se representarmos $||A^{-1}||\ ||A||$ por $\overline{M}(A)$ conforme a (4.48) e a razão $||\delta A||/||A||$ por δ, a relação (4.73) pode ser escrita como:

$$\frac{||\delta x||}{||x||} < \frac{\overline{M}(A)\delta}{1 - \overline{M}(A)\delta} \qquad (4.74)$$

onde δ é geralmente um número muito pequeno comparado com a unidade, entretanto se $\overline{M}(A)$ é grande, $\overline{M}(A).\delta$ pode tender a um e a variação relativa da solução tender para infinito. As relações acima bem como a (4.44), (4.46) e (4.47) dão, em termos de norma, a variação abscluta ou relativa do vetor solução, devida às perturbações nos coeficientes do sistema. Da (4.74) vemos que são fundamentais no estabelecimento do limite de variações da solução, tanto o número de condição como a perturbação δ, dos coeficientes, estando esta estreitamente ligada ao número t de bits da palavra de memória do computador. Portanto, a perturbação introduzida durante o armazenamento de um coeficiente é $1:2^b$. Esta perturbação se torna muito menor em precisão estendida. Freqüentemente, a norma mais usada na estimativa da perturbação é a norma infinita $||\delta A||_\infty$ que é proporcional à ordem da matriz. Assim:

$$||\delta A||_\infty = n\ 10^{t}$$

A perturbação introduzida pelo processamento é geralmente pequena comparada com as perturbações de origens experimentais.

4.6. Refinamento da Matriz Inversa e da Solução.

Este método consiste em, dada uma inversa aproximada A_o^{-1} da matriz A, obter inversas A_1^{-1}, A_2^{-1} etc, tais que

$$|I - AA_n^{-1}| < |I - AA_{n-1}^{-1}| \qquad (4.75)$$

O refinamento é limitado pelo nível de ruído do

computador, e para que o mesmo ocorra, se fizermos:

$$C_o = I - AA_o^{-1} \tag{4.76}$$

devemos ter [17]:

$$||C_o|| < 1 \tag{4.77}$$

É usual utilizar norma 1 ou ∞ por facilidade computacional. Faddeev [11] p. 159 demonstra a seguinte fórmula iterativa para refinamento da inversa:

$$A_m^{-1} = A_{m-1}^{-1} (I + C_{m-1}) \tag{4.78}$$

onde C_{m-1} tem o significado do (4.76) para a inversa A_{m-1}^{-1}. Usando as fórmulas anteriores demonstra-se, [16], a seguinte fórmula para o refinamento:

$$A_m^{-1} = 2A_{m-1}^{-1} - A_{m-1}^{-1} AA_{m-1}^{-1} \tag{4.79}$$

Representando por D_m a discrepância entre a inversa exata e a inversa calculada:

$$- D_m = A^{-1} - A_m^{-1} \tag{4.80}$$

temos, em termos de norma, o seguinte limite

$$||D_m|| \leq ||A_{m-1}^{-1}|| . ||(I-C_{m-1})^{-1}|| . ||C_o^{2^m}|| \tag{4.81}$$

Morgan [16], usou a (4.79) no refinamento matricial. Efetuou testes com matrizes de grandes dimensões e com segmentos finitos da matriz de Hilbert, que são muito mal condicionados, e concluiu que o refinamento é possível até o nível de

ruído do computador para matrizes cujo número de condição $P<10^5$. Wilkinson (ref |4|, p. 93) estabelece limite superior para D_m em função da inversa A^{-1} e do número t de dígitos da precisão utilizada. Da (4.80)

$$A_m^{-1} = A^{-1} + D_m \tag{4.82}$$

onde

$$|d_{ij}| \leq 2^{-t} |a_{ij}|$$

$$||D_m||_E \leq 2^{-t} ||A^{-1}||_E$$

ou (4.83)

$$||D_m||_S \leq 2^{-t} n^{1/2} ||A^{-1}||_S$$

Da igualdade (4.82) temos:

$$AA_m^{-1} = A (A^{-1} + D_m) = I + AD_m \tag{4.84}$$

Como

$$||AD_m||_S \leq ||A||_S ||D_m||_S$$

e substituindo nesta última o valor dado pela (4.83) temos:

$$||AD_m||_S \leq ||A||_S . 2^{-t} . n^{1/2} . ||A^{-1}||_S \leq 2^{-t} n^{1/2} k(A) \tag{4.85}$$

onde k(A) é o número de condição de A, definido como em (4.48), para a norma espectral e n a ordem da matriz A. Vemos que se:

$$k(A) \geq 2^t n^{-1/2}$$

a norma de AD_m pode ser grande e conseqüentemente AA_m^{-1} não será aproximadamente igual à unidade e A_m^{-1} não é uma boa aproximação da inversa de A.

Além do refinamento da inversa podemos ainda , num processo iterativo, obter refinamento para a solução. Do sistema Ax = b, se $\hat{x}$ é o vetor solução calculada, tal que $x = \hat{x} + \delta x$, temos:

$$A(\hat{x} + \delta x) = b \tag{4.86}$$

e

$$A\delta x = b - A\hat{x} = r \tag{4.87}$$

sendo r o resíduo da solução calculada.

A equação (4.87) pode ser resolvida aproximadamente em δx dando um $\delta\hat{x}$ que nos permite obter um novo valor de $\hat{x}$. Iterativamente teríamos:

$$\hat{x}_{i+1} = \hat{x}_i + \delta\hat{x}_i \tag{4.88}$$

onde

$$\delta\hat{x}_i = A^{-1} r_i \tag{4.89}$$

Os valores dos resíduos r_i devem ser calculados em precisão estendida. A solução $\hat{x}_1$, inicial do processo iterativo é a solução $\hat{x}$ do sistema Ax = b obtida em precisão simples. É importante notar que este refinamento da solução não acarreta um grande acréscimo de cálculo, pois a inversa A^{-1} é calculada somente a primeira vez. O acréscimo de cálculo é de n^2 multiplicações e mesmo número de adições, número pequeno comparado ao total de operações necessárias para a resolução do sistema ou para inversão matricial. Segundo Noble [12], se o processo converge, é linear a convergência e acresce um número fixo de dígitos por iteração até o nível de ruído.

Se o sistema é muito mal condicionado a solução inicial $\hat{x}_1$ obtida em precisão simples pode não ser satisfatória para a convergência. Neste caso é conveniente analisar se mesmo

a solução em precisão estendida não seria duvidosa.

Freqüentemente somos encorajados empíricamente a excessivo esmero na solução precisa, usando precisão estendida, refinamento e outros cuidados, quando o sistema é experimental e cuja solução, ainda que calculada exatamente, dependendo da condição do sistema, pode ser desprovida de qualquer grau de confiança. Segundo Noble [12], a melhor maneira de solucionar o mau condicionamento é evitá-lo. Isto nem sempre é possível, mas deve-se procurar obter um modelo matemático que evite o mau condicionamento do problema. Exemplifica sua sugestão com um segmento finito da matriz de Hilbert.

4.7. Uma Melhor Condição para $A^TAx = A^Tb$

Sabemos que o vetor x que minimiza a soma dos quadrados dos resíduos do sistema inconsistente Ax = b onde A é m x n, m > n, é:

$$x = (A^TA)^{-1} A^Tb$$

Vimos no cap. II que o sistema:

$$A^TAx = A^Tb \qquad (4.90)$$

é pior condicionado que o original Ax = b, A normalizada.

Pretendemos obter um sistema melhor condicionado que o sistema (4.90) e que dê a solução pelo método dos mínimos quadrados.

Consideremos então o sistema:

$$Ax = b$$

com A m x n e posto de A igual a n. Particionemos A e b na forma:

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \qquad e \qquad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{4.91}$$

tais que A_1 tenha dimensões n x n e b_1 n x 1. Sendo post (A) = n, existe uma ordenação das equações do sistema, tal que det $(A_1) \neq 0$ e na prática devemos obter $\det(A_1)$ tão grande quanto possível. Como A tem posto n podemos obter A_2 como combinação linear de A_1:

$$A_2 = PA_1 \tag{4.92}$$

e

$$A = \begin{bmatrix} I \\ P \end{bmatrix} A_1 \tag{4.93}$$

onde I é de ordem n e P de dimensões (m - n) x n. Substituindo os valores das igualdades (4.91) a (4.93) em (4.90) temos:

$$A_1^T \begin{bmatrix} I & P^T \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} A_1 x = A_1^T \begin{bmatrix} I & P^T \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{4.94}$$

ou

$$A_1^T \left[I + P^T P \right] A_1 x = A_1^T \left[b_1 + P^T b_2 \right] \tag{4.95}$$

e como $\text{post}(A_1) = n$, temos:

$$\left[I + P^T P \right] A_1 x = b_1 + P^T b_2 \tag{4.96}$$

Analisemos a condição deste último sistema através do determinante da matriz dos coeficientes:

$$\det \left([I + P^T P] A_1 \right) = \det [I + P^T P] . \det A_1 \tag{4.97}$$

Como

$$A^TA = A_1^T \left[I + P^TP\right] A_1 \qquad (4.98)$$

podemos escrever, usando a (2.10)

$$\det\left[I+P^TP\right] = \frac{\det(A^TA)}{(\det A_1)^2} = \frac{\sum_p (\det A_p)^2}{(\det A_1)^2} \qquad (4.99)$$

substituindo o resultado desta última na (4.97) temos:

$$\det(\left[I+P^TP\right] A_1) = \frac{\det (A^TA)}{\det A_1} \qquad (4.100)$$

ou

$$\det(\left[I+P^TP\right] A_1) = \frac{\sum_p (\det A_p)^2}{\det A_1} \qquad (4.101)$$

Lembrando que A é normalizada e que A_1 é uma das submatrizes A_p, a de maior determinante, da (4.99) é fácil concluir que $\det|I + P^TP| > 1$ e com este resultado, através da (4.100) concluímos que a condição do sistema (4.99) é pior que a condição do (4.96). A condição deste último será tanto melhor quanto maior for o determinante de A_1, razão pela qual devemos escolher $\det(A_1)$ tão grande quanto possível.

Sendo as m - n últimas equações combinações lineares das n primeiras, podemos escrever:

$$b_2 = P\, b_1 \qquad (4.102)$$

A (4.96) pode ser escrita como:

$$\left[I + P^TP\right] A_1x = \left[I+P^TP\right]b_1 + P^T\left[b_2 - Pb_1\right]$$

ou

$$A_1 x = b_1 + \left[I + P^T P\right]^{-1} P^T \left[b_2 - P b_1\right] \qquad (4.103)$$

Em se tratando de problemas com coeficientes experimentais não temos combinações lineares como indicadas em (4.102) e (4.32) que implicariam, pela (4.103), em equivalência entre $A_1 x = b_1$ e o sistema (4.96). Temos as $(m - n)$ equações como aproximadas combinações lineares das n primeiras e então a segunda parcela do segundo membro da (4.103) é não nula e pode ser considerada como correções ao vetor b_1 uma vez que $b_2 \simeq P b_1$

Uma das maneiras práticas de procedimento, usando as submatrizes referidas acima, seria considerar a matriz:

$$\begin{bmatrix} A_1 & b_1 & I \\ A_2 & b_2 & 0 \end{bmatrix} \qquad (4.104)$$

e através da eliminação Gauss-Jordam, por exemplo, fazer $A_1 = I$ e $A_2 = 0$ obtendo:

$$\begin{bmatrix} I & A_1^{-1} b_1 & A_1^{-1} \\ 0 & b_2 - P b_1 & -P \end{bmatrix} \qquad (4.105)$$

sendo P obtido pela (4.32). Esta matriz fornece os elementos necessários para a resolução da (4.96) ou da (4.103). Para obtermos $\det(A_1)$ grande, podemos usar a seleção parcial de pivô.

Este procedimento, como vimos, teoricamente conduz a sistema melhor condicionado para determinação de uma solução pelo método dos mínimos quadrados. Deve entretanto ser testado praticamente, para se constatar se o ganho em termos de condição é significativo para sistemas de diferentes dimensões,

apesar do volume de cálculos e dos conseqüentes erros de arredondamento introduzidos na computação. Experimentos recomendariam - ou não sua aplicação prática.

4.8. Remoção do Mau Condicionamento.

A remoção do mau condicionamento, para solução de sistemas de equações lineares, é um método, devido a Eisemann - [7], que se aplica aos sistemas mal condicionados onde poucas equações são aproximadamente combinações lineares de outras. O método não acarreta grande acréscimo de cálculos e aprimora sensivelmente a solução, entretanto não pode ser usado para resolver sistemas onde a quase totalidade das equações são aproximadamente combinações lineares das demais, como ocorre com os segmentos finitos da matriz de Hilbert.

A fundamentação matemática do método está baseada na eliminação de Gauss. O procedimento será analisado considerando a matriz A, quadrada e de ordem cinco. Usando o método da eliminação, após a realização do segundo estágio teremos a transformação $R^{(2)}$ de A tal que:

$$R^{(2)} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} \\ 0 & 1 & r_{23} & r_{24} & r_{25} \\ 0 & 0 & a_{33,2} & a_{34,2} & a_{35,2} \\ 0 & 0 & a_{43,2} & a_{44,2} & a_{45,2} \\ 0 & 0 & a_{53,2} & a_{54,2} & a_{55,2} \end{bmatrix} \qquad (4.106)$$

onde os elementos r e a de R são dados para um estágio qualquer k, por:

$$r_{kj} = \frac{a_{kj,\,k-1}}{a_{kk,\,k-1}} \qquad (4.107)$$

$$a_{ij,k} = a_{ij,\,k-1} - a_{ik,\,k-1}\, r_{kj}$$

onde $k = 1, \ldots, n-1$; j e $i = k+1, \ldots, n$

A matriz $R^{(k)}$, obtida da matriz A, de ordem n , após k estágios de tranformação, pode ser representada parti - cionadamente como:

$$R^{(k)} = \begin{bmatrix} R_{11} & R_{12} \\ 0 & A^{(k)} \end{bmatrix} \qquad (4.108)$$

onde R_{11} é a matriz triangular superior de ordem k; R_{12} é matriz de dimensões k x (n-k); 0 é a matriz nula de mesmas dimensões de R_{12}^T e $A^{(k)}$ é matriz quadrada de ordem (n-k) dos elementos $a_{ij,\,k}$.

Numa formulação matricial podemos obter $R^{(1)}$, $R^{(2)}$, ... de A pelas transformações L_1, L_2, ... Assim:

$$\begin{aligned} R^{(1)} &= L_1 A \\ R^{(2)} &= L_2 R^{(1)} = L_2 L_1 A \\ R^{(k)} &= L_k L_{k-1} \cdots L_1 A \end{aligned} \qquad (4.109)$$

onde:

$$L_1 = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \cdot & \cdot & \cdot & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & \cdot & \cdot & & \cdot \\ & & & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & & \cdot & \cdot & 0 \\ -\frac{a_{n1}}{a_{11}} & 0 & \cdot & \cdot & 0 & 1 \end{bmatrix}$$

e (4.110)

$$T_k = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & & & & \\ \vdots & & \ddots & \vdots & & \\ & & 0 & 1 & & \\ & & \vdots & -\frac{a_{k+1\,k,k-1}}{a_{kk,k-1}} & & \\ & & & \vdots & & \\ 0 & & \cdots & -\frac{a_{nk,k-1}}{a_{kk,k-1}} & \cdots & 1 \end{bmatrix}$$

$$\begin{aligned} A &= T_1 R^{(1)} \\ R^{(1)} &= T_2 R^{(2)} \end{aligned} \qquad (4.111)$$

$$A = R^{(0)} = T_1 \cdot T_2 \cdot \ldots \cdot T_k \cdot R^{(k)} = P^{(k)} \cdot R^{(k)}$$

onde $P^{(k)} = T_1 T_2 \ldots T_k$ e $T_k = L_k^{-1}$. Esta inversa é facilmente obtida a partir da L_k pela simples troca de sinais dos elementos que a distinguem da matriz identidade. Os elementos da k-ésima coluna da matriz T_k, que a tornam distinta da matriz identidade, são iguais aos elementos $a_{i\,k-1,\,k-1}$, i. e., são iguais aos elementos da primeira coluna da submatriz $A^{(k)}$ da partição (4.108).

A matriz $P^{(k)}$ é constituída das k colunas da matriz T_i. Por exemplo, para a ordem cinco temos:

$$P^{(2)} = T_1 . T_2 = \begin{bmatrix} p_{11} & 0 & 0 & 0 & 0 \\ p_{21} & 1 & 0 & 0 & 0 \\ p_{31} & 0 & 1 & 0 & 0 \\ p_{41} & 0 & 0 & 1 & 0 \\ p_{51} & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & p_{22} & 0 & 0 & 0 \\ 0 & p_{32} & 1 & 0 & 0 \\ 0 & p_{42} & 0 & 1 & 0 \\ 0 & p_{52} & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} p_{11} & 0 & 0 & 0 & 0 \\ p_{21} & p_{22} & 0 & 0 & 0 \\ p_{31} & p_{32} & 1 & 0 & 0 \\ p_{41} & p_{42} & 0 & 1 & 0 \\ p_{51} & p_{52} & 0 & 0 & 1 \end{bmatrix} \tag{4.112}$$

Podemos representar a matriz P, triangular inferior, de ordem n, para um estágio k, particionada como:

$$P^{(k)} = T_1 \, T_2 \, \ldots \, T_k = \begin{bmatrix} P_{11} & 0 \\ P_{21} & I \end{bmatrix} \tag{4.113}$$

onde P_{11} é triangular inferior de order k, P_{21} tem dimensões - (n-k) x k; 0 tem as dimensões de P_{21}^T e I é a matriz identidade de ordem (n-k). Os elementos p_{ij} de P são obtidos diretamente - da eliminação, sem envolver cálculo adicional. A matriz original A pode ser representada por:

$$A = P^{(k)} \, R^{(k)} \tag{4.114}$$

Esta igualdade é uma generalização da decomposição triangular. Podemos daqui por diante subentender o estágio genérico k e abandonar a notação correspondente. A submatriz $A^{(k)}$ representaremos por S, para evitar confusão.

Se num dado estágio todos os elementos de uma linha i de S, representada por S^i, são muito pequenos, isto revela mau condicionamento da matriz A e que a linha i é aproxima

damente combinação linear de outras. Consequëntemente na conclusão deste passo da eliminação ou num subsequënte se verificarão perdas de dígitos significativos pela divisão, por número muito pequeno dos elementos da linha S^i. Isto equivale a multiplicar os erros de arredondamento, até então ocorridos, por um número muito grande. Ocorre então uma súbita queda de precisão que se propaga a toda a solução. Torna-se necessário, portanto, reduzir os erros de arredondamento na obtenção de S^i para preservar a precisão. A linha S^i é equivalente à diferença entre uma combinação linear das k primeiras linhas da matriz original A e a i-ésima linha da mesma matriz A. Obtendo os coeficientes desta combinação linear podemos obter S^i a partir de A.

Como vimos da (4.114), subentendido o estágio k, A = PR e portanto:

$$R = P^{-1}A = L\,A$$

a linha S^i de S, submatriz de R, para i > k, pode ser expressa como:

$$S^i \sim R^i = L^iA = (L^i_{21},\ I^i)A = \ell_{i1}A^1 + \ldots + \ell_{ik}A^k + A^i \tag{4.115}$$

Sendo todos os elementos de S^i pequenos, podemos escrever:

$$A^i = -\ell_{i1}\,A^1 - \ell_{i2}\,A^2 - \ldots - \ell_{ik}\,A^k \tag{4.116}$$

onde os índices superiores indicam linha.

Devemos então obter os elementos ℓ_{ij} da linha L^i para i > k. Considerando as matrizes particionadas $L^{(k)}$ e $P^{(k)}$ e a identidade L P = I, temos:

$$\begin{vmatrix} L_{11} & 0 \\ L_{21} & I \end{vmatrix} \cdot \begin{vmatrix} P_{11} & 0 \\ P_{21} & I \end{vmatrix} = \begin{vmatrix} I & 0 \\ 0 & I \end{vmatrix} \tag{4.117}$$

donde:

$$L_{21}\, P_{11} + P_{21} = 0$$

e

$$L^{i}_{21}\, P_{11} + P^{i}_{21} = 0$$

o que nos permite escrever:

$$\begin{aligned} \ell_{i1}\, P_{11} + \ell_{i2}\, P_{21} + \ldots + \ell_{ik}\, P_{k1} &= -\, P_{i1} \\ \ell_{i2}\, P_{22} + \ldots + \ell_{ik}\, P_{k2} &= -\, P_{i2} \\ - - - - - - - - - - - - - - - & \\ \ell_{ik}\, P_{kk} &= -\, P_{ik} \end{aligned} \qquad (4.118)$$

Os coeficientes p_{ij} deste sistema já são conhecidos dos passos anteriores, apesar de estarem eivados de erros de arredondamento dos referidos passos. Obtemos então, por retro-substituição, valores ℓ'_{ij} aproximados em lugar de ℓ_{ij}, mas estes valores aproximados satisfazem as exigências do problema. Da (4.115) vemos que a partir dos elementos ℓ_{ij}, j = 1, ... , k, e A^j e A^i podemos obter S^i. Usando ℓ'_{ij} obteremos A'^i. Novamente aqui teríamos perdas de dígitos significativos, entretanto agora procuramos desenvolver os cálculos em "double precision", dando A'^i isenta de erros de arredondamento. A linha A'^i tem ainda pequenos elementos, mas estes podem ser multiplicados por um número M, grande para obtenção de pivô unitário, sem introduzir ou ampliar erros de arredondamento neste estágio, portanto a solução conserva um nível de precisão compatível com o dos passos anteriores.

Eliminando os k primeiros elementos da linha A'^i teremos a linha R^i.

O volume de cálculo adicional é pequeno compara-

do ao processamento total do problema, uma vez que as demais linhas de R permanecem inalteradas.

Nas linguagens de programação modernas é possível declarar as variáveis em módulos independentes. Assim o fato de desenvolvermos pequena parte do processamento em precisão estendida, dupla ou múltipla precisão, não constitui dificuldade.

Uma vez constatado o mau condicionamento da matriz, podemos resumir o método exposto em quatro etapas:

a) Solução do sistema (4.118) por retro-substituição para obter os ℓ_{ij}.

b) Cálculo de A'^{i} em dupla precisão.

c) Multiplicação de A'^{i} por um número M, grande, "left shift".

d) Eliminação dos k primeiros elementos da linha i.

O exemplo que segue serve para ilustrar o método e é transcrito da referência |7|. Por simplicidade não leva em conta o posicionamento por tamanho. Dado o sistema:

$$1{,}32x_1 - 4{,}73x_2 + 5{,}39x_3 - 2{,}84x_4 + 3{,}97x_5 = 1{,}33$$

$$5{,}68x_1 - 6{,}25x_2 + 1{,}40x_3 + 7{,}02x_4 + 4{,}50x_5 = -6{,}04$$

$$1{,}93x_1 + 1{,}34x_2 - 2{,}16x_3 + 3{,}81x_4 - 2{,}62x_5 = -1{,}75$$

$$2{,}85x_1 + 3{,}09x_2 + 4{,}41x_3 + 2{,}36x_4 + 3{,}14x_5 = 2{,}34$$

$$4{,}32x_1 + 0{,}20x_2 + 4{,}69x_3 - 1{,}63x_4 - 5{,}39x_5 = 4{,}50$$

Estabeleçamos que o maior elemento de uma linha qualquer não deve ser menor que 0,20. Isto ocorrendo a precisão está aquém da aceitável.

Aplicando a eliminação com arredondamento a duas decimais e desenvolvendo três estágios, k = 3, obtemos : $R^{(3)}x = b$ onde $A = P^{(3)} R^{(3)}$. O valor numérico da matriz aumentada será:

$$(R^{(3)}, b^{(3)}) = \begin{bmatrix} 1 & -3,58 & 4,08 & -2,15 & 3,01 & 1,01 \\ & 1 & -1,55 & 1,37 & -0,89 & -0,84 \\ & & 1 & -1,21 & -0,39 & 1,17 \\ & & & 6,47 & 11,61 & -5,03 \\ & & & -0,08 & -0,01 & 0,02 \end{bmatrix}$$

e

$$P^{(3)} = \begin{bmatrix} 1,32 & & & & \\ 5,68 & 14,08 & & & \\ 1,93 & 8,25 & 2,76 & & \\ 2,85 & 13,29 & 13,38 & 1 & \\ 4,32 & 15,67 & 11,35 & 0 & 1 \end{bmatrix}$$

Na eliminação, após o terceiro estágio, os elementos da quinta linha, i = 5, se tornaram menores que o mínimo aceitável, indicando que a quinta linha é aproximadamente combinação linear das três primeiras. Formamos então o sistema $P_{11}^T \ell^i + P^i = 0$, que é:

$$\begin{aligned} 1,32\ell_{51} + 5,68\ell_{52} + 1,93\ell_{53} &= -4,32 \\ 14,08\ell_{52} + 8,25\ell_{53} &= -15,67 \\ 2,76\ell_{53} &= -11,35 \end{aligned}$$

cuja solução arredondada para duas decimais dá: $\ell'_{51} = -2,86$; $\ell'_{52} = 1,30$ e $\ell'_{53} = -4,11$. Com estes valores, através da fórmula:

$$A'^5 = \ell'_{51} A^1 + \ell'_{52} A^2 + \ell'_{53} A^3 + A^5$$

calculada em dupla precisão, obtemos uma nova linha para A^5 que será:

$$-0,0035x_1 + 0,0954x_2 - 0,0278x_3 - 0,0407x_4 - 0,1260x_5 = 0,0367$$

que com um "left shift" e com a eliminação das três primeiras incógnitas dá para o final do terceiro estágio a seguinte quinta linha para a matriz aumentada $(R^{(3)}, b^{(3)})$:

$$(0;\ 0;\ 0;\ -2,26;\ 0,31;\ -2,47)$$

onde o valor numérico do maior elemento está acima do limite estabelecido, dentro da precisão aceitável.

A solução correta do sistema é o vetor x, tal que $x^T = \left[-2\ 0\ 2\ 1\ -1\right]$. Usando método comum temos para solução $x'^T = \left[-0,30\ 0,39\ 0,77\ -0,23\ -0,31\right]$ e usando o método acima a solução será: $x''^T = \left[-1,96\ 0,00\ 1,95\ 0,96\ -0,97\right]$ que indica excelente ganho em precisão.

CAPÍTULO V

CONCLUSÃO

Como frisamos em páginas anteriores, não existe uma solução para sistemas mal condicionados, no sentido geral. O que se pode fazer é identificá-lo; estimar a precisão da solução com base na condição e na dispersão dos coeficientes e adotar alguns procedimentos que dêem a solução mais precisa possível. Para isso deve-se:

a) Efetuar adequação de escala do sistema, antes de qualquer iniciativa de solução, usando um dos procedimentos indicados no ítem 3.4. do capítulo IV.

b) Identificar a condição do sistema em escala utilizando, se possível, elementos que o próprio método de solução forneça, tais como: determinante da matriz normalizada; elementos da matriz inversa; máximo elemento das (m - k) linhas de cada estágio k de um dos métodos de eliminação; ou um dos números de condição citados no capítulo IV. Como os indicadores de condição não são condições características é recomendável, se possível, usar mais que um indicador de condição.

c) Estimar a qualidade da solução não pelos resídu

os de sua substituição no sistema e sim através de uma das fórmulas (4.70) a (4.73) ou (4.74) do ítem 4.5 do capítulo IV, que dão a variação absoluta ou relativa da solução. Deve-se precaver com a escolha adequada do erro a ser analisado, se absoluto ou relativo.

d) Adotar, se o sistema for mal condicionado, os seguintes procedimentos: Tentar obter outro modelo matemático - bem condicionado (*); se for sistema de equações com coeficientes exatos usar a técnica da condição prévia exposta no ítem 4.3 do último capítulo; sendo os coeficientes oriundos de experimentação, usar o refinamento da matriz inversa e da solução se o método adotado usa inversão matricial, e, nos métodos de eliminação utilizar a remoção do mau condicionamento descrita no ítem 4.8 do capítulo IV.

Os procedimentos acima são recomendáveis, entretanto só a experimentação permitirá conclusões bem precisas acerca dos mesmos.

Vimos que os sistemas de equações normais são piores condicionados, (cap. II) e que a atribuição de pesos excessivamente discrepantes entre si prejudica a condição do sistema de equações normais, devendo portanto ser evitada (ítem 3.1 capítulo IV).

Vimos também no ítem 4.7 do último capítulo que teoricamente é possível obter u'a matriz melhor condicionada para as equações normais, devendo isto ser verificado experimentalmente.

(*) Pode-se objetar que isto não é solução e sim fuga ao problema, mas, na realidade é o melhor.

REFERÊNCIAS BIBLIOGRÁFICAS

|01| RALSTON, Anthony. The solution of simultaneous linear equations. In: --- A first course in numerical analysis. Tokyo, McGraw-Hill Koga Kusha, c 1965 p. 394-450. (International series in pure and applied Mathematics).

|02| MOULTON, F.R. On the solutions of linear equations having small determinants. Amer.Math. Monthly, (20): 242-249, 1973.

|03| WESTLAKE, Joan R. A handbook of numerical matrix inversion and solution of linear equations. New York, John Wiley, 1968. 167 p.

|04| WILKINSON, J.H. Rounding errors in algebraic processes. New Jersey, Prentice Hall, 1964. 160 p.

|05| GEMAEL, Camil. Aplicação do cálculo matricial em geodésia 2ª parte: Ajustamento de observações. Curitiba, Curso de Pós-Graduação em Ciências Geodésicas, Universidade Federal do Paraná, 1974. 85 p.

|06| TAUSSKY, O. Note on the condition of matrices, MTAC (4): 111-112, 1950.

|07| EISEMANN, Kurt. Removal of ill conditioning for matrices. New York, IBM Corporation, 225-231, 1956.

08| DE LUCA, Nelson. Curso de cálculo matricial. Curitiba, U.F.P. Depto. de Física. 1966. 96 p.

09| DWYER, Paul S. & MACPHAIL, M.S. Symbolic matrix derivatives. Cópias xerox remetidas pelo Dr. Uotila. 18 p. s. d.

10| TURING, A. M. Rounding off errors in matrix processes. Quart. I. Mech. Appl. Math., 1: 287-308, 1948.

11| FADDEEV. D.K. & FADDEEVA, V.N. Computational methods of linear algebra. Trad. Robert C. Willians. San Francisco, W. H. Freeman, 1960. 621 p.

12| NOBLE, Ben. Applied linear algebra. New Jersey, Prentice - Hall Inc, 1969. 523 p.

13| TRENCOV, I. Une propriete du nombre de Todd des matrices normales obtenues en compensation par la methode des moindres carres. Bulletin Géodésique, Paris, (107): 85-88, mar. 1973.

14| FORSYTHE, George E. & MULER, Cleve B. Computer solution of linear algebric systems. New Jersey, Prentice Hall, 1967. 148 p.

15| MENDELSOHN, N. S. Some elementary properties of ill conditioned matrices and linear equations. 18p.s.d. |fotocópia|

16| MORGAN, Peter James. An investigation into some problems

of linear orbiter photography system. Columbus, Ohio State University, 1971. 48 p. (Reports of the Department of Geodetic Science n. 162).

[17] FOX, L. An introdution to numerical linear algebra. Oxford, Clarendon Press, p. 136-157, 1964.