MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO DE PESQUISAS ESPACIAIS

AUTORIZAÇÃO PARA PUBLICAÇÃO
AUTHORIZATION FOR PUBLICATION

PALAVRAS CHAVES/KEY WORDS
AUTORES/AUTHORS: MÉTODOS AUTOCORRETIVOS, EQUAÇÕES NÃO-LINEARES, EQUAÇÕES LINEARES, EQUAÇÕES DIFERENCIAIS ORDINÁRIAS, EQUAÇÕES DE DERIVADAS PARCIAIS, EQUAÇÕES INTEGRAIS.

AUTORIZADA POR/AUTHORIZED BY
Volker W.J.H. Kirchhoff
Diretor Cienc.Esp.Atmos.

AUTOR RESPONSAVEL
RESPONSIBLE AUTHOR
Carlos J. Zamlutti

DISTRIBUIÇÃO/DISTRIBUTION
☐ INTERNA / INTERNAL
☒ EXTERNA / EXTERNAL
☐ RESTRITA / RESTRICTED

REVISADA POR / REVISED BY
Osmar Pinto Júnior
Editor Cienc Esp. Atmos.

CDU/UDC
519.6:681.322

DATA / DATE
Outubro 1990

PUBLICAÇÃO Nº
PUBLICATION NO
INPE-5196-MD/45

ORIGEM
ORIGIN
CEA/DAE

PROJETO
PROJECT
IONO

Nº DE PAG.
NO OF PAGES
195

ULTIMA PAG.
LAST PAGE
615

VERSÃO
VERSION

Nº DE MAPAS
NO OF MAPS

TÍTULO/TITLE
ANÁLISE DE MÉTODOS NUMÉRICOS
- MÉTODOS AUTOCORRETIVOS -
VOLUME 3

AUTORES/AUTHORSHIP
C.J. Zamlutti

RESUMO-NOTAS/ABSTRACT-NOTES

Apresenta-se um conjunto de métodos que, dentro da teoria das aproximações, são englobados num mesmo contexto com a denominação de métodos autocorretivos. A característica comum do conjunto é a possibilidade de determinar soluções que sucessivamente vão se aproximando do resultado exato desejado. Nos métodos apresentados, neste volume, o erro tolerado na aproximação pode ser estipulado previamente para qualquer método, embora não se possa ter uma noção exata do trabalho computacional necessário para obtenção da precisão desejada. Este volume contém os capítulos finais da terceira parte (Análise de Métodos Numéricos) do trabalho sobre os fundamentos da Análise Numérica, iniciado com a Publicação INPE-1937-MD/005.

OBSERVAÇÕES/REMARKS

Complemento das publicações INPE-1937-MD/005 (Volume 1) e INPE-5088-MD/043 (volume 2).

ABSTRACT

Under the denomination of self-correction methods, a set of methods are collected within the same context for the theory of approximation. The set is characterized by the possibility of refining successively the solutions until a prescribed accuracy is obtained. However, in the methods presented here the amount of computational work required to obtain the desired accuracy cannot be predicted. This volume presents the last chapters of the third part of the work "Análise de Métodos Numéricos" (Analysis of Numerical Methods), which started with the publication INPE-1937-MD/005.

SUMÁRIO

Pág.

TERCEIRA PARTE

ANÁLISE DE MÉTODOS NUMÉRICOS
- MÉTODOS AUTOCORRETIVOS -

CAPÍTULO 20

SOLUÇÃO DE EQUAÇÕES NÃO-LINEARES

20.1 - INTRODUÇÃO

Os tipos de equação de interesse prático em análise numérica podem ser colocados na forma geral:

$$\underline{\Psi}(\underline{v}) = \underline{0},$$

o que de imediato sugere a aplicação dos métodos iterativos e de relaxação para encontrar o valor de $\underline{v}$. Como estes métodos utilizam a própria equação $\underline{\Psi}(\underline{v}) = \underline{0}$ para aprimoramento de soluções aproximadas, eles são por vezes chamados autocorretivos.

A aplicação dos métodos iterativo e de relaxação nem sempre é simples e depende decisivamente de uma boa aproximação inicial para o seu sucesso. É interessante, então, dispor de métodos diretos que permitam a transformação inversa:

$$\underline{v} = \underline{\Psi}^{-1} \left[\underline{\Psi}(\underline{v})\right] = \underline{\Psi}^{-1}(\underline{0}),$$

ainda que apenas aproximadamente.

Neste capítulo, inicia-se o tratamento de equações, com o objetivo de encontrar a solução de equações não-lineares.

Inicialmente, prefere-se a abordagem unidimensional, a fim de proporcionar ao leitor um melhor entrosamento com o problema.

Na forma unidimensional, as equações não-lineares englobam basicamente dois tipos de problemas de maior incidência prática:

a) determinação de raízes reais de equações transcendentais da forma $f(x) = 0$;

b) determinação de todas raízes, reais ou imaginárias, da equação polinomial

$$p_n(z) = 0,$$

sendo n o grau do polinômio.

A ocorrência de raízes imaginárias em equações transcendentais constitui caso raro e, por isto, não será aqui discutida.

O problema de sistemas de equações não-lineares e programação não-linear será brevemente discutido, após o tratamento do caso unidimensional.

20.2 - SOLUÇÃO DE EQUAÇÕES TRANSCENDENTAIS UNIDIMENSIONAIS

As equações transcendentais unidimensionais podem ser colocadas na forma genérica:

$$f(x) = 0,$$

cuja solução é:

$$x = f^{-1}\left[f(x)\right] = f^{-1}(0) \quad ,$$

sempre que exista a função inversa f^{-1}. Nestes casos, pode-se aplicar a interpolação inversa e obter a solução diretamente.

Como nem sempre é possível usar o método direto, os métodos de aproximações sucessivas adquirem importância à medida que possuam convergência garantida e rápida. Estes métodos são apresentados a seguir.

20.2.1 - APROXIMAÇÕES SUCESSIVAS INDEPENDENTES

No caso de aproximações sucessivas independentes destaca-se pela sua importância o método do bisseccionamento, que constitui a aplicação imediata do teorema de Bolzano (Apostol, 1957);

Teorema: Se $f(x)$ é contínua em $[a, b]$, e o produto $f(a)\,f(b)<0$, então $f(x)$ possui pelo menos uma raiz em $[a, b]$.

1) Método do bisseccionamento

Seja o intervalo $[a, b]$, tal que $f(a)\,f(b) < 0$. Tem-se então pelo menos, uma raiz de $f(x)$ dentro deste intervalo. Chamando $x_1 = a$ e $x_2 = b$, calcula-se $f(x_m)$ para $x_m = \frac{x_1 + x_2}{2}$. Se $f(x_1)\,f(x_m) > 0$, faz-se $x_1 = x_m$ e repete-se o processo. Em caso contrário, faz-se $x_2 = x_m$ e repete-se o processo. Assim, o intervalo dentro do qual a raiz se encontra é reduzido à metade cada vez que o processo é repetido.

Após n repetições do processo, a amplitude do intervalo, em que se encontra uma raiz de $f(x)$ estará reduzida a:

$$\Delta x = (b - a)/2^n \ .$$

Assim, os valores de x_1 e x_2 para a enésima repetição aproximarão a raiz com um erro que não excede à amplitude Δx.

Pode eventualmente ocorrer que durante o processo se obtenha $f(x_m) = 0$. Neste caso, interrompe-se o processo porque o valor de x_m é a raiz procurada.

O método do bisseccionamento é sempre convergente, mas não se aplica a raízes de multiplicidade par, pois neste caso não são satisfeitas as condições do teorema de Bolzano nas vizinhaças dessas raízes.

2) Método da falsa posição (Regula Falsi)

Este método também é uma aplicação do teorema de Bolzano e pode ser considerado como uma variante do método do bissecionamento. A diferença entre esses métodos está no cálculo do ponto x_m que, no presente caso, é dado pela intersecção da corda que une os pontos $(x_1, f(x_1))$ a $(x_2, f(x_2))$ com o eixo dos x. O valor de x_m será então calculado pela expressão:

$$x_m = x_1 + \frac{|f(x_1)|}{|f(x_1)| + |f(x_2)|} (x_2 - x_1)$$

A convergência do método da falsa posição é garantida pelo seguinte teorema:

Teorema: o método da falsa posição converge em todos os casos de sua aplicação.

Prova: basta provar que $x_1 < x_m < x_2$, o que implica a redução da amplitude do intervalo $(x_2 - x_1)$ a cada repetição do processo.

Como por construção $x_1 < x_m < x_2$, exceto quando $f(x_1) = 0$ ou $f(x_2) = 0$, o teorema está provado, pois se $f(x_1)$ ou $f(x_2)$ se anula, o valor desejado da raiz já foi atingido. Uma demonstração mais rigorosa pode ser encontrada em Young and Gregory (1972).

Para examinar a rapidez da convergência, basta reescrever a expressão para o cálculo de x_m na forma:

$$x_m = x_1 + \frac{|f(x_1)|}{|f'(\xi)|}, \quad x_1 < \xi < x_2.$$

Assim, a velocidade de convergência depende do comportamento da relação:

$$\frac{|f(x_1)|}{|f'(\xi)|}.$$

O método da falsa posição apresenta as mesmas restrições que o do bisseccionamento para o caso de raízes de multiplicidade par.

Quando a função f(x) apresenta um comportamento monotônico, no intervalo (x_1, x_2), a amplitude do intervalo após um grande número de repetições do processo tende a um valor limite, dado por:

$$|x_2 - x_1| \equiv d \neq 0.$$

Nestes casos, a convergência é, em geral, mais lenta do que nos casos em que a amplitude do intervalo $|x_2 - x_1|$, tende a 0.

3) Modificação dos métodos para o caso de raízes múltiplas

À primeira vista pode parecer ao leitor que, no caso de raízes de multiplicidade par, não se dispõe de nenhum método para encontrar uma aproximação inicial para a raiz da equação f(x) = 0. A seguir, mostra-se um artifício que permite contornar este problema.

Teorema: A equação

$$u(x) = 0$$

com

$$u(x) = \frac{f(x)}{f'(x)}$$

possui as mesmas raízes da equação $f(x) = 0$, mas com multiplicidade simples.

Prova: Seja x o ponto para o qual $f(x)=0$ com multiplicidade m. A função f pode então ser decomposta como o produto de duas funções, na forma:

$$f(y) = (y - x)^m \, g(y)$$

com $g(x) \neq 0$. Neste caso, a função u(y) assume a forma:

$$u(y) = (y - x) \left[\frac{g(y)}{mg(y) + (y - x)\, g'(y)} \right]$$

Como $g(x) \neq 0$, o limite para $y \to x$ é o mesmo nos dois casos, f(y) e u(y), e vale 0. Assim, u(y) possui as mesmas raízes de f(y), mas com multiplicidade simples.

A utilização da função u(x), no lugar da função f(x), permite empregar os métodos do bisseccionamento e da falsa posição, mesmo nos casos de multiplicidade par das raízes. Elimina-se assim a restrição de aplicabilidade desses métodos. O artificio aqui apresentado pode ser usado também em conjunção com todos os métodos que se seguem, razão pela qual não se discutirá o problema de raízes multiplas nos métodos subsequentes.

20.2.2 - APROXIMAÇÕES ENCADEADAS, BASEADAS EM RELAXAÇÃO

Os métodos baseados em relaxação utilizam a própria igualdade:

$$f(x) = 0$$

para corrigir um valor aproximado, x_k, da solução, x, do problema.

Considerando o desenvolvimento de Taylor até o termo linear, pode-se escrever:

$$f(x) = f(x_k) + (x - x_k)\, f'(\xi) = 0,$$

sendo ξ um ponto intermediário entre x_k e x.

Como não se conhece o valor de ξ, não se pode determinar exatamente o valor de x, mas é possível, utilizando a expressão acima, calcular um "refinamento", x_{k+1}, para a aproximação da solução. Assim, tem-se:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(\xi_k)},$$

onde ξ_k é uma aproximação para ξ, com a restrição de que no limite:

$$x_k \to x, \quad \xi_k \to x.$$

Da forma utilizada para calcular $f'(\xi_k)$ derivam os diferentes métodos. Destacam-se pela sua importância os métodos da secante e de Newton.

1) Método da Secante

Este método usa a aproximação:

$$f'(\xi_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}},$$

e, então, a fórmula de recorrência torna-se:

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}.$$

Para que haja convergência local, é necessário que:

$$\frac{|x_{k+1} - x_k|}{|x_k - x_{k-1}|} < 1,$$

o que ocorre somente quando $f(x_k)$ e $f(x_{k-1})$ possuem sinais opostos. Entretanto considerando a convergência global, é possível, em alguns casos, dispensar esta condição (Ralston, 1965)

Para problemas localmente convergentes, é simples determinar uma estimativa para o erro, que é dado por:

$$\varepsilon = |x - x_{k+1}| = \left| x - x_k + \frac{f(x_k)}{f'(\xi_k)} \right|$$

Recordando que:

$$x = x_k - \frac{f(x_k)}{f'(\xi)}$$

e desenvolvendo $f'(\xi)$ em série de Taylor em torno do ponto ξ_k, tem-se:

$$\varepsilon = \left| \frac{f(x_k)\ f''(\eta)}{f'(\xi)\ f'(\xi_k)} (\xi - \xi_k) \right|$$

Como foi admitido que a função possui sinais opostos para pontos consecutivos, então ξ, ξ_k e η estão compreendidos no intervalo limitado pelos pontos x_k e x_{k-1}. Supondo-se também ξ e ξ_k muito próximos, pode-se limitar o erro por:

$$\varepsilon < |x_k - x_{k-1}| \left| \frac{f(x_k)\ f''(\eta_k)}{[f'(\xi_k)]^2} \right| < |x_k - x_{k-1}|^2 \left| \frac{f''(\eta_k)}{f'(\xi_k)} \right|,$$

onde $f''(\eta_k)$ é uma estimativa de $f''(\eta)$, que pode ser calculada por:

$$f''(\eta_k) = \frac{f'(\xi_k) - f'(\xi_{k-1})}{x_k - x_{k-1}} .$$

Para analisar a estabilidade do método, são feitas as seguintes considerações iniciais:

a) a precisão, ε, requerida para o cálculo de x é bem maior numericamente que o erro, δ, introduzido nos arredondamentos do computador;

b) o erro δ não altera muito os valores ξ_k a ponto de perturbar a estrutura do método.

Nessas condições, o valor ξ_k é subistituído por um valor ζ_k, que ainda tem a propriedade:

$$\lim_{x_k \to x} \zeta_k = x$$

A equação básica para o cálculo da propagação de erros é:

$$\varepsilon_{i+1} = \frac{\varepsilon_i - \varepsilon_i f'(x_k)}{f'(\zeta_k) + \delta} = \delta \quad ,$$

onde se supõe que as primeiras derivadas podem ser confundidas, havendo assim o cancelamento. Esta última expressão mostra que o método da secante é um método estável (ver Capítulo 6).

Este método possui algumas vantagens que o tornam preferido, mesmo em comparação com métodos mais sofisticados. As principais vantagens são:

- o método não exige o conhecimento prévio das derivadas de $f(x)$, que são calculadas numericamente;

- o método permite também o cálculo da raiz de equações do tipo $f(x, y_1, \ldots, y_n) = 0$, onde $y_i = \phi_i(x)$ $i = 1, \ldots, n$;

- o método não requer o uso de dupla precisão, apesar de envolver diferenças de números próximos quando se aproxima do valor desejado (Ralston, 1965).

2) Método de Newton

O método de Newton também chamado Newton Raphson, ou método das tangentes, emprega a aproximação:

$$f'(\xi_k) = f'(x_k) \quad ,$$

que resulta na fórmula de recorrência:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad .$$

Para que haja convergência local, deve-se ter:

$$\frac{|x_{k+1} - x_k|}{|x_k - x_{k-1}|} = \frac{|f(x_k)\ f'(x_{k-1})|}{|f(x_{k-1})\ f'(x_k)|} < 1,$$

que pode ser facilmente testada no decorrer do programa.

O cálculo do erro deste método é idêntico ao cálculo já efetuado para o método da secante. A única restrição é que o ponto ξ estará compreendido entre os pontos x_{k-1} e x. O limite do erro pode então ser escrito diretamente como:

$$\varepsilon < |x - x_{k-1}|^2 \left|\frac{f''(\eta_k)}{f'(x_k)}\right| .$$

A análise de estabilidade é idêntica à do método da secante e a conclusão também é a mesma, ou seja, o método de Newton é estável.

O método de Newton não apresenta as mesmas vantagens do método da secante, pois requer o cálculo da primeira derivada em cada ponto calculado.

20.2.3 - APROXIMAÇÕES SUCESSIVAS ENCADEADAS, GERADAS POR ITERAÇÃO

Os métodos de iteração baseiam-se na construção de uma relação de dependência funcional entre a solução e uma função dessa solução. Esta dependência funcional pode ser colocada na forma:

$$x = \phi\ (x)$$

que é conhecida como iteração com um ponto, sem memória.

Outras formas de iteração são construídas com o intuito de acelerar o processo de convergência ou melhor utilizar as informações conhecidas. destacam-se as formas:

a) Iteração com um ponto e memória:

$$x = \phi(x, \underline{\tilde{x}}),$$

onde $\underline{\tilde{x}}$ é o vetor construído com um conjunto de pontos aproximadores da solução.

b) Iteração com múltiplos pontos sem memória:

$$x = \phi\left[x, \underline{v}(x)\right],$$

onde $\underline{v}(x)$ indica um conjunto de funções que devem ser recalculadas para cada ponto novo. Difere da forma anterior por não usar resultado já calculado.

c) Iteração com múltiplos pontos e memória:

$$x = \phi\left[x, \underline{v}(x); \underline{\tilde{x}}, \underline{v}(\underline{\tilde{x}})\right]$$

onde $\underline{\tilde{x}}, \underline{v}(\underline{\tilde{x}})$ especificam a contribuição de outros pontos aproximadores da solução, em geral já calculados no processo de recorrência.

A seguir, apresentam-se os processos de construção de formas iterativas, com uma breve discussão de seus aspectos principais.

1) Método Iterativo com um ponto, sem memória

Considere-se a equação $f(x) = 0$. Então, tem-se também a equação:

$$h(x)\ f(x) = 0,$$

onde h(x) é uma função que não se anula no intervalo (a, b), dentro do qual o valor de x está compreendido.

Pode-se então escrever que:

$$x = x + h(x)\ f(x) = \phi(x),$$

e o processo de recorrência é gerado pela seqüência:

$$x_{k+1} = \phi(x_k) = x_k + h(x_k)\ f(x_k)$$

Para que haja convergência, é necessário que $\phi(x)$ satisfaça a condição de Lipschitz no intervalo (a, b), com $N < 1$ (ver Capítulo 6).

Introduz-se agora o conceito de "Ordem de Iteração".

Definição: Diz-se que uma iteração é de ordem m, se:

$$\phi(x) = \ldots = \phi^{(m-1)}(x) = 0,$$

onde x é a raiz da equação $f(x) = 0$.

Desenvolvendo $\phi(y)$ em série de Taylor em torno do ponto $y_0 = x$, mostra-se que para iteração de ordem m tem-se:

$$\phi(y) - x = \frac{(y - x)^m}{m!}\ \phi^{(m)}(\xi),$$

onde ξ é um ponto compreendido entre x e y. Segue-se então que:

$$x_{k+1} - x = \frac{(x_k - x)^m}{m!}\ \phi^{(m)}(\xi_k).$$

Esta última expressão permite analisar a convergência das iterações de ordem superior (Berezin and Zhidkov, 1965). O valor de m é também chamado ordem de convergência do processo.

A forma para gerar iterações de ordem superior é simples. Basta impor a condição

$$\phi^{(k)}(x) = 0, \quad k = 1, \ldots, m,$$

para obter o valor de h(y) e suas derivadas no ponto x. Pode-se, assim, usar a fórmula de Taylor para representar h(y). Como exemplo, impondo-se a condição $\phi'(x) = 0$, obtém-se $h(y) = -1/f'(\xi)$ que resulta na mesma expressão dos métodos de relaxação.

Uma discussão sobre vantagens e desvantagens da aplicação de iterações de ordem superior pode ser encontrada em Traub (1964).

A expressão que se obtém com o método aqui apresentado, para geração de iterações de ordem superior, coincide com a expressão do chamado método de Schröder (Korganoff, 1962).

A estabilidade dos métodos iterativos já foi discutida no caso geral (Capítulo 6). Sua particularização para os casos em discussão é imediata.

2) Método iterativo com um ponto e memória

A obtenção de métodos iterativos com memória é feita a partir de fórmulas interpoladoras, ou do uso da fórmula de Taylor, empregando pontos já calculados como "memória" para obtenção de estimativa das derivadas (Traub, 1964). Apresenta-se aqui apenas uma forma simples, baseada na fórmula interpoladora de Newton.

Considere-se a fórmula interpoladora de Newton:

$$N(x) = f[x_k] + (x - x_k)\, f[x_k, x_{k-1}] + (x - x_k)(x - x_{k-1})\, f[x_k, x_{k-1}, x] \quad .$$

Fazendo $N(x) = 0$ e reagrupando os termos, resulta em:

$$x = x_k - \frac{f(x_k)}{f[x_k, x_{k-1}] + (x - x_{k-1})\, f[x_k, x_{k-1}, x]} \quad .$$

Usando $x = x_{k+1}$ do lado esquerdo desta expressão e $x = x_{k-2}$ do lado direito, obtém-se a fórmula de recorrência:

$$x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, x_{k-1}] + (x_{k-2} - x_{k-1})\, f[x_k, x_{k-1}, x_{k-2}]} \quad ,$$

que possui uma identidade formal com as fórmulas resultantes de relaxação. De fato, o denominador da expressão acima nada mais é do que uma forma mais aprimorada para calcular $f'(\xi)$.

Assim, desprezando a última parcela do denominador, obtém-se a conhecida expressão do método da secante.

Pode-se demonstrar facilmente que o erro deste método possui o mesmo limite que o do método da secante, não apresentando assim decisiva vantagem. De fato, para conseguir melhores resultados, devem ser usadas iterações de ordem superior conjuntamente com diferenças finitas para o cálculo das derivadas. Estes métodos não são aqui discutidos.

3) Método iterativo com vários pontos sem memória

Uma forma possível e simples de gerar estes métodos baseia-se nas duas equações primitivas:

$$f(x) = 0 \ ,$$

$$x = \phi(x),$$

de onde se pode escrever que:

$$\phi(x) = \phi\ (x) + h(x)\ f[\ \phi(x)\ ],$$

onde h (x) é uma função que não muda de sinal no intervalo (a, b), dentro do qual se encontra o valor desejado, x, da raiz.

A função iterativa é então gerada pela sequência:

$$\phi_{k+1}(x) = \phi_k(x) + h(x)\ f\ [\ \phi_k(x)\ ] \ ,$$

$$\phi_1(x) = x,$$

sendo h(x) escolhida para aumento da ordem de convergência.

Como exemplo seja:

$$h(x) = -1/f'(x) \text{ e } k = 3.$$

Tem-se então:

$$\phi_3(x) = x - \frac{f(x)}{f'(x)} - \frac{f[\ x - f(x)/f'(x)\ ]}{f'(x)},$$

e o processo de recorrência dado por:

$$x_{i+1} = \phi_3(x_i) \quad .$$

A função $\phi_3(x)$ pode ser colocada na estrutura formal $\phi[x, \underline{v}(x)]$, fazendo-se:

$$v_1(x) = - \frac{f(x)}{f'(x)} \quad ,$$

$$v_2(x) = - \frac{f[x - f(x)/f'(x)]}{f'(x)} \quad .$$

Para a implementação do método em computadores, não é necessário explicitar a forma analítica de $\phi_k(x)$, como foi feito no exemplo acima. Os valores $\phi_k(x_i)$ são obtidos para cada ponto x_i, usando-se processo de recorrência para o cálculo de $\phi_k(x)$.

O efeito da recorrência de $\phi_k(x)$, na convergência do método, pode ser facilmente mostrado. Chamando:

$$E_{k,i} = x_i - \phi_k(x_i)$$

e usando a fórmula de recorrência para cálculo das $\phi_k(x)$, tem-se:

$$E_{k+1,i} = E_{k,i} + E_{k,i}\, h(x_i)\, f'(x_i) - \frac{E_{k,i}^2}{2} h(x_i)\, f''(x_i) + \ldots \quad ,$$

onde $f[\phi_k(x)]$ foi desenvolvida em série de Taylor em torno do ponto x_i e $h(x_i)\, f(x_i)$ foi suposto nulo.

Para $h(x) = - 1/f'(x)$, resulta aproximadamente em:

$$E_{k+1,i} \cong \frac{E_{k,i}^2}{2} \frac{f''(x_i)}{f'(x_i)}$$

e finalmente:

$$E_{k+1,i} = \left[\frac{E_{1,i}^2}{2} \quad \frac{f''(x_i)}{f'(x_i)} \right]^k .$$

Esta última expressão ilustra o efeito da utilização de de vários pontos em confronto com a iteração que usa um único ponto.

4) Método iterativo com vários pontos e memória

A função iterativa, neste caso, pode ser gerada pela mesma fórmula de recorrência do método anterior, ou seja:

$$\phi_{k+1}(x) = \phi_k(x) + h(x)\, f[\, \phi_k(x)\,].$$

Neste caso, entretanto, a função inicial $\phi_1(x)$ é dotada de memória. Pode-se usar, por exemplo, a expressão do método da secante:

$$\phi_1(x_i) = x_1 - \frac{f(x_i)}{f[\, x_i, x_{i-1}\,]} .$$

Este método possui essencialmente as características de convergência do método anterior, acrescidas das vantagens de uma função iterativa inicial, rapidamente convergente.

20.3 - SOLUÇÃO DE SISTEMAS DE EQUAÇÕESTRANSCENDENTAIS, PROGRAMAÇÃO NÃO-LINEAR.

O sistema de equações não-lineares:

$$\Psi_i(v_i, \ldots, v_n) = 0, \qquad i = 1, \ldots, n$$

é escrito em notação vetorial:

$$\underline{\Psi}\ (\underline{v}) = \underline{0}$$

que é a equação básica para o desenvolvimento do método de relaxação. A solução do problema, colocado nesta forma, já foi discutida no Capítulo 6, que engloba os métodos iterativos e de relaxação.

O problema de encontrar a solução de um sistema de equações não-lineares pode ser convertido no problema de otimização de funções não-lineares de várias variáveis. Para isto, basta construir a função:

$g(\underline{v}) = < \underline{\Psi}(\underline{v}), \underline{\Psi}(\underline{v}) > = |\underline{\Psi}(\underline{v})|^2$ que possui um mínimo $\underline{\Psi}(\underline{v}) = \underline{0}$.

A recíproca é verdadeira, pois a procura de valores extremos locais é equivalente à solução do problema de equações não-lineares:

$$\frac{\partial g(\underline{v})}{\partial v_i} = 0 \quad , \qquad\qquad i = 1, \ldots, n \ .$$

20.3.1 - MÉTODO DO GRADIENTE

O méotodo do gradiente é também denominado método do declive máximo (steepest descent).

O gradiente da função g(x) é um vetor $\underline{d}$ de componentes:

$$d_i = \frac{\partial g(\underline{v})}{\partial v_i} \quad ,$$

que indica a direção de máxima variação da função.

Inicia-se o processo de recorrência supondo que o vetor $\underline{v}_0$ constitui uma aproximação da solução. Faz-se o aprimoramento da solução pelo vetor:

$$\underline{v}_1 = \underline{v}_0 - t\,\underline{d} \quad .$$

Obtém-se o valor de t impondo a condição da função:

$$f(t) = g\,(\underline{v}_0 - t\,\underline{d})$$

ser mínima.

Para obter este valor de t, é necessário calcular o valor da função g para um conjunto de pontos.

Para obtenção do mínimo o método do gradiente pode ser usado recursivamente, agora sobre a função f(t).

Obtém-se o processo iterativo a partir da recorrência, da aproximação de ordem k para a aproximação de ordem k+1, resultando em:

$$\underline{v}_{k+1} = \underline{v}_k - t_k\,\underline{d}_k \quad ,$$

$$f_k(t_k) = g(\underline{v}_k - t_k\,\underline{d}_k) \quad .$$

Uma grande quantidade de operações está envolvida na obtenção de cada aproximação no método do gradiente. Para simplificá-lo, prefere-se, por vezes, usar apenas a máxima derivada de cada ponto, ao invés de todas as derivadas. Para a aproximação de ordem k, tem-se:

$$d_{j,k} = \max_i d_{i,k}, \qquad i = 1, \ldots, n ,$$

$$d_{i,k} = \left.\frac{\partial g(\underline{v})}{\partial v_i}\right|_{\underline{v} = \underline{v}_k} ,$$

$$\underline{v}_{k+1} = \underline{v}_k - t_k \, d_{j,k} \, \underline{\ell}_j ,$$

onde $\underline{\ell}_j$ é o versor da direção j. Como apenas a componente desta direção é afetada, seu valor $v_{k+1,j}$ pode ser calculado diretamente, pois ele é a solução da equação:

$$\left.\frac{\partial}{\partial v_j} g(\underline{v})\right|_{\underline{v} = \underline{v}_{k+1}} = 0 \quad .$$

Dispensa-se assim o processo iterativo para o cálculo do valor t_k.

Embora nesta variante do método sejam necessárias mais iterações, no compto geral ela torna-se competitiva por economizar a recorrência no cálculo de t_k.

20.4 - SOLUÇÃO DE EQUAÇÕES ALGÉBRICAS.

Os métodos discutidos anteriormente podem ser aplicados também à equação do tipo:

$$P_n(z) = 0,$$

onde $P_n(z)$ é um polinômio de grau n. Como, entretanto, várias raízes diferentes podem estar contidas dentro de um intervalo relativamente reduzido, este problema possui maior complexidade.

O tratamento das equações algébricas será aqui desenvolvido em 3 etapas:

a) separação das raízes,

b) aproximação segura de uma solução,

c) aprimoramento de soluções aproximadas.

20.4.1 - SEPARAÇÃO DAS RAÍZES

Historicamente, a regra de sinais de Descartes parece ter sido a forma mais eficaz de que se dispunha para a separação das raízes. Atualmente, sua relevância é devida à sua simplicidade e importância para a compreensão de métodos mais eficazes.

- Regras de Sinais de Descartes

O número de raízes positivas de $P_n(z) = 0$ nunca é maior que o número de variações de sinais na seqüência de coeficientes a_0, a_1, a_2, ..., a_n omitidos os coeficientes nulos.

O limite do número de raízes negativas pode ser obtido usando este mesmo teorema para os coeficientes a_i', $i = 0,...n$, do polinômio $P_n(t)$, onde $t = -z$, pois as raízes positivas de $P_n(t) = 0$ são as raízes negativas de $P_n(z) = 0$.

A regra de sinais pode também ser usada para determinar se as raízes excedem um valor escolhido M, quando aplicada ao polinômio $P_n(z-M)$.

O limite do número de raízes positivas e negativas, dado pela regra de Descartes, nem sempre é atingido, daí a necessidade de métodos mais eficientes para a determinação do número exato de raízes reais. O teorema de Sturm é usado com este propósito.

Antes do teorema de Sturm, apresenta-se o cálculo do máximo divisor comum entre dois polinômios.

1) Máximo Divisor Comum entre dois polinômios

Dados dois polinômios: $P_n(z)$ de grau n, e $T_m(z)$ de grau m, chama-se máximo divisor comum entre eles o polinômio $Q_k(z)$, do maior grau k, que divide exatamente tanto $P_n(z)$ como $T_m(z)$.

O máximo divisor comum (m.d.c) não é único, pois se $Q_k(z)$ é o m.d.c. entre $P_n(z)$ e $T_m(z)$ então $c\,Q_k(z)$ com c uma constante arbitrária também é o m.d.c. entre $P_n(z)$ e $T_m(z)$.

O processo para encontrar o m.d.c. entre dois polinômios é análogo ao processo para encontrar o m.d.c. entre dois números.

Seja $n \geq m$. Pode-se dizer que:

$$P_n(z) = T_m(z)\, q_1(z) + \Gamma_1(z) \quad ,$$

ou reescrevendo-se:

$$P_n(z) = \Gamma_1(z) \left[\frac{T_m(z)}{\Gamma_1(z)}\, q_1(z) + 1 \right] \quad .$$

De modo análogo, pode-se obter sucessivamente:

$$T_m(z) = \Gamma_1(z)\, q_2(z) + \Gamma_2(z) = \Gamma_2(z) \left[\frac{\Gamma_1(z)}{\Gamma_2(z)}\, q_2(z) + 1 \right] \quad ,$$

$$\Gamma_1(z) = \Gamma_2(z)\, q_3(z) + \Gamma_3(z) = \Gamma_3(z) \left[\frac{\Gamma_2(z)}{\Gamma_3(z)}\, q_3(z) + 1 \right] \quad ,$$

$$\Gamma_p(z) = \Gamma_{p+1}(z)\, q_{p+2}(z) + \Gamma_{p+2} = \Gamma_{p+2}(z) \left[\frac{\Gamma_{p+1}(z)}{\Gamma_{p+2}(z)}\, q_{p+2}(z) + 1 \right] .$$

Seja i o menor índice, tal que $\Gamma_{i+2}(z) = 0$. Segue-se, que

$$\Gamma_i(z) = \Gamma_{i+1}(z)\, q_{i+2}(z)$$

e assim:

$$\Gamma_{i-1}(z) = \Gamma_i(z)q_{i+1}(z) + \Gamma_{i+1}(z) = \Gamma_{i+1}(z)[q_{i+1}(z)q_{i+2}(z) + 1] \ ;$$

então pela lei de recorrência, o fato de $\Gamma_i(z)$ ser múltiplo de $\Gamma_{i+1}(z)$ implica que $\Gamma_{i-1}(z)$ também o seja, e assim sucessivamente $\Gamma_{i-1}(z), \Gamma_{i-3}(z)$ etc., serão múltimplos de $\Gamma_{i+1}(z)$. Segue-se que $T_m(z)$ e $P_n(z)$ também serão múltiplos de $\Gamma_{i+1}(z)$. O polinômio $\Gamma_{i+1}(z)$, assim encontrado, é o polinômio de maior grau que divide exatamente $P_n(z)$ e $T_m(z)$. Logo, $\Gamma_{i+1}(z) = Q_k(z) = \text{m.d.c.}\ [P_n(z), T_m(z)]$. O esquema abaixo é a forma prática de execução das divisões sucessivas para o cálculo do m.d.c..

	QUOCIENTE	$q_1(z)$	$q_2(z)$	$q_3(z)$	.....	$q_{i+1}(z)$	$q_{i+2}(z)$
	$P_n(z)$	$T_m(z)$	$\Gamma_1(z)$	$\Gamma_2(z)$	.....	$\Gamma_i(z)$	$\Gamma_{i+1}(z)$
RESTO →	$\Gamma_i(z)$	$\Gamma_2(z)$	$\Gamma_3(z)$	$\Gamma_4(z)$	.....	0	

Pela própria definição do m.d.c., multiplicações por constantes durante o processo de divisões sucessivas não alteram o valor do m.d.c.. Assim, para não trabalhar com números fracionários, usualmente multiplica-se o resto $\Gamma_j(z)$ por uma constante a_j, tal que o polinômio $a_j\Gamma_j(z)$ só possua coeficientes inteiros.

2) Seqüências de Sturm

Definição 1 - Seja $f_1(z), \ldots, f_n(z)$ uma seqüência de polinômios. Esta é a chamada seqüência Sturm em (a,b), podendo a e b serem infinitos, se satisfazer as duas condições:

- $f_n(z)$ não se anula em (a, b).

- Para todo zero de $f_K(z)$, $K = 2, \ldots, n-1$, as duas funções vizinhas são diferentes de zero e têm sinais opostos, i.e., $f_{K-1}(z)\, f_{K+1}(z) < 0$.

Definição 2 - Seja $\{f_j(z)\}$ uma seqüência de Sturm em (a, b) com n elementos. Considere-se um ponto z_0 de (a,b), tal que $f_1(z_0) \neq 0$. Define-se então $V(z_0)$ como o número de variações de sinal de seqüência $\{f_j(z_0)\}$ desprezando-se seus valores nulos. Se a é finito, V(a) é definida como $V(a+\varepsilon)$ com ε suficientemente pequeno para que nenhuma das funções $f_j(z)$ se anule no intervalo $(a, a+\varepsilon)$. Analogamente, define-se V(b) para b finito. Se $a = -\infty$, V(a) é definido pela variação de sinais $\{\lim_{x\to-\infty} f_j(z)\}$. Analogamente, define-se V(b) com $b = +\infty$.

Definição 3 - Seja R(z) uma função racional. Define-se o índice de Cauchy de R(z) em (a, b), $C_R(a, b)$ como a diferença entre o número de vezes que R(z) salta de $-\infty$ para $+\infty$ e o número de vezes que R(z) salta de $+\infty$ para $-\infty$, quando z varia de a até b.

- *Teorema de Sturm* - Seja $\{f_j(z)\}$ uma seqüência de Sturm em (a,b) e $R(z) = f_2(z)/f_1(z)$. Se $f_1(a) \neq 0$ e $f_1(b) \neq 0$, então:

$$C_R(a, b) = V(a) - V(b) .$$

Prova: O valor de V(z) não se altera quando z passa por um zero de $f_K(z)$, $K = 2, \ldots, n-1$ por causa da restrição imposta na Definição 1 acima. Assim, o valor de V(z) só se altera quando $f_1(z)$ passa por um zero.

Por outro lado, sendo z_0 um zero de $f_1(z)$, ele não é um zero de $f_2(z)$ pela definição 1. Assim, se z_0 é um zero de multiplicidade par, $V(z)$ não muda quando z passa por z_0. Entretanto, se z_0 é um zero de multiplicidade ímpar, $V(z)$ aumenta de uma unidade, se $f_1(z)$ e $f_2(z)$ possuem o mesmo sinal à esquerda de z_0, e decresce de uma unidade, se $f_1(z)$ e $f_2(z)$ possuem sinais diferentes à esquerda de z_0. Também o índice de Cauchy não se altera se a multiplicidade da raiz de $f_1(z) = 0$ for par, e varia de uma unidade se ela for ímpar, exatamente como a variação de $V(z)$, o que demonstra o teorema.

Pode-se agora aplicar o teorema de Sturm para a determinação do número de raízes reais de uma equação $P_n(z) = 0$.

Considere-se o processo para encontrar o m.d.c entre $P_n(z)$ e sua derivada $P_n'(z)$:

$$P_n(z) = q_i(z)\, P_n'(z) + \Gamma_1(z) \quad ,$$

$$\Gamma_p(z) = q_{p+2}(z)\, \Gamma_{p+1}(z) + \Gamma_{p+2}(z) \quad ,$$

$$\Gamma_{i-1}(z) = q_{i+1}(z)\, \Gamma_i(z) \quad .$$

Seja a seqüência $f_j(z)$, $j = 1, \ldots$ de polinômios:

$$f_1(z) = P_n(z) \quad ,$$

$$f_2(z) = P_n'(z) \quad ,$$

$$f_{\kappa+2}(z) = q_{\kappa+2}(z)\, f_{\kappa+3} - f_{\kappa+4}(z) \quad , \qquad \kappa = 1, \ldots, i-4 \quad ,$$

$$f_{i-1}(z) = q_{i+1}(z)\, f_i(z) \quad .$$

Assim, os $f_j(z)$ são os restos $r_j(z)$ do processo do cálculo do m.d.c. afetados por uma constante multiplicativa.

Conforme mostrado anteriormente, nessas condições $f_i(z)$ é o m.d.c entre $P_n(z)$ e $P_n'(z)$ e divide exatamente todos os $f_\kappa(z)$, $\kappa = 1, \ldots, i$.

Supõe-se que $f_i(z)$ não se anula em (a,b) para satisfazer a primeira condição de definição 1. Nestas condições, se $f_\kappa(z) = 0$ para um z qualquer dentro de (a, b), a segunda condição da definição 1 é automaticamente satisfeita, pois se segue que:

$$f_{\kappa-1}(z) = - f_{\kappa+1}(z) , \qquad \kappa = 2, \ldots, i - 2$$

Adicionalmente, se $f_\kappa(z) = 0$, necessariamente $f_{\kappa+1}(z) \neq 0$, pois caso contrário o m.d.c entre os polinômios é nulo para este particular z, logo $f_j(z)$ é nulo, contrariando a hipótese original.

Com essa construção, as $\{f_j(z)\}$ constituem uma seqüência de Sturm em (a, b), podendo-se aplicar o teorema de Sturm ao polinômio $P_n(z)$ e sua derivada $P_n'(z)$.

Se $f_i(z)$ se anular em (a,b), dividem-se todas as $f_j(z)$, $j = 1, \ldots, i$ por $f_i(z)$ resultando assim uma seqüência de Sturm, dada por $\{f_j(z)/f_i(z)\}$, $j = 1, \ldots, i$.

Para a seqüência de Sturm construída, tem-se:

$$R(x) = \frac{f_2(z)}{f_1(z)} = \frac{P_n'(z)}{P_n(z)} = \sum_{j=1}^{p} \frac{m_j}{z - z_j} + R_1(z),$$

onde z_j são as raízes reais distintas de $P_n(z)$, m_j são as multiplicidades destas raízes, e $R_1(z)$ é um polinômio que não possui raízes reais.

Como os m_j são positivos, o índice de Cauchy, $C_R(a, b)$, fornece o número de raízes reais distintas dentro de (a, b). Fazendo $a = -\infty$ e $b = +\infty$, tem-se o número total de raízes distintas. A multiplicidade destas raízes pode ser obtida, lembrando que o m.d.c. entre $P_n(z)$ e $P_n'(z)$ é o produto das raízes múltiplas de $P_n(z)$ com as multiplicidades reduzidas de uma unidade.

O leitor deve notar que a forma de teste das seqüências de Sturm apresenta grande semelhança com a regra de sinais de Descartes.

As seqüências de Sturm resolvem completamente o problema da determinação do número de raízes reais contidas num intervalo (a,b) escolhido. Variando a localização do intervalo, consegue-se separar todas as raízes reais.

3) Critério de Lehmer Schur

O critério de Lehmer Schur completa a solução do problema de separação das raízes, determinando se em um dado círculo do plano complexo existe uma raiz de $P_n(z)$. Uma vez que as raízes reais já foram separadas pelo método de Sturm, as raízes adicionais, determinadas por este método, são todas complexas.

Considerando o polinômio $P_n(z)$, dado por:

$$P_n(z) = \sum_{i=0}^{n} a_i z^i \quad ,$$

constrói-se o polinômio $P_n^{\sim}(z)$ pela relação:

$$P_n^{\sim}(z) = \sum_{i=0}^{n} a_{n-i}^{*} z^i = z^n P_n^{*} \left[(z^*)^{-1} \right] \quad ,$$

onde se usa o asterisco para indicar complexo conjugado.

Definindo a função redutora de grau T por:

$$T\left[P_n(z)\right] = a_0^* P_n(z) - a_n \tilde{P_n}(z),$$

resulta, por recorrência, em:

$$T^j\left[P_n(z)\right] = T\{T^{j-1}\left[P_n(z)\right]\}, \quad j = 2, 3 \ldots$$

Antes de passar ao critério de Lehmer - Schur, é necessário recordar três teoremas da teoria de variáveis complexas.

Teorema 1: (Cauchy) - para o círculo de raio unitário, é válido o resultado

$$\int_C \frac{dz}{z - a} = \begin{cases} 2\pi i & \text{se } |a| < 1 \\ 0 & \text{se } |a| > 1, \end{cases}$$

onde C é a circunferência de raio unitário, orientada no sentido positivo, e $i = \sqrt{-1}$.

Prova - basta usar a transformação de variáveis $z = a + r\,e^{i\theta}$, obtendo-se uma integral em θ com variação de 0 a 2π.

Teorema 2 - Seja p(z) um polinômio que não possui raízes na circunferência de raio unitário. O número de raízes de p(z) no círculo de raio unitário é dado por:

$$N = \frac{1}{2\pi i} \int_C \frac{p'(z)}{p(z)}\, dz \quad .$$

Prova - escrevendo $p(z)$ como o produto de monômios na forma:

$$p(z) = A \prod_{j=1}^{n} (z - z_j),$$

onde A é uma constante e z_j são as raízes de $p(z)$, obtém-se:

$$\int_C \frac{p'(z)}{p(z)}\, dz = \sum_{j=1}^{n} \int_C \frac{dz}{z - z_j} \quad .$$

Aplicando o Teorema 1, o teorema fica demonstrado.

Teorema 3 - (Rouchê): Seja $p(z)$ e $q(z)$ dois polinômios que satisfazem a condição:

$$|q(z)| < |p(z)| \quad \text{para} \quad |z| = 1 \quad ,$$

então $p(z)$ e $p(z) + q(z)$ possuem o mesmo número de raízes no círculo de raio unitário.

Prova - é suficiente provar que $p(z) + q(z)$ não possui nenhuma raiz adicional no círculo de raio unitário e, então, pelo teorema 2 tem-se a tese.

Colocando $p(z) + q(z)$ na forma:

$$p(z) + q(z) = p(z) \left[1 + \frac{q(z)}{p(z)}\right] \quad ,$$

vê-se que, para que haja uma raiz adicional no círculo de raio unitário, é necessário que a expressão entre colchetes se anule. Isto entretanto não é possível, pois $|q(z)| < |p(z)|$ para $|z| = 1$. Logo, $p(z) + q(z)$ possui somente as raízes de $p(z)$ no círculo de raio unitário.

Pode-se agora demonstrar o teorema abaixo, (Lehmer, 1961) de capital importância para o critério de Lehmer Schur.

Teorema 4 - Seja $P_n(z)$ um polinômio de grau n que não possui raízes na circunferência de raio unitário. Se $T[P_n(0)] \neq 0$, então:

a) $P_n(z)$ e $T[P_n(z)]$ possuem o mesmo número de raízes no círculo de raio unitário, se $T[P_n(0)] > 0$;

b) $P_n^{\sim}(z)$ e $T[P_n(z)]$ possuem o mesmo número de raízes no círculo de raio unitário, se $T[P_n(0)] < 0$

Prova - para circunferência de raio unitário, os módulos de $P_n(z)$ e $P_n^{\sim}(z)$ são iguais.

Examina-se agora a possibilidade de $T[P_n(z)]$ possuir uma raiz na circunferência de raio unitário.

Seja β esta raiz. Então, tem-se:

$$T[P_n(\beta)] = a_0^* P_n(\beta) - a_n P_n^{\sim}(\beta) = 0$$

Como $|P_n(\beta)| = |P_n^{\sim}(\beta)| \neq 0$, deve-se ter então:

$$|a_0^*| = |a_n|.$$

Para que isto seja possível, deve-se ter também:

$$T[P_n(0)] = |a_0^*|^2 - |a_n|^2 = 0,$$

que contraria a hipótese inicial. Logo $T[P_n(z)]$ não se anula na circunferência de raio unitário.

Considere-se agora o caso $T[P_n(0)] > 0$. Seja

$$p(z) = a_o^* P_n(z) \ ,$$

$$q(z) = - a_n \tilde{P_n}(z),$$

de modo que $T[P_n(z)] = p(z) + q(z)$.

Para aplicar o Teorema 3 aos polinômios $T[P_n(z)]$ e $p(z)$, deve-se mostrar que para a circunferência de raio unitário $|p(z)| > |q(z)|$. Isto é simples, pois para esta circunferência os módulos de $P_n(z)$ e $\tilde{P_n}(z)$ são iguais. Como $T[P_n(0)] > 0$ tem-se $|a_o^*| > |a_n|$, de onde se segue que $|p(z) > |q(z)|$ no contorno de raio unitário.

Pelo Teorema 3, $T[P_n(z)]$ e $p(z)$ possuem o mesmo número de raízes no circulo de raio unitário. Portanto, $T[P_n(z)]$ e $P_n(z)$ possuem o mesmo número de raízes neste círculo.

Quando $T[P_n(0)] < 0$, faz-se:

$$p(z) = - a_n \tilde{P_n}(z) \ ,$$

$$q(z) = a_o^* \ P_n(z) \ ,$$

e aplicando o Teorema 3 tem-se que $T[P_n(z)]$ e $\tilde{P_n}(z)$ possuem o mesmo número de raízes no círculo de raio unitário.

É interessante observar que cada raiz de $\tilde{P_n}(z)$ dentro do círculo de raio unitário corresponde univocamente a uma raiz de $P_n(z)$ fora deste círculo.

A seguir apresenta-se o Teorema de Schur que constitui o fundamento do método de Lehmer-Schur (Lehmer, 1961).

Teorema 5: (Schur) - seja $P_n(z)$ um polinômio que não possui raízes na circunferência de raio unitário. Suponha-se que $P_n(0) \neq 0$. Seja m o menor valor, tal que:

$T^m[P_n(0)] = 0$, então:

a) se $T^k[P_n(0)] > 0$ para $k = 1, \ldots, m - 1$ e $T^{m-1}[P_n(z)]$ é uma constante, o polinômio $P_n(z)$ não possui raízes no círculo de raio unitário;

b) se $T^k[P_n(0)] > 0$, $k = 0, \ldots, i - 1$ e $T^i[P_n(0)] < 0$, o polinômio $P_n(z)$ possui, pelo menos, uma raíz dentro do círculo de raio unitário.

Prova:

a) pelo Teorema 4, todos os polinômios da seqüência:

$$P_n(z),\ T[P_n(z)],\ \ldots,\ T^{m-1}[P_n(z)]$$

possuem o mesmo número de raízes no círculo de raio unitário. Como $T^{m-1}[P_n(z)]$ é uma constante não-nula, não possui raízes neste círculo. O mesmo ocorre com todos os polinômios da sequência. Logo, $P_n(z)$ não possui raízes neste círculo.

b) pelo Teorema 4 todos os componentes da seqüência:

$$P_n(z), T[P_n(z)],\ \ldots,\ T^{i-1}[P_n(z)]$$

possuem o mesmo número n_1 de raízes dentro do círculo de raio unitário. Por outro lado, seja t_j o grau do polinômio:

$$T^j[P_n(z)], \quad j = 0, 1, 2, \ldots .$$

Em virtude da redução sucessiva de graus, tem-se:

$$n = t_0 > t_1 > \ldots$$

Assim, na seqüência acima o menor grau é t_{i-1}. O número de raízes de $T^{i-1}[P_n(z)]$ fora do círculo de raio unitário é $t_{i-1} - n_1$, e este é o número de raízes de $T^i[P_n(z)]$ dentro deste círculo. Tem-se então a relação:

$$t_i \geq t_{i-1} - n_1,$$

de onde se segue que:

$$n_1 \geq t_{i-1} - t_i \geq 1,$$

o que mostra que, pelo menos, uma raiz de $P_n(z)$ está contida no círculo de raio unitário.

A aplicação do Teorema 5, para a pesquisa de raízes no plano complexo deve-se a Lehmer. Apresenta-se aqui uma variante para a separação de uma coroa circular do plano complexo, dentro da qual se encontra uma raiz do polinômio. O método é idêntico ao do bissecionamento apresentado na Seção 20.2.

Primeiramente, observe-se que se $P_n(zR)$ possui uma raiz dentro do círculo de raio unitário, então $P_n(z)$ possui uma raiz num círculo de raio R. Diz-se que o valor R é o fator de escala.

Variando o fator de escala, por exemplo, em potências positivas ou negativas de 2, pode-se determinar uma coroa circular, dentro da qual existe pelo menos uma raiz do polinômio $P_n(z)$. Seja R_i o raio interno desta coroa e R_e o seu raio externo, de modo que:

a) $P_n(z)$ não possui nenhuma raiz no círculo de R_i,

b) $P_n(z)$ possui pelo menos uma raiz dentro do círculo de R_e.

Procede-se, em seguida, ao bisseccionamento da coroa circular em duas outras, uma interna e outra externa, de áreas iguais, escolhendo um raio R_m dado por:

$$R_m = + \sqrt{(R_e^2 + R_i^2)/2} \ .$$

Se $P_n(z)$ possui uma raiz dentro do círculo de raio R_m, escolhe-se o novo valor de R_e igual a R_m. Caso contrário, R_m substitui o valor de R_i.

Repetindo o processo, reduz-se a faixa de incerteza, $R_e - R_i$, dentro da qual se encontra a raiz desejada. Pode-se, em seguida, usar o método de aproximações sucessivas para encontrar o valor da raiz.

4) Modificação dos métodos para o caso de raízes múltiplas

Como no caso de equações transcendentais, prefere-se também neste caso evitar a ocorrência de raízes múltiplas.

A modificação sugerida no item 3 da Seção 20.2.1 também é aplicável a equações polinomiais, mas acarreta o problema de trabalhar com funções racionais no lugar de polinômios. Para evitar este inconveniente, o artifício a seguir é recomendável.

Considere-se $Q_k(z)$ como o m.d.c. entre os polinômios $P_n(z)$ e $P_n'(z)$. O polinômio:

$$T_m(z) = P_n(z)/Q_k(z)$$

não possui raízes múltiplas.

Utilizando $T_m(z)$ em lugar de $P_n(z)$, nos métodos para solução de equações polinomiais, elimina-se o problema de raízes múltiplas.

Portanto, nos métodos que são apresentados a seguir, a ocorrência de raízes múltiplas não necessita ser considerada.

20.4.2 - APROXIMAÇÕES SUCESSIVAS INDEPENDENTES

Da mesma forma que para equações transcendentais, existe, neste caso, o interesse em dispor de métodos que garantidamente forneçam uma aproximação da raiz.

Para o caso de raízes reais, pode-se usar o bissecionamento de intervalo conjuntamente com as seqüências de Sturm. Semelhante procedimento é aplicável para o caso de raízes complexas, como mostra o Item 3 da Seção 20.4.1.

Apresenta-se, aqui, um método alternativo para o propósito em pauta. Este método conhecido como o do quadrado das raízes (ou de Graeffe) é discutido a seguir.

1) Método do quadrado das raízes

Este método baseia-se na elevação sucessiva das raízes de uma equação polinomial ao quadrado, separando-as assim gradativamente. O processo é bastante simples. Dado o polinômio $P_n(z)$, produz-se um polinômio $B_n(Z)$ pela relação:

$$B_n(Z) = (-1)^n P_n(z) P_n(-z), \qquad Z = z^2 ,$$

cujas raízes são o quadrado das raízes de $P_n(z)$.

Em termos de coeficientes polinomiais, se

$$P_n(z) = \sum_{j=0}^{n} a_j z^j$$

e

$$B_n(Z) = \sum_{j=1}^{n} A_j Z^j,$$

tem-se a relação:

$$A_j = (-1)^{n-j} a_j^2 + 2 \sum_{k=1}^{K} (-1)^k a_{j-k} a_{j+k},$$

onde $K = \min(j, n-j)$.

Se z_j são as raízes de $P_n(z)$, então as raízes Z_j de $B_n(Z)$ valem:

$$Z_j = z_j^2$$

Considerem-se agora as raízes escritas na forma polar:

$$z_j = \rho_j \exp(i\theta_j) , \qquad j = 1, \ldots, n,$$

onde $i = \sqrt{-1}$. Supõe-se também que estas raízes estão ordenadas em ordem decrescente de módulo, isto é:

$$\rho_1 > \rho_2 > \cdots > \rho_n .$$

Após m repetições do processo de elevação das raízes ao quadrado, tem-se as seguintes relações entre coeficientes e raízes:

$$z_1^{2^m} + z_2^{2^m} + \ldots z_n^{2^m} = - A_{n-1}/A_n ,$$

$$(z_1 z_2)^{2^m} + (z_1 z_3)^{2^m} + \ldots (z_{n-1} z_n)^{2^m} = A_{n-2}/A_n ,$$

$$\vdots$$

$$(z_1 z_2 \ldots z_n)^{2^m} = (-1)^n A_o/A_n ,$$

onde os A_j referem-se ao polinômio $B_n(Z)$, obtido na recorrência de ordem m. Então, pode-se escrever que:

$$z_1^{2^m} \left[1 + \sum_{j=2}^{n} (z_j/z_1)^{2^m} \right] = - A_{n-1}/A_n ,$$

e no limite para $m \to \infty$, tem-se:

$$|z_1| = \rho_1 = \lim_{m \to \infty} |-A_{n-1}/A_n|^{1/2^m} .$$

Usando o mesmo processo para a segunda relação, o valor aproximado resulta em:

$$\rho_1\rho_2 \cong |A_{n-2}/A_n|^{1/2^m},$$

de onde:

$$\rho_2 \cong |A_{n-2}/A_{n-1}|^{1/2^m}$$

e assim sucessivamente até:

$$\rho_n = |A_o/A_1|^{1/2^m} \quad .$$

Na prática para 3 ou 4 repetições do processo, as raízes já estão suficientemente separadas, podendo-se aplicar as relações acima.

Uma dificuldade tradicional do método do quadrado das raizes está na ocorrência de raízes múltiplas. Utilizando, entretanto, o artifício sugerido no item 4 da Seção 20.4.1, evita-se este problema.

Uma objeção a este método é que o valor dos A_j cresce exponencialmente, podendo sobrepujar a capacidade de representação dos registros da memória para m relativamente pequeno. Este problema é contornado dividindo todos os A_j por uma constante arbitrária a cada repetição do processo. Pode-se observar que este procedimento não afeta o cálculo das raízes.

Cumpre observar que o método de Graeffe deve ser encarado como um gerador de valores iniciais, a fim de serem usados em processos autocorretivos para obtenção das raízes. Neste aspecto, este método é análogo ao de Lehmer-Schur, possuindo contudo a vantagem de prover uma localização precária de todas as raízes, quando isto for de interesse.

Via de regra este método é usado apenas para fornecer valores iniciais para outros métodos; entretanto, os problemas relativos à sua precisão não serão discutidos.

20.4.3 - APROXIMAÇÕES SUCESSIVAS ENCADEADAS

Na resolução de equações polinomiais, como na resolução de qualquer equação, não se pode dispensar as vantagens dos métodos autocorretivos quando se requer precisão nos resultados.

Todos os métodos apresentados para equações transcendentais aplicam-se também a equações polinomiais. Discute-se aqui apenas o método de Newton a título de exemplo.

1) Método de Newton

A restrição para o uso mais difundido do método de Newton em equações transcendentais reside na dificuldade de calcular a derivada da função no ponto desejado. Esta dificuldade não existe no caso de equações algébricas, pois a derivada de um polinômio é facilmente determinada.

O método de Newton apresenta, neste caso, uma grande popularidade pela sua fácil implementação em computadores.

Estabelecido um valor inicial $(z_j)_o$, constrói-se o processo de recorrência pela relação:

$$(z_j)_{k+1} = (z_j)_k - \frac{P_n[(z_j)_k]}{P'_n[(z_j)_k]}, \qquad k = 1, 2, \ldots$$

Como esta equação envolve a variável complexa z, ela se dividirá em duas outras, uma para a parte real e outra para a parte imaginária.

Sendo conhecidas as expressões analíticas das derivadas $P_n'(z)$, pode-se facilmente determinar se o método converge ou não nas vizinhaças do ponto inicial considerado.

Os outros métodos apresentados para equações transcendentais podem ser adaptados do mesmo modo para equações algébricas.

20.4.4 - ALGORITMOS PARA DIVISÃO SINTÉTICA

Uma vez conhecida uma aproximação satisfatória para o valor z_j de uma raiz de $P_n(z)$, pode-se reduzir o grau do polinômio usando os algoritmos de divisão sintética.

Pela simplicidade de sua aplicação, esses algoritmos são usados conjuntamente com os métodos de aproximações sucessivas encadeadas.

Discute-se, a seguir, o desenvolvimento desses algoritmos.

1) Divisão sintética com fatores lineares

Seja $\tilde{z}_j$ o valor aproximado de uma raiz de $P_n(z) = 0$. Pode-se escrever que:

$$P_n(z) = (z - \tilde{z}_j)\, P_{n-1}(z) + R,$$

onde $P_{n-1}(z)$ é um polinômio de grau $n - 1$ com coeficientes $b_j, j = 1, \ldots, n - 1$, e o R é o resto da divisão de $P_n(z)$ por $(z - \tilde{z}_j)$.

A identidade acima fornece a relação:

$$P_n(z) = \sum_{k=0}^{n} a_k z^k = (z - \tilde{z}_j) \sum_{i=0}^{n-1} b_i z^i + R,$$

de onde $P_n(\tilde{z}_j) = R$, que desenvolvida produz:

$$\sum_{k=0}^{n} a_k z^k = \sum_{i=1}^{n} b_{i-1} z^i - \sum_{i=0}^{n-1} \tilde{z}_j b_i z^i + R \quad .$$

Igualando os coeficientes de mesma potência, tem-se:

$$a_n = b_{n-1} \quad ,$$

$$a_k = b_{k-1} - b_k \tilde{z}_j \quad , \qquad , k = 1, \ldots, n-1 \quad ,$$

$$a_o = b_o \tilde{z}_j + R \quad .$$

Assim o valor $R = P_n(\tilde{z}_j)$ pode ser encontrado pelas fórmulas de recorrência acima, que permitem o cálculo dos b_j a partir dos a_j. Este método é conhecido como divisão sintética (ou regra de Horner) e pode ser esquematizado por:

	a_n	a_{n-1}	a_{n-2}	a_1	a_o	$\tilde{z}_j$
+		$b_{n-1}\tilde{z}_j$	$b_{n-2}\tilde{z}_j$	$b_1 z_j$	$b_o \tilde{z}_j$	
	b_{n-1}	b_{n-2}	b_{n-3}	b_o	R	

onde os elementos da terceira linha são obtidos por adição dos elementos das duas primeiras linhas. Os elementos da segunda linha são obtidos da esquerda para a direita, pela multiplicação do termo precedente da terceira linha por $\tilde{z}_j$, começando-se o processo com $b_{n-1} = a_n$. Verifica-se que este algoritmo reproduz as relações de recorrência acima. Nesta forma, o algoritmo é conhecido como algoritmo de Briot-Ruffini.

Analogamente, pode-se escrever:

$$P_{n-1}(z) = (z - \tilde{z}_j)\, P_{n-2}(z) + R',$$

onde $P_{n-2}(z)$ é um polinômio de grau n - 2 com coeficientes C_i, i = 1, ..., n - 2. Aplicando as mesmas fórmulas de recorrência, obtêm-se:

$$b_{n-1} = C_{n-2} \ ,$$

$$b_k = C_{k-1} - C_k\, \tilde{z}_j \ , \qquad , k = 1, \ldots, n - 2 \ ,$$

$$b_o = - C_o\, \tilde{z}_j + R',$$

onde o algoritmo de Briot-Ruffini pode ser usado para calcular R'.

Por outro lado,

$$P_n(z) = (z - \tilde{z}_j)^2\, P_{n-2}(z) + R'\, (z - \tilde{z}_j) + R$$

e segue-se que:

$$R' = P_n'(\tilde{z}_j).$$

Quando se utiliza $\tilde{z}_1 = z_1 + \varepsilon$, onde ε é o erro tolerado, obtém-se uma redução no grau do polinômio e pode-se prosseguir na busca das raízes com o polinômio $P_{n-1}(z)$.

Outra aplicação do método de divisão sintética consiste em utilizar $\tilde{z}_j = (z_j)_k$, k = 1, 2, ... em conjução com o método de Newton. Nestas condições, o processo de recorrência torna-se:

$$(z_j)_{k+1} = (z_j)_k - R/R',$$

e a recorrência é interrompida quando $|R|$ se torna suficientemente pequeno.

2) Divisão sintética com fatores quadráticos

Neste caso, divide-se

$P_n(z)$ por $z^2 + \tilde{p}z + \tilde{q}$

e tem-se:

$$P_n(z) = (z^2 + \tilde{p}z + \tilde{q})P_{n-2}(z) + zR + S.$$

Usando o mesmo desenvolvimento anterior, as relações de recorrência resultam em:

$$a_n = b_{n-2} ,$$

$$a_{n-1} = b_{n-3} + \tilde{p}b_{n-2} ,$$

$$a_k = b_{k-2} + \tilde{p}b_{k-1} + \tilde{q}b_k , \qquad k = 2, \ldots, n-2 ,$$

$$a_1 = \tilde{p}b_o + \tilde{q}b_1 + R ,$$

$$a_o = \tilde{q}b_o + S,$$

que permitem calcular os valores dos b_k, $k = 0, \ldots, n - 2$ e os valores de R e S.

Também neste caso pode-se estabelecer uma forma esquemática para o algoritmo, como a mostrada a seguir.

a_n	a_{n-1}	a_{n-2}	.....		a_o	
	$-\tilde{p}b_{n-2}$	$-\tilde{p}b_{n-3}$	.....	$-\tilde{p}b_o$		$-\tilde{p}$
		$-\tilde{q}b_{n-2}$	.....		$-\tilde{q}b_o$	$-\tilde{q}$
b_{n-2}	b_{n-3}	b_{n-4}	.....	R	S	

Quando $\tilde{p} = p + \varepsilon_1$ e $\tilde{q} = q + \varepsilon_2$, onde p e q estabelecem o valor quadrático resultante do produto de duas raízes de $P_n(z)$, obtém-se um polinômio de grau n-2.

Analogamente, pode-se associar um método de aproximações sucessivas com a divisão sintética usando fatores quadráticos. Quando os coeficientes de $P_n(z)$ forem todos reais, obtém-se o método de Lin-Bairstow anulando R e S e calculando novos valores para $\tilde{p}$ e $\tilde{q}$.

Considerando $\tilde{p}_k$ e $\tilde{q}_k$ valores aproximados para p e q, obtém-se um refinamento desta solução fazendo R = 0 e S = 0. Os novos valores são então dados por:

$$\tilde{q}_{k+1} = \frac{a_o}{b_o} \qquad \text{e} \qquad \tilde{p}_{k+1} = \frac{a_1 - \tilde{q}_k b_1}{b_o} \quad .$$

O processo de recorrência é interrompido quando $|R|$ e $|S|$ são suficientemente pequenos.

EXERCÍCIOS

1. O cálculo da raiz da equação $f(x) = 0$ é prejudicado em computadores por truncamento e arrendondamento. Discuta o efeito do uso de uma função aproximadora $\bar{f}(x)$ no lugar de $f(x)$, para a determinação da raiz da equação $f(x) = 0$ (veja Young and Gregory, 1972).

2. O método do bisseccionamento não se aplica ao caso $f(x) = (x - 2)^2 = 0$, no intervalo (1, 3). Discuta a utilização do artifício do item 3 da Seção 20.2.1 e calcule o valor da raiz por esse método. Use $\varepsilon = 0{,}01$.

3. Calcule a raiz de $f(x) = 1 - \tan x = 0$ no intervalo $(0{,}3\ \pi/8)$, usando o método da falsa posição. Mostre que, neste caso, o valor de x_2 permanece constante e igual a $3\ \pi/8$ no decorrer do processo. Use $\varepsilon = 0{,}01$.

4. Repita os Exercícios 2 e 3 usando os métodos da secante e de Newton. Discuta a eficiência na obtenção do resultado.

5. Mostre que o cálculo da derivada, usado pelo método da secante, dispensa a utilização de dupla precisão, apesar de envolver diferenças de números próximos. (*Sugestão:* use o desenvolvimento de Taylor para $f(x)$ e verifique o cancelamento de efeitos da diferença, Δx, de números próximos).

6. Encontre uma expressão aproximada para $h(x)$ dos métodos iterativos com um ponto sem memória para que se tenha $\phi'(x) = 0$ e $\phi''(x) = 0$.

7. Discuta a convergência e estabilidade do método iterativo, determinado no Exercício 6.

8. Usando $h(x) = -1/f'(x)$ e a função $\phi_3(x)$, do exemplo do ítem 3 da Seção 20.2.3, resolva a equação do Exercício 3. Compare a eficiência dos dois métodos. Use $\varepsilon = 0{,}01$.

9. Repita o exercício anterior, sendo $\phi_3(x)$ gerada a partir da função:

$$\phi_1(x_i) = x_i - \frac{f(x_i)}{f[x_i, x_{i-1}]} \quad .$$

Compare a eficiência deste método com a do exercício anterior.

10. Discuta a possibilidade de generalizar o método do bisseccionamento para resolver sistemas de equações não-lineares.

11. Considere a função $T_m(z) = P_n(z)/Q_k(z)$, onde $Q_k(z)$ é o m.d.c. entre $P_n(z)$ e $P_n'(z)$. Mostre que:

a) $T_m(z)$ não possui raízes múltiplas,

b) $T_m(z)$ possui todas raízes distintas de $P_n(z)$.

12. Estabeleça um método para encontrar o valor de uma raiz real de $P_n(z) = 0$, com base na seqüência de Sturm e no bisseccionamento do intervalo dentro do qual foi localizada uma raiz.

13. Estabeleça um método para encontrar o valor de uma raiz complexa de $P_n(z) = 0$, com base no bisseccionamento de uma coroa circular do plano complexo. Use o critério de Lehmer Schur.

Uma vez determinado o raio no plano complexo, calcule o valor da raiz fazendo:

$$z = R_m \exp(i\theta)$$

e determine o valor de θ que satisfaz $P_n(z) = 0$.

14. Escreva um programa de computador para calcular o valor aproximado de todas as raízes de uma equação algébrica. Empregue:

a) o método do quadrado das raízes,

b) o artifício para eliminar raízes múltiplas (Exercício 11).

15. Compare a eficácia dos métodos dos dois últimos exercícios, como forma de prover valores iniciais para processos autocorretivos.

16. Mostre que, para o método de Newton, se a raiz (z_j) é complexa, o valor inicial $(z_j)_0$ não pode ser real quando os coeficientes do polinômio forem todos reais.

17. Discuta o resultado do exercício anterior para uma aplicação mais generalizada.

18. Escreva um programa de computador para o cálculo das raízes, pelo método de Newton, usando divisão sintética.

19. Discuta o problema da propagação de erros no processo de divisões sintéticas sucessivas com redução gradativa do grau do polinômio.

20. Escreva um programa de computador para o cálculo das raízes de um polinômio, utilizando divisão sintética em conjunção com o método de Newton.

21. Considerando pares de raízes escreva um programa de computador para implementar o método de Lin-Bairstow, utilizando os restos de divisão sintética.

BIBLIOGRAFIA

APOSTOL, T.M. *Mathematical Analysis.* Reading, MA, Addison-Wesley, 1957.

BEREZIN, I.S.; ZHIDKOV, N.P. *Computing Methods.* Reading, MA, Addison-Wesley, 1965. v. 2.

CARNAHAN, B.; LUTHER, H.A. and WILKES, J.O. *Applied Numerical Methods.* New York, John Wiley, 1969.

DICKSON, L.E. *New First Course in the Theory of Equations.* New York, John Wiley, 1939.

GANTMACHER, F.R. *The Theory of Matrices.* New York, Chelsea, 1959. v. 2.

HENRICI, P. *Applied and Computational Complex Analysis.* New York, John Wiley, 1974. v. 1.

HILDEBRAND, F.B. *Introduction to Numerical Analysis.* New York, McGraw-Hill, 1956.

KORGANOFF, A. *Methodes de Calcul Numérique.* Paris, Dunod, 1962. Tome 1.

LEHMER, D.H. *A Machine method for solving polynomial equations.* J. Assoc. Comput. Mach. 8,151-161, 1961.

RALSTON, A. *A first course in numerical analysis.* New York, McGraw-Hill, 1965.

RALSTON, A.; WILF, H.S. *Mathematical Methods for Digital Computer.* New York, John Wiley, 1967. v. 2.

SZIDAROVSZKY, F.; YAKOWITZ, S. *Principles and procedures of numerical analysis.* Plenum Press, New York.

TRAUB, J.F. *Iterative methods for the solution of equations.* New Jersey, Prentice-Hall, 1964.

YOUNG, D.M.; GREGORY, R.T. *A survey of numerical mathematics.* Reading, MA, Addison-Wesley.

CAPÍTULO 21

EQUAÇÕES LINEARES SIMULTÂNEAS E MATRIZES

21.1 - INTRODUÇÃO

Um conjunto de equações lineares simultâneas é representado por:

$$\sum_{j=1}^{N} a_{ij}\, x_j = v_i\,, \qquad i = 1, \ldots, M\,.$$

Considera-se apenas o caso de maior interesse prático em que M = N. As equações lineares podem ser escritas na forma matricial:

$$\underline{\underline{A}}\ \underline{x} = \underline{v}\,,$$

onde $\underline{\underline{A}}$ é a matriz dos coeficientes (a_{ij}), $\underline{x}^t = (x_1, \ldots, x_N)$ o vetor das incógnitas e $\underline{v}^t = (v_1, \ldots, v_N)$ o vetor constante. Esta última forma, $\underline{\underline{A}}\ \underline{x} = \underline{v}$, estabelece uma estreita ligação entre os problemas de álgebra matricial e os de solução de equações lineares.

O sistema de equações lineares pode ser considerado como uma função vetorial de variável vetorial na forma:

$$\underline{f}(\underline{x}) = \underline{\underline{A}}\ \underline{x} = \underline{v}\,,$$

cuja solução é escrita simbolicamente como:

$$\underline{x} = \underline{f}^{-1}(\underline{v}) = \underline{\underline{A}}^{-1}\ \underline{v}.$$

Conhecendo a inversa da matriz $\underline{\underline{A}}$, pode-se determinar a solução do sistema de equações lineares. Isto entretanto nem sempre

constitui a forma mais prática para resolver o problema. Por esta razão são apresentados primeiramente os métodos para solução de equações lineares e, em seguida, tratados os problemas de álgebra matricial.

A apresentação dos métodos segue a mesma estrutura do capítulo anterior, procurando primeiramente uma solução segura para o problema e posteriormente o aperfeiçoamento dessa solução.

As matrizes consideradas neste capítulo são todas da forma:

$$\underline{\underline{A}} = \sum_j \sum_i a_{ji}\, \hat{e}_i\, \hat{e}_j \quad ,$$

e os vetores definidos no espaço vetorial, caracterizado pelas propriedades $\hat{e}_i$.

Como o espaço vetorial no qual se trabalhará é único, pode-se simplificar a notação dispensando os $\hat{e}_i$, pois os índices, neste caso, caracterizam univocamente a propriedade associada. Entretanto, esta notação exige cuidados especiais no tratamento do produto matricial (ver Capítulo 5).

21.2 - SOLUÇÃO DE EQUAÇÕES LINEARES SIMULTÂNEAS

Enquanto para a solução de equações não-lineares não se dispõe de métodos simples diretos, no presente caso eles são bastante populares e aplicáveis na maioria dos casos práticos ($N < 30$). Para sistemas com grande número de equações, os métodos de aproximações sucessivas devem ser preferidos.

Os métodos numéricos para solução de equações lineares simultâneas são apresentados a seguir.

21.2.1 - MÉTODOS DIRETOS

Dentre os métodos diretos, os que mais se destacam podem ser englobados em dois tipos principais:

- métodos de eliminação,
- métodos de ortogonalização.

1) Métodos de eliminação

Dentre os métodos de eliminação, destaca-se o de Gauss em duas versões: uma mais apropriada para pequenos computadores e outra mais utilizada em grandes computadores. Estas formas são apresentadas a seguir sem os detalhes computacionais que se encontram muito bem documentados na bibliografia.

Considere-se o sistema de equações escrito na forma:

$$\sum_{j=1}^{N} a_{ij}\, x_j = v_i \qquad i = 1, \ldots, N.$$

Isolando a primeira incógnita na primeira equação resulta em:

$$x_1 = (v_1)_1 - \sum_{j=2}^{N} (a_{1j})_1\, x_j ,$$

onde:

$$(v_1)_1 = a_{11}^{-1}\, v_1 ,$$

$$(a_{1j})_1 = a_{11}^{-1}\, a_{1j} .$$

Eliminando-se essa incógnita nas equações seguintes, tem-se um sistema de N-1 equações com N-1 incógnitas, dado por:

$$\sum_{j=2}^{N} (a_{ij})_1 \, x_j = (v_i)_1 , \qquad i = 2, \ldots, N ,$$

onde:

$$\left.\begin{aligned} (v_i)_1 &= v_i - a_{i1}(v_1)_1 \\ (a_{ij})_1 &= a_{ij} - a_{i1}(a_{1j})_1 \end{aligned}\right\} \qquad i, j = 2, \ldots, N .$$

Em seguida, eliminando x_2 a partir da terceira equação, tem-se um sistema de N-2 equações com N-2 incógnitas. Procede-se assim sucessivamente até obter uma equação com uma incógnita.

Os valores do sistema reduzido pela eliminação da incógnita x_k são calculados recursivamente por:

$$(v_k)_k = (a_{kk})_{k-1}^{-1} \, (v_k)_{k-1} ,$$

$$(a_{kj})_k = (a_{kk})_{k-1}^{-1} \, (a_{kj})_{k-1} ,$$

$$(v_i)_k = (v_i)_{k-1} - (a_{ik})_{k-1} \, (v_k)_k ,$$

$$(a_{ij})_k = (a_{ij})_{k-1} - (a_{ik})_{k-1} \, (a_{kj})_k$$

para i, j=k+1, ..., N com:

$$\left.\begin{aligned} (v_j)_0 &= v_i \\ (a_{ij})_0 &= a_{ij} \end{aligned}\right\} \qquad i, j = 1, \ldots, N.$$

As incógnitas eliminadas formam um sistema triangular de equações:

$$\sum_{j=i}^{N} (a_{ij})_i \, x_j = (v_i)_i \, , \qquad i = 1, \ldots, N,$$

onde:

$$(v_i)_i = (a_{ii})_{i-1}^{-1} \, (v_i)_{i-1} \, ,$$

$$(a_{ij})_i = (a_{ii})_{i-1}^{-1} \, (a_{ij})_{i-1} \, ,$$

para i=1, ..., N; j=1, ..., N.

A solução desse sistema triangular de equações é obtida calculando sucessivamente as incógnitas, partindo-se da última equação para a primeira. Esse sistema é também chamado triangular superior em oposição ao sistema triangular de equações:

$$\sum_{j=1}^{i} b_{ij} \, y_j = v_i \, , \qquad i = 1, \ldots, N,$$

dito sistema triangular inferior de equações.

Na implementação do método de eliminação de Gauss em computadores, procura-se maximizar os valores de $(a_{ii})_{i-1}$ por conveniente mudança de linhas e colunas. Este processo pode afetar a ordem de obtenção das incógnitas, mas não o seu valor. O valor assim maximizado de $(a_{ii})_{i-1}$ é chamado "pivot". Detalhes sobre esta forma do método de eliminação podem ser encontrados, por exemplo, em: Ralston (1975), Pennington (1970), Bakhvalov (1977).

Pode-se mostrar (Ralston, 1965) que o método de eliminação de Gauss é equivalente a decompor o sistema original:

$$\underline{\underline{A}}\ \underline{x} = \underline{v}$$

em dois sistemas de equações triangulares: um triangular inferior dado por $\underline{\underline{B}}\ \underline{y} = \underline{v}$; outro triangular superior dado por $\underline{\underline{C}}\ \underline{x} = \underline{y}$. Daí decorre que $\underline{\underline{A}} = \underline{\underline{B}}\ \underline{\underline{C}}$. A matriz $\underline{\underline{B}}$ possui como característica adicional elementos unitários na diagonal.

Esta forma triangular é mais apropriada para utilização em computadores. O problema resume-se na determinação dos b_{ij} e c_{ij}, $i, j = 1, \ldots, N$.

Para determinar os elementos das matrizes $\underline{\underline{B}}$ e $\underline{\underline{C}}$, usa-se a equação $\underline{\underline{A}} = \underline{\underline{B}}\ \underline{\underline{C}}$, obtendo-se:

$$a_{ij} = \sum_{k=1}^{K} b_{ik}\, c_{kj},$$

onde $K = \min(i,j)$. Adicionalmente tem-se a propriedade $b_{ii} = 1$, $i = 1, \ldots, N$.

O processo de determinação dos b_{ij} e c_{ij} é feito recursivamente considerando cada coluna da matriz $\underline{\underline{A}}$. Assim, tomando a primeira coluna, obtêm-se:

$$a_{11} = c_{11},$$

$$a_{i1} = b_{i1}\, c_{11}, \qquad i = 2, \ldots, N,$$

pois $b_{11} = 1$. Do mesmo modo, tomando a segunda coluna tem-se:

$$a_{12} = c_{12},$$

$$a_{22} = b_{21}\, c_{12} + C_{22}\,,$$

$$a_{i3} = b_{i1}\, c_{12} + b_{i2}\, c_{22}\,, \qquad i = 3, \ldots, N$$

e assim por diante. Genericamente tem-se:

$$c_{ij} = a_{ij} - \sum_{k=1}^{i-1} b_{ik}\, c_{kj}\,, \qquad i \leq j\,,$$

$$b_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} b_{ik}\, c_{kj}}{c_{jj}}\,, \qquad i > j\,.$$

A utilização de "pivots" também é possível neste caso visando maximizar c_{jj}, o que é conseguido maximizando a_{jj}. A troca de linhas e colunas pode ser feita antes de iniciar o processo para cálculo dos b_{ij} e c_{ij}. Outra alternativa consiste somente na troca de linhas durante o decorrer dos cálculos (Ralston, 1965).

Uma variante do método de Gauss, conhecida como método de Gauss-Jordan, consiste em eliminar em todas as outras equações a variável isolada em cada equação escolhida. Este procedimento produz um sistema diagonal de equações, mas o número de operações envolvidas acarreta um aumento de erros computacionais. O método pode, entretanto, ser usado em conjunção com a aritmética de resíduos, que será apresentada posteriormente neste capítulo.

Os métodos diretos produzem resultados satisfatórios quando a solução, $\underline{x} = \underline{\underline{A}}^{-1}\, \underline{v}$, satisfaz a condição de Lipschitz:

$$|\Delta \underline{x}| < M\, |\Delta \underline{v}|\,,$$

onde M é um número que depende de $\underline{v}$ e $\Delta\underline{v}$. Quanto maior o valor de M, mais sensível será a solução aos efeitos de arredondamento.

Em geral, os métodos diretos são usados para produzir uma aproximação inicial para utilização em métodos autocorretivos. Para M < 1 e um pequeno número N de equações, pode-se utilizar a solução dos métodos diretos sem refinamento. O aumento do número de equações produz um aumento em $|\Delta\underline{v}|$, que, por vezes, torna os métodos diretos inadequados.

2) Métodos de ortogonalização

O sistema de equações $\underline{\underline{A}}\,\underline{x} = \underline{v}$ pode ser reescrito na forma:

$$\underline{\underline{G}}\,\underline{y} = \underline{0} \ ,$$

onde $\underline{\underline{G}}$ é uma matriz N x (N+1) e $\underline{y}$ é um vetor com N+1 componentes.

Usando a matriz $\underline{\underline{G}}$ decomposta em vetores, pode-se escrever que:

$$\begin{bmatrix} \underline{g}_1^t \\ \vdots \\ \underline{g}_N^t \end{bmatrix} \underline{y} = \underline{0} \ ,$$

onde se identificam os vetores componentes como:

$$\underline{g}_i^t = (a_{i1}, \ldots, a_{i,N}, -v_i),$$

$$\underline{y}^t = (x_1, \ldots, x_N, 1).$$

Assim, a solução do sistema de equações reduz-se ao problema de encontrar um vetor $\underline{y}$ ortogonal a N vetores $\underline{g}_i$, i=1, ..., N.

Os vetores $\underline{g}_i$ formam uma base do espaço N-dimensional. Para obter uma base do espaço N+1 dimensional, completa-se o conjunto com o vetor:

$$\underline{g}^t_{N+1} = (0, \ldots, 0,1)$$

e pode-se mostrar que o conjunto assim formado é linearmente independente.

Procede-se em seguida ao processo de ortonormalização construindo, a partir dos vetores $\underline{g}_i$, um conjunto de vetores ortonormais $\underline{u}_1, \ldots, \underline{u}_{N+1}$. Este processo é desenvolvido recursivamente pelas expressões:

$$\underline{u}'_i = \underline{g}_i - \sum_{k=1}^{j-1} < \underline{g}_i, \ \underline{u}_k > \underline{u}_k ,$$

$$\underline{u}_i = \underline{u}'_i / |\underline{u}'_i| , \qquad i = 1, \ldots, N+1 ,$$

onde na elaboração de cada novo vetor é subtraída a projeção dos vetores já construídos, tornando-o ortogonal a todos os precedentes.

Considere-se o último vetor construído da seguinte forma:

$$\underline{u}_{N+1} = (z_1, \ldots, z_{N+1}).$$

Este vetor é ortogonal a todos os vetores g_i, i=1, ..., N e assim obtém-se o vetor $\underline{y}$ fazendo:

$$\underline{y} = \frac{1}{z_{N+1}} \underline{u}_{N+1} \ .$$

Este método de ortogonalização envolve menos operações que o método de Gauss, mas é menos preciso.

Uma possível modificação consiste em ordenar os vetores $\underline{g}_i$, i=1, ..., N em ordem crescente de módulo, antes de iniciar o processo de ortogonalização. De qualquer modo este método é recomendado para obtenção de uma solução inicial a ser aperfeiçoada.

21.2.2 - APROXIMAÇÕES SUCESSIVAS ENCADEADAS

Na solução de equações lineares simultâneas também não se podem dispensar as vantagens oferecidas pelos métodos autocorretivos. Os métodos iterativos ocupam, neste particular, uma posição de destaque pela simplicidade de implementação em computadores. Os métodos de relaxação não são competitivos na maioria dos casos.

1) Métodos de relaxação

O sistema de equações lineares é colocado na forma:

$$\underline{\psi}(\underline{x}) = \underline{\underline{A}}\ \underline{x} - \underline{v} = 0 \ ,$$

e o processo de recorrência é formado por:

$$\underline{x}_{k+1} = \underline{x}_k - \underline{\underline{A}}^{-1}(\underline{\underline{A}}\ \underline{x}_k - \underline{v}).$$

Chamando $\underline{c}_k = \underline{\underline{A}}\ \underline{x}_k - \underline{v}$, o termo corretivo será dado por $\underline{\underline{A}}^{-1}\ \underline{c}_k$, o que exige a solução de um sistema de equações lineares (por métodos diretos) a cada repetição do processo.

Para contornar a dificuldade de solução de um sistema de equações, a cada passo prefere-se considerar apenas o elemento domi

nante da diferença $\underline{\underline{A}}\,\underline{x}_k - \underline{v}$. Seja i o índice da componente de maior valor. A correção simplifica-se então (Ralston, 1965) para:

$$(x_j)_{k+1} = (x_j)_k - \frac{(c_i)_k}{a_{ij}} ,$$

onde a_{ij} é a componente de maior valor absoluto na linha i da matriz $\underline{\underline{A}}$. A relaxação nesta forma simplificada é consideravelmente mais lenta do que no processo anterior. Em termos de quantidade de operações, entretanto, pode-se esperar uma certa compensação. No caso de matrizes esparsas, esta variante do método de relaxação pode eventualmente tornar-se competitiva.

O problema de convergência já foi discutido no Capítulo 6 para o caso geral. Na variante simplificada pode-se mostrar que o processo converge quando:

$$\left| 1 - \frac{(c_i)_k}{a_{ij}(x_j)_k} \right| < 1 .$$

Em geral, o processo de relaxação é estável quando se usa uma boa aproximação inicial (ver Capítulo 6).

2) Métodos iterativos

Os métodos iterativos para solução de equações de uma variável já foram tratados no capítulo anterior. A sua extensão para o caso de sistemas lineares é possível na maioria dos casos.

Fundamentalmente (ver Capítulo 6), os métodos iterativos consistem em transformar a equação:

$$\underline{f}(\underline{x}) = \underline{\underline{A}}\,\underline{x} - \underline{v} = \underline{0}$$

numa equação da forma:

$$\underline{x} = \underline{\phi}(\underline{x}).$$

Apresentam-se, aqui, as formas mais simples para essa transformação.

a) Métodos de Iteração Simples

Equivalente à forma unidimensional (ver Capítulo 20):

$$x = x + h(x)\, f(x),$$

tem-se a forma vetorial:

$$\underline{x} = \underline{x} - \underline{\underline{G}}(\underline{\underline{A}}\,\underline{x} - \underline{v}),$$

onde $\underline{\underline{G}}$ é uma matriz de determinante não-nulo. No caso mais simples $\underline{\underline{G}} = \underline{\underline{I}}$, que resulta na forma:

$$\underline{x} = (\underline{\underline{I}} + \underline{\underline{A}})\underline{x} - \underline{v},$$

onde $\underline{\underline{I}}$ é a matriz identidade. O processo de recorrência é então gerado por:

$$\underline{x}_{k+1} = (\underline{\underline{I}} + \underline{\underline{A}})\underline{x}_k - \underline{v}.$$

Quando $\underline{\underline{G}} = \underline{\underline{A}}^{-1}$ obtém-se o método de relaxação.

b) Método de Jacobi e Variante de Gauss-Seidel

O método de Jacobi consiste em decompor a matriz $\underline{\underline{A}}$ na soma de 3 outras matrizes:

$$\underline{\underline{A}} = \underline{\underline{D}} + \underline{\underline{B}} + \underline{\underline{C}},$$

onde:

- $\underline{\underline{D}}$ é uma matriz diagonal,

- $\underline{\underline{B}}$ é uma matriz triangular inferior com elementos da diagonal nulos,

- $\underline{\underline{C}}$ é uma matriz triangular superior com elementos da diagonal nulos.

O sistema de equações torna-se então:

$$\underline{\underline{D}}\,\underline{x} = -(\underline{\underline{B}} + \underline{\underline{C}})\underline{x} + \underline{v},$$

de onde a fórmula de recorrência:

$$\underline{x}_{k+1} = -\underline{\underline{D}}^{-1}(\underline{\underline{B}} + \underline{\underline{C}})\underline{x}_k + \underline{\underline{D}}^{-1}\,\underline{v}.$$

Em geral, o método de Jacobi colocado nessa forma é pouco usado, pois a implementação conhecida como método de Gauss-Seidel lhe é superior.

O método de Gauss-Seidel usa essencialmente a mesma fórmula de recorrência do método de Jacobi, mas utiliza o valor refinado das componentes já calculadas. Assim, a expressão de recorrência para o cálculo de cada componente toma a forma:

$$(x_i)_{k+1} = -\frac{1}{a_{ii}}\left[\sum_{j=1}^{i-1} a_{ij}(x_j)_{k+1} + \sum_{j=i+1}^{N} a_{ij}(x_j)_k - v_i\right], \quad i = 1,\ldots,N,$$

o que pode ser escrito em forma vetorial como:

$$\underline{x}_{k+1} = -\underline{\underline{D}}^{-1}\,[\underline{\underline{B}}\,\underline{x}_{k+1} + \underline{\underline{C}}\,\underline{x}_k] + \underline{\underline{D}}^{-1}\,\underline{v}.$$

Rearranjando esta expressão obtém-se:

$$\underline{x}_{k+1} = -(\underline{\underline{D}}+\underline{\underline{B}})^{-1}\ \underline{\underline{C}}\ \underline{x}_k + (\underline{\underline{D}}+\underline{\underline{B}})^{-1}\ \underline{v}\ .$$

A condição de convergência estrita local requer que (Capítulo 6):

$$\left|-(\underline{\underline{D}}+\underline{\underline{B}})^{-1}\ \underline{\underline{C}}\ \Delta\underline{x}_k\right| < \left|\underline{\underline{I}}\ \Delta\underline{x}_k\right|\ .$$

Assim, para que haja convergência, todos os autovalores λ_i da matriz:

$$\underline{\underline{Q}} = -(\underline{\underline{D}}+\underline{\underline{B}})^{-1}\ \underline{\underline{C}}$$

devem satisfazer as relações:

$$|\lambda_i| < 1\ , \qquad\qquad i = 1, \ldots, N\ .$$

Quando a matriz $\underline{\underline{A}}$ é real e simétrica, pode-se provar (Berezin and Zhidkov, 1965; Ralston, 1965) que o método de Gauss-Seidel converge independentemente do valor inicial.

O efeito de erros de arredondamento (estabilidade do método) pode ser examinado particularizando os resultados do Capítulo 6. No presente caso, a condição de estabilidade é idêntica à de convergência, ou seja, todos os autovalores tem valor absoluto menor que 1.

c) Método de Sobre-relaxação

O método de sobre-relaxação consiste em multiplicar o sistema original pela matriz $\underline{\underline{G}} = \omega\underline{\underline{D}}^{-1}$ e, em seguida, utilizar o método de iteração simples com o esquema do método de Gauss-Seidel. Assim resulta em:

$$\underline{x}_{k+1} = (1-\omega)\underline{x}_k - \omega(\underline{\underline{B}}\,\underline{x}_{k+1} + \underline{\underline{C}}\,\underline{x}_k) + \omega\underline{\underline{D}}^{-1}\,\underline{v}\,,$$

pois

$$\underline{\underline{D}}^{-1}\,\underline{\underline{B}} = \underline{\underline{B}} \text{ e } \underline{\underline{D}}^{-1}\,\underline{\underline{C}} = \underline{\underline{C}}\,.$$

Rearranjando os termos tem-se:

$$\underline{x}_{k+1} = (\underline{\underline{I}} + \underline{\underline{B}})^{-1}\,[(1-\omega)\underline{\underline{I}} - \omega\underline{\underline{C}}]\,\underline{x}_k + \omega\underline{\underline{D}}^{-1}\,\underline{v}\,.$$

Variando o valor de ω consegue-se alterar os autovalores da matriz:

$$\underline{\underline{Q}} = (\underline{\underline{I}} + \omega\underline{\underline{B}})^{-1}\,[(1-\omega)\underline{\underline{I}} - \omega\underline{\underline{C}}]$$

para que a convergência seja possível. Para aumentar a rapidez de convergência, escolhe-se o valor de ω de modo que os autovalores sejam mínimos.

Para $\omega = 0$ há uma degeneração do esquema iterativo. Para $\omega=1$ obtém-se o método de Gauss-Seidel com as equações normalizadas em relação aos elementos da diagonal principal. Prova-se (Young and Gregory, 1973) que a convergência só é obtida para $0 < \omega < 2$. Em geral, a melhoria de convergência, em relação ao método de Gauss-Seidel, é obtida para $\omega > 1$ (Young and Gregory, 1973).

3) Métodos baseados em minimização de funções

Da mesma forma que para solução de sistemas de equações não-lineares, podem-se usar, na solução de equações lineares, métodos baseados na minimização de funções escalares da variável vetorial.

Uma possível função escalar é:

$$g(\underline{x}) = |\underline{\underline{A}}\,\underline{x} - \underline{v}|^2\,,$$

a qual se pode aplicar o método do Gradiente (Seção 20.3.1).

Os métodos baseados em minimização de funções, no presente caso, não são competitivos em relação aos métodos já discutidos. Por esta razão eles não serão aqui discutidos com maiores detalhes. O leitor interessado pode facilmente implementar o método do Gradiente (Seção 20.3.1) para o caso de equações lineares.

21.2.3 - MÉTODOS BASEADOS NA ARITMÉTICA DE RESÍDUOS

Os métodos diretos apresentados anteriormente são seriamente afetados por erros de arredondamento, principalmente com o aumento do número de equações. Os métodos auto-corretivos podem exigir condições muito severas para a convergência, não sendo aplicáveis nos casos mais críticos.

A utilização da aritmética de resíduos na solução de sistemas de equações lineares veio complementar o conjunto de métodos com as seguintes vantagens adicionais:

a) todas as operações são executadas "exatamente", pois trabalha-se apenas com números inteiros;

b) a utilização de resíduos aritméticos evita o problema do resultado das operações cair fora do domínio condicional dos números inteiros no computador.

Assim, mesmo para os problemas mais críticos, é possível obter uma solução confiável usando a aritmética de resíduos.

1) Operações básicas utilizando resíduos

Definição 1 - Chama-se resíduo, $r_m(x)$, ao resto inteiro da divisão de x por m, ambos inteiros positivos. Pode-se escrever que:

$$r_m(x) = x - q\,m, \quad x \geq 0,$$

onde q é o quociente da divisão de x por m.

O resíduo assim definido é chamado resíduo de x no módulo m e, como conseqüência, tem-se:

$$0 \leq r_m(x) < m .$$

Para englobar os números negativos, a definição é estendida admitindo que o quociente seja negativo e considerando:

$$r_m(x) = x - q\,m, \quad x < 0 ,$$

com valores de q, de tal modo que se obtenha:

$$0 \leq r_m(x) < m .$$

Decorrem dessas definições algumas propriedades, tais como:

a) $r_m(k\,m) = 0$, k inteiro,

b) $r_m(x) = x$, $0 \leq x < m$,

c) $r_m[x \pm y] = r_m\,[r_m(x) \pm r_m(y)]$,

d) $r_m[x\,y] = r_m[r_m(x)\;r_m(y)]$.

Quando se tem $r_m[x\,y] = 1$, diz-se que x é o inverso de y no módulo m, se $0 \leq x \leq m$. Representa-se este inverso multiplicativo por $x = y_m^{-1}$.

Nota-se que não existe uma correspondência biunívoca entre o resíduo de um número e o próprio número. Na utilização da aritmética de resíduos, deve-se pois tomar cuidado para que esta perda de informação possa ser recuperada.

Pode-se constatar que, se o valor de m é maior que as quantidades envolvidas nas operações com inteiros positivos, nenhuma informação é perdida, pois o resíduo coincide com o próprio número utilizado. Em compensação toda simplicidade da aritmética de resíduos é perdida, pois neste caso trabalha-se com os próprios números ao invés de seus resíduos.

A seguir, apresenta-se uma forma para a recuperação de informação quando se trabalha com resíduos.

2) Recuperação de informação quando se utilizam resíduos

A recuperação de informações quando são utilizados resíduos baseia-se em dois teoremas fundamentais. O primeiro teorema estabelece o mínimo valor do módulo m para que não haja perda de informação. O segundo mostra uma forma para decompor o módulo m em um conjunto de módulos menores, a fim de que as vantagens da simplicidade da aritmética de resíduos possa ser utilizada.

Teorema 21.1 - Se o módulo m é escolhido de forma que

$$m > 2|x|$$

e se é construída uma quantidade $\tilde{x}$ a partir de $r_m(x)$ com as características:

a) $r_m(\tilde{x}) = r_m(x)$,

b) $|\tilde{x}| < m/2$,

então:

$$\overline{x} = x .$$

Prova: Sendo os resíduos idênticos, pode-se escrever que:

$$\bar{x} - x = m\,q\,,$$

onde q é um número inteiro; logo:

$$|mq| = |\tilde{x} - x|\,,$$

de onde se segue que:

$$m|q| \leq |\tilde{x}| + |x| < m/2 + m/2,$$

pois m é positivo. Assim resulta em:

$$m|q| < m$$

que é satisfeita por um único valor inteiro, q=0. De onde se conclui que $\tilde{x} = x$.

Este teorema, como se pode notar, mostra que a informação só não é perdida quando $m > |x|$, mas, como discutido anteriormente, não apresenta valor prático. O teorema que se segue é de capital importância para o problema de recuperação de informação quando se utilizam módulos $m_i < |x|$.

Teorema 21.2 - (Teorema Chinês dos resíduos).

Sejam m_i, $i = 1, \ldots, n$ vários módulos escolhidos com as seguintes características:

a) quaisquer dois módulos distintos deste conjunto são primos entre si, ou seja, m.d.c. $(m_i, m_j) = 1$ para $i \neq j$;

b) o produto destes módulos reproduz um valor desejado, m, para o módulo com que se deseja trabalhar.

Nestas condições, cada um dos módulos m_i pode ser encarado como uma propriedade para caracterizar um vetor, cuja quantidade associada é o resíduo do número neste módulo. A projeção desses vetores, segundo a propriedade m, fornece o valor desejado do vetor com a propriedade m. Então, a quantidade associada a esta propriedade é dada por:

$$r_m(x) = r_m \left[\sum_{i=1}^{m} M_i \; r_{m_i}(x) \; r_{m_i}(M_i^{-1}) \right] ,$$

onde $M_i = m/m_i$.

Prova - Considere-se o cálculo do resíduo do número $x > 0$ no módulo m_i.

$$r_{m_i}(x) = x - q_i m_i , \qquad i = 1, \ldots, n .$$

Este cálculo não é alterado se cada equação for reescrita na forma:

$$M_i \; r_{m_i}(x) \; r_{m_i}(M_i^{-1}) = x - qM_i \; m_i \; r_{m_i}(M_i^{-1}) ,$$

o que pode ser verificado calculando o resíduo de ambos os lados da equação no módulo m_i. Como $M_i . m_i = m$, a projeção da componente de m_i, segundo x, pode ser obtida calculando o resíduo no módulo m de ambos os lados da equação escrita nessa última forma. Obtém-se assim:

$$r_m[M_i \; r_{m_i}(x) \; r_{m_i}(M_i^{-1})] = [r_m(x)]_i , \qquad i = 1, \ldots, N .$$

Adicionando todas as projeções tem-se:

$$r_m(x) = r_m \left[\sum_{i=1}^{n} M_i \; r_{m_i}(x) \; r_{m_i}(M_i^{-1}) \right] .$$

O mesmo desenvolvimento é usado para $x < 0$, obtendo-se o mesmo resultado.

Uma demonstração mais rigorosa do Teorema Chinês dos resíduos pode ser encontrada em Zamlutti (1982).

Resolvendo várias vezes o mesmo problema em módulos distintos, primos entre si, isto será equivalente, pelo segundo teorema, à resolução do problema num módulo m dado pelo produto dos módulos considerados. Pelo primeiro teorema, se m for duas vezes maior que o valor absoluto de todas as quantidades envolvidas nos processos aritméticos, então a solução x do problema coincide com o seu resíduo na base m, valor este calculado pelo segundo teorema.

3) Utilização de resíduos na resolução de sistemas de equações lineares

Uma vez que nas operações aritméticas que envolvem resíduos trabalha-se sempre com valores exatos de quantidades inteiras, os métodos diretos já apresentados podem ser usados sem restrições. Transformam-se, para isso, todas as equações envolvidas em equações de resíduos.

A maior dificuldade na utilização de resíduos está no cálculo do valor de m a ser utilizado. Este problema é discutido a seguir:

Considere-se o sistema de equações escrito na forma:

$$\underline{\underline{A}}\ \underline{x} = \underline{v},$$

cuja solução pode ser escrita como:

$$\underline{x} = \frac{1}{d}\ \underline{\underline{A}}^{adj}\ \underline{v} = \frac{1}{d}\ \underline{y},$$

onde:

d = determinante de $\underline{\underline{A}}$,

$\underline{\underline{A}}^{adj}$ = matriz adjunta de $\underline{\underline{A}}$.

Então existem duas quantidades básicas a serem consideradas na determinação do módulo m. Estas quantidades são d e $\max_{1\leq i\leq N}|y_i|$. Escolhe-se m de acordo com o Teorema 1 do item 2 da Seção 21.2.3.

$$m = 2 \max(|d|, \max_{1\leq i\leq N}|y_i|).$$

Os valores d e y não são facilmente calculados; então, deve-se obter uma estimativa para essas quantidades. A estimativa para d é fornecida pela desigualdade de Hadamard:

$$|d|^2 \leq \prod_{i=1}^{N} \sum_{j=1}^{N} a_{ij}^2,$$

que pode ser reescrita na forma:

$$|d| \leq \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2\right)^{1/2}.$$

É interessante notar que cada elemento a_{ij}^{adj} da adjunta da matriz $\underline{\underline{A}}$ também é limitado por este valor. Assim:

$$|a_{ij}^{adj}| < \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2\right)^{1/2},$$

então:

$$\max_{1\leq k\leq N} |y_k| < \sum_{k=1}^{N} \left[\prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2 \right)^{1/2} \right] |v_k|,$$

de onde se segue que:

$$\max_{1\leq k\leq N} |y_k| < \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2 \right)^{1/2} \sum_{k=1}^{N} |v_k|.$$

Pode-se notar que este segundo valor engloba o primeiro, de onde a estimativa para m é dada por:

$$m > 2 \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2 \right)^{1/2} \sum_{k=1}^{N} |v_k|.$$

Quando $\sum_{k=1}^{N} |v_k| \leq 1$, a estimativa de $|d|$ limita o valor de m que, neste caso, é dado por:

$$m > 2 \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2 \right)^{1/2}.$$

Em geral são escolhidos, sempre que possível, os n maiores números primos inferiores a m. No caso em que, para determinado módulo, o sistema de equações:

$$r_{m_j}(\underline{\underline{A}}\ \underline{x}) = r_{m_j}(\underline{v})$$

não admite solução, deve-se escolher outro valor para o módulo.

Os detalhes da implementação deste método, bem como os resultados numéricos obtidos, podem ser encontrados em Young e Gregory

(1973). Para ilustrar a utilização dos métodos diretos, em conjunção com a aritmética de resíduos, considere-se o sistema 2 x 2 de equações:

$$a_{11}\, x_1 + a_{12}\, x_2 = v_1,$$

$$a_{21}\, x_1 + a_{22}\, x_2 = v_2.$$

Para resolver esse sistema num módulo genérico m, pelo método de Gauss, deve-se inicialmente multiplicar a primeira equação pelo inverso multiplicativo de a_{11} no módulo escolhido. Isto produz:

$$(x_1)_m + [(a_{11})_m^{-1}(a_{12})_m]_m\, (x_2)_m = [(a_{11})_m^{-1}(v_1)_m]_m,$$

$$(a_{21})_m\, (x_1)_m + (a_{22})_m\, (x_2)_m = (v_2)_m.$$

Para simplificar, utilizando a notação:

$$(a_{12})'_m = [(a_{11})_m^{-1}\, (a_{12})_m]_m, \qquad (v_1)'_m = [(a_{11})_m^{-1}\, (v_1)_m]_m,$$

e eliminando a variável $(x_1)_m$ na segunda equação, obtêm-se:

$$(x_1)_m + (a_{12})'_m\ (x_2)_m = (v_1)'_m\ ,$$

$$\{(a_{22})_m - [(a_{21})_m\, (a_{12})'_m]_m\}_m\, (x_2)_m = \{(v_2)_m - [(a_{21})_m\, (v_1)'_m]_m\}_m.$$

Simplificando a notação por:

$$(a_{22})'_m = \{(a_{22})_m - [(a_{21})_m\, (a_{12})'_m]_m\}_m\ ,$$

$$(v_2)'_m = \{(v_2)_m - [(a_{21})_m\, (v_1)'_m]_m\}_m\ ,$$

e multiplicando a segunda equação pelo inverso multiplicativo de $(a_{22})'_m$, tem-se:

$$(x_1)_m + (a_{12})'_m\ (x_2)_m = (v_1)'_m,$$

$$(x_2)_m = \{[(a_{22})'_m]^{-1}_m\ (v_2)'_m\ \}_m.$$

Resolve-se o sistema triangular de equações substituindo, na primeira equação, o valor de $(x_2)_m$ obtido na segunda equação. Assim, segue se que:

$$(x_1)_m = \{(v_1)'_m - [(a_{12})'_m\ (x_2)_m]_m\}_m\ .$$

Variando o valor do módulo m obtém-se um conjunto de soluções para cada variável. Utilizando em seguida o Teorema Chinês dos resíduos pode-se recuperar cada variável isoladamente, completando desta forma a solução do sistema.

21.3 - ÁLGEBRA MATRICIAL

A álgebra matricial abrange uma grande quantidade de tópicos distintos. No intuito de condensar a matéria, são aqui apresentados somente os problemas de:

a) inversão de matrizes,

b) cálculo de determinantes,

c) cálculo de autovalores e autovetores.

21.3.1 - INVERSÃO DE MATRIZES

O problema de inversão de matrizes pode ser encarado como uma generalização da resolução de equações lineares simultâneas. Pa

ra isto, considere-se o conjunto de sistemas de equações expresso na forma:

$$\underline{\underline{A}}\ \underline{x}_k = \underline{u}_k, \qquad k = 1, \ldots, N,$$

onde $\underline{u}_k$ é um vetor cuja única componente não-nula é a k-ésima componente que tem valor 1. Este conjunto de sistemas pode ser escrito numa forma mais compacta como:

$$\underline{\underline{A}}\ \underline{\underline{X}} = \underline{\underline{I}}\ ,$$

onde $\underline{\underline{I}}$ é a matriz identidade de ordem N. A solução do problema assim colocado expressa-se como:

$$\underline{\underline{X}} = \underline{\underline{A}}^{-1}.$$

Os métodos já estudados podem então ser aplicados ao problema de inversão de matrizes utilizando a matriz $\underline{\underline{I}}$ no lugar do vetor $\underline{v}$. Entretanto, neste particular, alguns métodos mais apropriados foram desenvolvidos, entre os quais se destacam:

a) inversão por decomposição triangular,

b) inversão por partição,

c) inversão por destruição do grau.

1) Inversão por decomposição triangular

Neste caso a matriz $\underline{\underline{A}}$, a ser invertida, é decomposta no produto de uma matriz triangular inferior $\underline{\underline{B}}$ por uma matriz triangular superior $\underline{\underline{C}}$, como no item 1 da Seção 21.2.1. Assim, têm-se:

$$\underline{\underline{A}} = \underline{\underline{B}}\ \underline{\underline{C}}\ ,$$

$$\underline{\underline{A}}^{-1} = \underline{\underline{C}}^{-1}\,\underline{\underline{B}}^{-1},$$

e o problema reduz-se à inversão das matrizes $\underline{\underline{B}}$ e $\underline{\underline{C}}$.

Mostra-se que a inversa de uma matriz triangular é outra matriz triangular de mesma forma.

Sejam b_{ij} um elemento genérico da matriz $\underline{\underline{B}}$, e p_{ij} um elemento genérico da matriz B^{-1}. De onde se tem as relações:

$$p_{kk} = 1,$$

$$\sum_{i=j}^{k} p_{ki}\ b_{ij} = 1, \qquad j = 1, \ldots, k-1$$

que permitem a obtenção dos elementos p_{ki}, i=1, ..., k. O valor de k varia de 1 até N e nas relações acima, como no item 1 da Seção 21.2.1, usou-se $b_{kk} = 1$.

Analogamente obtém-se a matriz $\underline{\underline{C}}^{-1}$, devendo-se notar que neste caso os elementos c_{kk} da matriz $\underline{\underline{C}}$ não são necessariamente unitários. Assim, completa-se o processo de inversão.

2) Inversão por partição

Neste caso, a matriz $\underline{\underline{A}}$ é decomposta em matrizes menores por partição. Assim, escreve-se:

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{Q}}_1 & \underline{\underline{P}}_1 \\ \underline{\underline{P}}_2 & \underline{\underline{Q}}_2 \end{bmatrix}$$

onde $\underline{\underline{Q}}_1$ e $\underline{\underline{Q}}_2$ são matrizes quadradas e $\underline{\underline{P}}_1$ e $\underline{\underline{P}}_2$ matrizes retangulares.

A matriz inversa é escrita na forma:

$$\underline{\underline{A}}^{-1} = \begin{bmatrix} \underline{\underline{S}}_1 & \underline{\underline{T}}_1 \\ \underline{\underline{T}}_2 & \underline{\underline{S}}_2 \end{bmatrix}$$

onde $\underline{\underline{S}}_1$ e $\underline{\underline{S}}_2$ são matrizes quadradas e $\underline{\underline{T}}_1$ e $\underline{\underline{T}}_2$ matrizes retangulares. Como

$$\underline{\underline{A}}\ \underline{\underline{A}}^{-1} = \underline{\underline{I}} \ ,$$

tem-se as relações:

$$\underline{\underline{Q}}_1\ \underline{\underline{S}}_1 + \underline{\underline{P}}_1\ \underline{\underline{T}}_2 = \underline{\underline{I}} \ ,$$

$$\underline{\underline{P}}_2\ \underline{\underline{S}}_1 + \underline{\underline{Q}}_2\ \underline{\underline{T}}_2 = \underline{\underline{0}} \ ,$$

$$\underline{\underline{Q}}_1\ \underline{\underline{T}}_1 + \underline{\underline{P}}_1\ \underline{\underline{S}}_2 = \underline{\underline{0}} \ ,$$

$$\underline{\underline{P}}_2\ \underline{\underline{T}}_1 + \underline{\underline{Q}}_2\ \underline{\underline{S}}_2 = \underline{\underline{I}} \ ,$$

onde as matrizes identidades e nulas são de mesma ordem que as matrizes do lado esquerdo das igualdades acima.

Eliminando $\underline{\underline{T}}_1$ e $\underline{\underline{T}}_2$ no conjunto de relações, obtêm-se:

$$\underline{\underline{T}}_1 = -\underline{\underline{Q}}_1^{-1}\ \underline{\underline{P}}_1\ \underline{\underline{S}}_2 \ ,$$

$$\underline{\underline{T}}_2 = -\underline{\underline{Q}}_2^{-1}\ \underline{\underline{P}}_2\ \underline{\underline{S}}_1 \ ,$$

de onde:

$$(\underline{\underline{Q}}_1 - \underline{\underline{P}}_1 \underline{\underline{Q}}_2^{-1} \underline{\underline{P}}_2)\, \underline{\underline{S}}_1 = \underline{\underline{I}} \ ,$$

$$(\underline{\underline{Q}}_2 - \underline{\underline{P}}_2 \underline{\underline{Q}}_1^{-1} \underline{\underline{P}}_1)\, \underline{\underline{S}}_2 = \underline{\underline{I}} \ ,$$

e finalmente:

$$\underline{\underline{S}}_1 = (\underline{\underline{Q}}_1 - \underline{\underline{P}}_1 \underline{\underline{Q}}_2^{-1} \underline{\underline{P}}_2)^{-1} \ ,$$

$$\underline{\underline{S}}_2 = (\underline{\underline{Q}}_2 - \underline{\underline{P}}_2 \underline{\underline{Q}}_1^{-1} \underline{\underline{P}}_1)^{-1} \ .$$

Pode-se notar que são necessárias quatro inversões de matrizes para obter $\underline{\underline{A}}^{-1}$. Considerando, entretanto, as relações derivadas de

$$\underline{\underline{A}}^{-1}\, \underline{\underline{A}} = \underline{\underline{I}} \ ,$$

têm-se

$$\underline{\underline{T}}_2\, \underline{\underline{Q}}_1 + \underline{\underline{S}}_2\, \underline{\underline{P}}_2 = \underline{\underline{O}} \ ,$$

$$\underline{\underline{S}}_1\, \underline{\underline{P}}_1 + \underline{\underline{T}}_1\, \underline{\underline{Q}}_2 = \underline{\underline{O}} \ ,$$

que fornecem:

$$\underline{\underline{T}}_1 = -\underline{\underline{S}}_1\, \underline{\underline{P}}_1\, \underline{\underline{Q}}_2^{-1} \ ,$$

$$\underline{\underline{T}}_2 = -\underline{\underline{S}}_2\, \underline{\underline{P}}_2\, \underline{\underline{Q}}_1^{-1} \ .$$

Combinando essas relações com as anteriores que envolvem as matrizes identidade, obtêm-se finalmente:

$$\underline{\underline{S}}_1 = \underline{\underline{Q}}_1^{-1} - \underline{\underline{Q}}_1^{-1}\, \underline{\underline{P}}_1\, \underline{\underline{T}}_2 \ ,$$

$$\underline{\underline{T}}_2 = -\underline{\underline{S}}_2 \, \underline{\underline{P}}_2 \, \underline{\underline{Q}}_1^{-1} \, ,$$

$$\underline{\underline{S}}_2 = (\underline{\underline{Q}}_2 - \underline{\underline{P}}_2 \, \underline{\underline{Q}}_1^{-1} \, \underline{\underline{P}}_1)^{-1} \, .$$

que formam um conjunto, o qual requer apenas duas inversões de matrizes para sua solução.

O processo de partição apresenta a vantagem de reduzir os erros computacionais, pois transforma a inversão de uma matriz numa inversão de matrizes menores. Nas matrizes menores a serem invertidas, podem-se usar métodos autocorretivos para refinamento da solução.

3) Inversão por destruição do grau

Este método baseia-se no da seção anterior. Consideram-se as relações:

$$\underline{\underline{S}}_1 = (\underline{\underline{Q}}_1 - \underline{\underline{P}}_1 \, \underline{\underline{Q}}_2^{-1} \, \underline{\underline{P}}_2)^{-1} \, ,$$

$$\underline{\underline{S}}_1 = \underline{\underline{Q}}_1^{-1} - \underline{\underline{Q}}_1^{-1} \, \underline{\underline{P}}_1 \, \underline{\underline{T}}_2 \, ,$$

$$\underline{\underline{T}}_2 = -\underline{\underline{S}}_2 \, \underline{\underline{P}}_2 \, \underline{\underline{Q}}_1^{-1} \, ,$$

$$\underline{\underline{S}}_2 = (\underline{\underline{Q}}_2 - \underline{\underline{P}}_2 \, \underline{\underline{Q}}_1^{-1} \, \underline{\underline{P}}_1)^{-1} \, .$$

Tomando $\underline{\underline{Q}}_2$ como uma matriz 1 x 1, ou seja, um escalar $\underline{\underline{Q}}_2 = \alpha$, segue-se que $\underline{\underline{P}}_1$ e $\underline{\underline{P}}_2$ são vetores colocados, respectivamente, nas formas coluna e linha. Para designá-los usa-se a notação:

$$\underline{v} = \underline{\underline{P}}_1 \, ,$$

$$\underline{w}^t = \underline{\underline{P}}_2 \, .$$

Nestas condições, as quatro relações anteriores são combinadas, resultando na fórmula de inversão:

$$(\underline{\underline{Q}}_1 - \alpha^{-1}\,\underline{v}\,\underline{w}^t)^{-1} = Q_1^{-1} + \frac{\underline{\underline{Q}}_1^{-1}\,\underline{v}\,\underline{w}^t\,\underline{\underline{Q}}_1^{-1}}{(\alpha - \underline{w}^t\,\underline{\underline{Q}}_1^{-1}\,\underline{v})} .$$

Esta fórmula pode ser usada recursivamente. Assim, se:

$$\underline{\underline{M}} = \underline{\underline{D}} - \sum_{i=1}^{N} \alpha_i^{-1}\,\underline{v}_i\,\underline{w}_i^t ,$$

onde $\underline{\underline{D}}$ é uma matriz diagonal não-singular, é possível estabelecer um algoritmo para sua inversão. Para isto considere-se a soma parcial:

$$\underline{\underline{G}}_k = [\underline{\underline{D}} - \sum_{i=1}^{k} \alpha_i^{-1}\,\underline{v}_i\,\underline{w}_i^t]^{-1} ,$$

cuja recorrência é dada por:

$$\underline{\underline{G}}_k = [\underline{\underline{G}}_{k-1}^{-1} - \alpha_k^{-1}\,\underline{v}_k\,\underline{w}_k^t]^{-1} ,$$

que pela fórmula de inversão transforma-se em:

$$\underline{\underline{G}}_k = \underline{\underline{G}}_{k-1} + \frac{\underline{\underline{G}}_{k-1}\,\underline{v}_k\,\underline{w}_k^t\,\underline{\underline{G}}_{k-1}}{\alpha_k - \underline{w}_k^t\,G_{k-1}\,\underline{v}_k} .$$

Observa-se que a recorrência inicia com $\underline{\underline{G}}_0 = \underline{\underline{D}}^{-1}$ e termina com $\underline{\underline{G}}_N = \underline{\underline{M}}^{-1}$.

Para determinar uma decomposição basta considerar $\underline{w}_k = \underline{u}_k$, onde $\underline{u}_k$ é o vetor cuja única componente não-nula é a k-ésima que vale 1. Usando $\alpha_k = 1$, os vetores $\underline{v}_k$ são as colunas da matriz diferença:

$\underline{\underline{V}} = \underline{\underline{D}} - \underline{\underline{M}}$.

Uma forma alternativa para decomposição pode ser encontrada em Ralston e Wilf (1968).

A eficiência desse método depende da particular decomposição adotada, sendo mais recomendado nos casos em que a matriz a ser invertida difere pouco (por exemplo por uma coluna) de outra, cuja inversa já é conhecida.

21.3.2 - CÁLCULO DE DETERMINANTE

O método de eliminação para solução de equações lineares simultâneas fornece uma maneira para calcular o determinante da matriz dos coeficientes. De fato ao reduzir o sistema à forma triangular, o determinante da matriz dos coeficientes é dado pelo produto da diagonal principal da matriz na forma triangular.

Na decomposição da matriz dos coeficientes como produto de uma matriz triangular inferior $\underline{\underline{B}}$, com elementos unitários na diagonal, por uma matriz triangular superior $\underline{\underline{C}}$, o determinante é o produto dos elementos da diagonal principal de $\underline{\underline{C}}$.

21.3.3 - CÁLCULO DE AUTOVALORES

Nenhum dos métodos apresentados pode fornecer informações sobre os autovalores. Entretanto conhecidos os autovalores, podem-se determinar os autovetores usando os métodos já discutidos.

Portanto, a preocupação consiste no estabelecimento de métodos para determinação dos autovalores, pois, sendo estes conhecidos, os autovetores podem ser calculados pela resolução do sistema de equações:

$$(\underline{\underline{A}} - \lambda_i \underline{\underline{I}})\, \underline{v}_i = \underline{0}, \qquad i = 1, \ldots, N ,$$

onde λ_i são os autovalores e $\underline{v}_i$ são os autovetores, incógnitas dos sistemas.

Dada a importância do cálculo dos autovalores e autovetores, desenvolveram-se métodos mais apropriados para sua determinação do que a resolução da equação características e a solução do sistema homogêneo de equações.

No desenvolvimento dos métodos para o cálculo de autovalores consideram-se três etapas:

a) localização dos autovalores,

b) métodos para matrizes hermitianas,

c) métodos para matrizes não-hermitianas.

É importante ressaltar que não existem métodos diretos para obtenção dos autovalores, cujo problema é idêntico ao da solução de uma equação algébrica (a equação característica).

1) Localização dos autovalores

O teorema geral para localização dos autovalores de uma matriz complexa foi estabelecido por Gershgorin:

Teorema 21.3 - (Gershgorin)

a) Todos os autovalores de uma matriz $\underline{\underline{A}}$ pertencem à união dos discos circulares fechados, C_i, no plano complexo, com centros a_{ii} e raios dados por:

$$\rho_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} |a_{ij}|.$$

b) Se a união de m dos círculos C_i forma um domínio conexo, então este domínio contém exatamente m dos autovalores.

Prova:

a) Seja λ um autovalor de $\underline{\underline{A}}$. Então exite um vetor $\underline{v} \neq \underline{0}$, tal que:

$$\underline{\underline{A}}\ \underline{v} = \lambda\ \underline{v}.$$

Desenvolvendo esta expressão, obtém-se para as componentes:

$$(\lambda - a_{ii})\ v_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}\ v_j.$$

Seja v_p a componente de maior valor absoluto do vetor $\underline{v}$, então:

$$|\lambda - a_{pp}||v_p| = \sum_{\substack{j=1 \\ j\neq p}}^{N} |a_{pj}\ v_j|\ ,$$

de onde:

$$|\lambda - a_{pp}||v_p| \leq \sum_{\substack{j=1 \\ j\neq p}}^{N} |a_{pj}||v_p|$$

e finalmente:

$$|\lambda - a_{pp}| \leq \sum_{\substack{j=1 \\ j\neq p}}^{N} |a_{pj}|\ ,$$

logo λ pertence a C_p.

b) Decompõe-se a matriz $\underline{\underline{A}}$ na soma da matriz obtida da diagonal $\underline{\underline{D}}$ e de uma matriz $\underline{\underline{S}}$ com os elementos fora da diagonal principal. Constrói-se em seguida o polinômio $P(\lambda, t)$ pela relação:

$$P(\lambda,t) = \det (\lambda \underline{\underline{I}} - \underline{\underline{D}} - t\underline{\underline{S}}) = 0 .$$

Pode-se observar que, para $t = 1$, obtém-se a equação característica referente à matriz $\underline{\underline{A}}$.

Considerem-se as raízes de $P(\lambda,t)$ para $0 \leq t \leq 1$. Estas raízes encontram-se na união dos círculos com centros a_{ii} e raios $t\rho_i$, de acordo com a primeira parte do teorema.

Denomine-se C a componente conexa da união dos círculos C_i e C_ε que é esta mesma componente com os círculos acrescidos da quantidade ε. Escolhe-se esta quantidade suficientemente pequena para que C_ε não possua nenhum ponto em comum com $U_i\ C_i - C$. Para um ponto λ, no contorno de C_ε tem-se $|\lambda - a_{ii}| > \rho_i$, $i = 1, \ldots, N$. Então, λ não pode ser raiz de $P(\lambda,t)$ para $t\varepsilon[0,1]$. Neste intervalo, $P(\lambda,t)$ é analítica. Assim, pode-se calcular o número de raízes de $P(\lambda,t) = 0$ dentro da região C_ε utilizando a fórmula de Cauchy:

$$n = \frac{1}{2\pi} \oint_{C_\varepsilon} \frac{P'(\lambda,t)}{P(\lambda,t)}\, d\lambda .$$

Segue-se que n é contínuo para $t \varepsilon [0,1]$. Como para $t = 0$ existem exatamente m raízes em C, este será o número de raízes para $t = 1$ em C_ε. Fazendo $\varepsilon \to 0$ e $C_\varepsilon \to C$, existem em conseqüência m autovalores em C.

Corolário - Se um dos círculos C_k não possui nenhum ponto em comum com qualquer dos outros círculos, então ele contém exatamente um autovalor da matriz $\underline{\underline{A}}$.

A aplicação do teorema de Gershgorin permite estabelecer estimativas para os autovalores quando se deseja utilizar métodos iterativos.

2) Métodos para matrizes hermitianas

As matrizes hermitianas gozam da propriedade:

$$\underline{\underline{H}}^H = \underline{\underline{H}} \ ,$$

onde o expoente H indica transposta conjugada.

Para fins de cálculo de autovalores e autovetores, basta considerar o caso simplificado de matrizes reais e simétricas. Isto pode ser facilmente demonstrado lembrando que os autovalores de matrizes hermitianas são sempre reais. Nestas condições, decompondo a matriz hermitiana complexa, pode-se escrever:

$$\underline{\underline{H}} = \underline{\underline{G}}_s + i\underline{\underline{G}}_a \ ,$$

onde $\underline{\underline{G}}_s$ e $\underline{\underline{G}}_a$ são matrizes reais, sendo $\underline{\underline{G}}_s$ simétrica e $\underline{\underline{G}}_a$ anti-simétrica. Tomando o autovalor λ, o correspondente autovetor $\underline{v} + i\,\underline{w}$ é obtido da equação:

$$(\underline{\underline{G}}_s + i\,\underline{\underline{G}}_a)\,(\underline{v} + i\,\underline{w}) = \lambda(\underline{v} + i\,\underline{w}) \ ,$$

que pode ser reescrita na forma de um problema real e simétrico como:

$$\begin{bmatrix} G_s & -G_a \\ G_a & G_b \end{bmatrix} \begin{bmatrix} \underline{v} \\ \underline{w} \end{bmatrix} = \lambda \begin{bmatrix} \underline{v} \\ \underline{w} \end{bmatrix}$$

É importante notar que nesta segunda forma os autovalores têm suas multiplicidades dobradas.

No desenvolvimento dos métodos para matrizes simétricas, utilizam-se transformações de similaridade empregando matrizes ortogonais. Estas transformações não alteram os autovalores (Gantmacher, 1959). Assim, dada uma matriz $\underline{\underline{A}}$ pode-se transformá-la numa forma mais conveniente pela relação:

$$\underline{\underline{P}}^t \, \underline{\underline{A}} \, \underline{\underline{P}} = \underline{\underline{M}} \ ,$$

onde $\underline{\underline{P}}$ é uma matriz ortogonal.

As matrizes ortogonais têm a propriedade de coincidência da transposta com a inversa. Utilizando a fórmula de inversão do item 3 da Seção 21.3.1, pode-se verificar que matrizes da forma:

$$\underline{\underline{P}} = \underline{\underline{I}} - 2\underline{v} \, \underline{v}^t$$

possuem esta propriedade, sendo portanto ortogonais. Adicionalmente, $\underline{\underline{P}}^t = \underline{\underline{P}}$.

a) Método Givens-Householder

O método mais recomendado atualmente teve seu desenvolvimento relacionado com os nomes de Givens e Householder. Basicamente, o método utiliza transformações de similaridade para colocar a matriz dada numa forma tridiagonal e, a partir desta forma, calcular os autovalores.

O desenvolvimento do método requer preliminarmente a demonstração de 3 teoremas (Ralston and Wilf, 1968), que são apresentados a seguir.

Teorema 21.4 - Se as primeiras i componentes do vetor $\underline{v}$ são nulas, a matriz $\underline{\underline{P}}$ possui a forma:

$$\underline{\underline{P}} = \begin{pmatrix} \underline{\underline{I}}_i & \underline{\underline{0}}_{i,n-i} \\ \underline{\underline{0}}_{n-i,i} & \underline{\underline{V}}_{n-i} \end{pmatrix}$$

onde n é a dimensão da matriz $\underline{\underline{A}}$, $\underline{\underline{I}}_i$ é a matriz identidade de ordem i, $\underline{\underline{0}}_{i,n-1}$ é uma matriz nula com dimensão i x (n - i), e $\underline{\underline{V}}_{n-1}$ é a matriz gerada por:

$$\underline{\underline{V}}_{n-1} = \underline{\underline{I}}_{n-1} - 2\,\underline{w}\,\underline{w}^t ,$$

onde $\underline{w}$ é um vetor com n - i componentes, obtido pela supressão das i primeiras componentes nulas do vetor $\underline{v}$.

Prova - ela é imediata, bastando escrever o vetor $\underline{v}$ na forma:

$$\underline{v}^t = (0, \ldots, 0, v_{i+1}, \ldots, v_n) ,$$

de onde resulta a expressão desejada da matriz $\underline{\underline{P}}$ com:

$$\underline{w}^t = (v_{i+1}, \ldots, v_n) .$$

Corolário - Se i = 1, a primeira coluna da matriz $\underline{\underline{P}}$ é dada pelo vetor:

$$\underline{u}_1 = (1, 0, \ldots, 0)^t .$$

Teorema 21.5 - Sejam $\underline{x}$ e $\underline{z}$ dois vetores de mesmo módulo, mas com diferentes componentes. Então existe um vetor $\underline{v}$ de módulo unitário, tal que:

$$(\underline{\underline{I}} - 2\underline{v}\ \underline{v}^t)\underline{x} = \underline{z}$$

O vetor $\underline{v}$ que satisfaz esta equação é único (com exceção do sinal) e é dadó pór:

$$\underline{v} = \frac{\underline{x} - \underline{z}}{<(\underline{x} - \underline{z}),\ (\underline{x} - \underline{z})>} .$$

Prova - Como $\underline{x}^t\ \underline{x} = \underline{z}^t\ \underline{z}$ e $\underline{x}^t\ \underline{z} = \underline{z}^t\ \underline{x}$, tem-se:

$$\left[1 - \frac{2(\underline{x} - \underline{z})(\underline{x} - \underline{z})^t}{|x - z|^2}\right] \underline{x} = \underline{x} - \frac{2(\underline{x}^t\underline{x} - \underline{z}^t\underline{x})}{2(\underline{x}^t\underline{x} - \underline{z}^t\underline{x})}\ (\underline{x} - \underline{z}) = \underline{z}$$

que demonstra a existência do vetor $\underline{v}$ nas condições do teorema. Para provar a unicidade deste vetor, basta considerar a existência de outro vetor $\underline{v}_1$ nas mesmas condições. Então:

$$(\underline{\underline{I}} - 2\underline{v}\ \underline{v}^t)\underline{x} = (\underline{\underline{I}} - 2\underline{v}_1\ \underline{v}_1{}^t)\underline{x} ,$$

de onde:

$$\underline{v}(\underline{v}^t\ \underline{x}) = \underline{v}_1(\underline{v}_1{}^t\ \underline{x})$$

e como $\underline{x} \neq \underline{z}$, $\underline{v}^t\ \underline{x} \neq 0$, tem-se $\underline{v}_1 = \alpha\underline{v} = \pm\underline{v}$.

Teorema 21.6 - Num processo de transformações de similaridade encadeadas para tridiagonalização de uma matriz, a transformação de ordem $i+1$ não afeta os resultados já obtidos com as i primeiras transformações.

Prova - Após i transformações executadas sobre a matriz $\underline{\underline{A}}$, esta toma a forma:

$$A_i = \begin{pmatrix} \underline{\underline{T}}_{i+1} & \begin{pmatrix} \underline{\underline{0}}_{i,n-i-1} \\ \underline{b}^t \end{pmatrix} \\ (\underline{\underline{0}}_{n-i-1,i}\ \underline{b}) & \underline{\underline{R}}_{n-i-1} \end{pmatrix}$$

onde $\underline{\underline{T}}_{i+1}$ é uma matriz (i+1) x (i+1) tridiagonal, $\underline{b}$ é um vetor não-nulo, $\underline{\underline{0}}_{i,n-i-1}$ é uma matriz i x (n-i-1) nula e $\underline{\underline{R}}_{n-i-1}$ é uma matriz quadrada simétrica.

Na transformação de ordem i+1, utiliza-se um vetor $\underline{v}$ com as i+1 primeiras componentes nulas. Logo, a matriz $\underline{\underline{P}}$ possui estruturalmente a forma do Teorema 21.4. A transformação:

$$\underline{\underline{A}}_{i+1} = \underline{\underline{P}}\ \underline{\underline{A}}_i\ \underline{\underline{P}}$$

resulta então em:

$$\underline{\underline{A}}_{i+1} = \begin{pmatrix} \underline{\underline{T}}_{i+1} & \begin{pmatrix} \underline{\underline{0}}_{i,n-1} \\ \underline{d}^t \end{pmatrix} \\ (\underline{\underline{0}}_{n-i-1,i}\ \underline{d}) & (\underline{\underline{V}}\ \underline{\underline{R}}\ \underline{\underline{V}})_{n-i-1} \end{pmatrix}$$

onde $\underline{d} = \underline{\underline{V}}_{n-i-1}\ \underline{b}$ e $(\underline{\underline{V}}\ \underline{\underline{R}}\ \underline{\underline{V}})_{n-i-1}$ é a transformação de similaridade executada pela matriz $\underline{\underline{V}}_{n-i-1}$ sobre a matriz $\underline{\underline{R}}_{n-i-1}$.

Pode-se notar que a transformação executada sobre a matriz $\underline{\underline{A}}_i$ não afeta a forma obtida dos i primeiros resultados.

O processo para obtenção da forma tridiagonal baseia-se na anulação dos elementos da primeira linha e da primeira coluna fora desta estrutura. O Teorema 21.6 permite a utilização de recorrência no processo.

Considere-se o vetor $\underline{s} = (0\ \underline{w}^t)^t$ de módulo unitário e a transformação de similaridade executada pela matriz $\underline{\underline{P}} = \underline{\underline{I}} - 2\ \underline{s}\ \underline{s}^t$. Pelo corolário do Teorema 21.4, a primeira coluna da matriz $\underline{\underline{P}}$ é o vetor $\underline{u} = (1,0\ \ldots,0)^t$. Sendo $\underline{a}$ a primeira coluna da matriz $\underline{\underline{A}}$, e $\underline{b}$ a primeira coluna da matriz transformada, tem-se:

$$(\underline{\underline{I}} - 2\ \underline{s}\ \underline{s}^t)\underline{a} = \underline{b}\ ,$$

onde $\underline{b}$ toma a forma:

$$\underline{b} = (b_1,\ b_2,\ 0,\ \ldots,\ 0)^t\ ,$$

e o problema consiste na determinação do vetor $\underline{s}$.

Pelo Teorema 21.5, os vetores $\underline{b}$ e $\underline{a}$ possuem o mesmo módulo. Levando em conta a forma do vetor $\underline{s}$, obtém-se:

$$b_1 = a_{11}\ ,$$

de onde resulta em:

$$b_2^2 = \sum_{j=2}^{n} a_{j1}\ .$$

Usando estes resultados e o Teorema 21.5, obtém-se:

$$\underline{s}^t = \frac{(0, a_{21} \pm b_2,\ a_{31},\ \ldots,\ a_{n1})}{\langle(\underline{a} - \underline{b}),(\underline{a} - \underline{b})\rangle}\ .$$

Após essa transformação executada sobre a matriz $\underline{\underline{A}}$, esta adquire a forma:

$$\underline{\underline{A}}_2 = \begin{bmatrix} b_1 & \pm b_2 & 0 & \cdots & 0 \\ \pm b_2 & * & & \cdots & * \\ 0 & \vdots & & & \vdots \\ \vdots & & & & \\ 0 & * & & \cdots & * \end{bmatrix}$$

De acordo com o Teorema 21.6 pode-se trabalhar sobre a matriz reduzida, cujos elementos estão caracterizados por um * na forma acima, utilizando o mesmo processo. Calcula-se um vetor $\underline{s} = (0\ \underline{w}_2^t)^t$ para a matriz reduzida determinando a nova matriz $\underline{\underline{P}} = \underline{\underline{I}} - 2\ \underline{v}\ \underline{v}^t$, onde $\underline{v} = (0\ \underline{s}_2^t)^t = (0\ 0\ \underline{w}_2^t)^t$. Pela transformação, $\underline{\underline{P}}\ \underline{\underline{A}}_2\ \underline{\underline{P}}$, obtém-se uma matriz reduzida de dimensões (n-1) x (n-2). Esta recorrência é utilizada até a obtenção da matriz $\underline{\underline{A}}_{n-2}$ na forma tridiagonal.

Uma vez obtida a forma tridiagonal, pode-se utilizar o teorema de Gershgorin determinando os intervalos, no eixo real, dentro dos quais se encontram os autovalores:

$$(b_{11} - |b_{21}|;\ b_{11} + |b_{21}|)\ ,$$

$$(b_{1i} - |b_{2,i-1}| - |b_{2i}|;\ b_{1i} + |b_{2,i-1}| + |b_{2i}|)\ ,$$

$$i = 2, \ldots, n-1\ ,$$

$$(b_{1n} - |b_{2,n-1}|,\ b_{1n} + |b_{2,n-1}|)\ ,$$

onde b_{1k} e b_{2k} denotam os valores de b_1 e b_2 obtidos nas várias etapas do processo de tridiagonalização.

Estabelecida uma aproximação para um autovalor, podem-se usar métodos autocorretivos para refinamento da solução. Estes métodos serão discutidos posteriormente.

Outra abordagem consiste na utilização do polinômio característico da matriz na forma tridiagonal. Este método é discutido por exemplo em Ralston e Wilf (1968) e Young e Gregory (1973).

3) Método para matrizes não-hermitianas

Aplica-se este método para qualquer matriz complexa. Ele pode ser usado também para complementação do método de tridiagonalização visando um refinamento da solução. É conhecido na literatura como método QR, sendo mais apropriado quando a matriz se encontra numa forma particular, conhecida como forma de Hessenberg.

No desenvolvimento do método, são utilizadas matrizes unitárias Q que têm a propriedade:

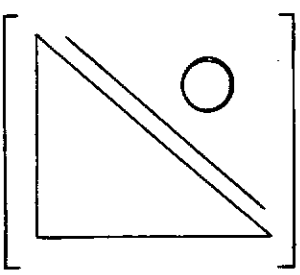
$$\underline{\underline{Q}}^H \underline{\underline{Q}} = \underline{\underline{I}} \quad ,$$

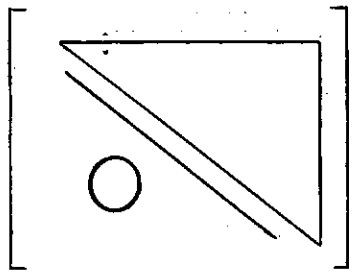
onde o expoente H indica a transposta conjugada da matriz (Gantmacher, 1959).

a) Redução de matrizes à forma de Hessenberg

A forma de Hessenberg é a forma triangular acrescida de mais uma diagonal. As matrizes $\underline{\underline{H}}_b$ e $\underline{\underline{H}}_c$ abaixo encontram-se, respectivamente, nas formas inferior e superior de Hessenberg.

$$\underline{\underline{H}}_b = \left[\quad \right]$$



$\underline{\underline{H}}_C =$ 

O processo para a redução de matrizes a essas formas consiste na utilização do método de eliminação de Gauss para anular os elementos externos ao triângulo e diagonal adicional. A seguir, apresenta-se um método para redução à forma superior de Hessenberg.

Na aplicação do método de eliminação de Gauss para obtenção da forma de Hessenberg alguns cuidados devem ser tomados, pois este método, na forma exposta no item 1 da Seção 21.2.1, modifica os autovalores. Para que isto não aconteça, o processo deve ser ligeiramente alterado a fim de construir uma transformação de similaridade.

Primeiramente, analisa-se o aspecto de mudança de linhas e colunas necessária para o correto posicionamento do "pivot". Para isto, considere-se a matriz:

$$\underline{\underline{U}}_{ji} = (\underline{u}_1, \ldots, \underline{u}_{i-1}, \underline{u}_j, \underline{u}_{i+1}, \ldots, \underline{u}_{j-1}, \underline{u}_i, \underline{u}_{j+1}, \ldots, \underline{u}_n)$$

para $j > i$, onde os vetores $\underline{u}_k$ possuem somente a k-ésima componente não-nula e com valor 1. Pode-se observar que na posição original $\underline{\underline{U}}_{ij} = \underline{\underline{I}}$. A matriz $\underline{\underline{U}}_{ji}$ tem as seguintes propriedades:

- $\det \underline{\underline{U}}_{ji} = -1$.

- $\underline{\underline{U}}_{ji}$ é auto-inversa, isto é: $\underline{\underline{U}}_{ji}\ \underline{\underline{U}}_{ij} \equiv \underline{\underline{I}}$.

- O produto $\underline{\underline{A}}\ \underline{\underline{U}}_{ji}$ troca as posições das colunas i e j da matriz $\underline{\underline{A}}$.

- $\underline{\underline{U}}^t_{ji} = \underline{\underline{U}}_{ji}$,

- O produto $\underline{\underline{U}}^t_{ji}\ \underline{\underline{A}}$ troca as posições das linhas i e j da matriz $\underline{\underline{A}}$.

O posicionamento do "pivot" deve então ser feito pela transformação:

$$\underline{\underline{X}}' = \underline{\underline{U}}_{ji}\ \underline{\underline{X}}\ \underline{\underline{U}}_{ji} \ ,$$

onde $\underline{\underline{X}}$ é a matriz sobre a qual se faz a modificação.

No processo para redução do problema à forma de Hessenberg, escolhe-se o maior elemento de cada coluna abaixo da diagonal principal. A transformação requerida obedece à forma:

$$\underline{\underline{X}}' = \underline{\underline{U}}_{i,k+1}\ \underline{\underline{X}}\ \underline{\underline{U}}_{i,k+1} \ ,$$

onde k é a coluna considerada e $i \geq k + 1$, sendo k obtido de:

$$|x_{ik}| = \max_{j>k} \{|x_{jk}|\} .$$

Assim, pode-se passar ao processo de eliminação propriamente dito. Para anular os elementos indesejáveis da coluna k, usa-se uma matriz semelhante à do Exercício 3 (apresentado no final deste capítulo) denotada por:

$$M_{k+1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & m_{k+2,\,k+1} & & \\ & m_{k+3,\,k+1} & & \\ & \vdots & & \\ & m_{n,k+i} & & 1 \end{bmatrix}$$

onde:

$$m_{j,k+1} = - \frac{x'_{jk}}{x'_{k+1,k}} ,$$

cujo produto $\underline{\underline{M}}_{k+1}\ \underline{\underline{X}}'$ produz a supressão dos elementos indesejados. A fim de não alterar os autovalores, completa-se a transformação de similaridade fazendo:

$$\underline{\underline{X}}'' = \underline{\underline{M}}_{k+1}\ \underline{\underline{X}}'\ \underline{\underline{M}}^{-1}_{k+1} .$$

Mostra-se que:

$$\underline{\underline{M}}^{-1}_{k+1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & - m_{k+2,\,k+1} & & \\ & - m_{k+3,\,k+1} & & \\ & \vdots & & \\ & - m_{n,k+1} & & 1 \end{bmatrix}$$

A redução do problema à forma de Hessenberg é então resumida nas equações:

$$\underline{\underline{A}}_1 = \underline{\underline{A}} ,$$

$$\underline{\underline{X}} = \underline{\underline{A}}_1 ,$$

$$\left.\begin{array}{l} \underline{\underline{X}}' = \underline{\underline{U}}_{i,k+1}\ \underline{\underline{X}}\ \underline{\underline{U}}_{i,k+1} \\ \underline{\underline{X}}'' = \underline{\underline{M}}_{k+1}\ \underline{\underline{X}}'\ \underline{\underline{M}}_{k+1}^{-1} \\ \underline{\underline{A}}_{k+1} = \underline{\underline{X}}'' \\ \underline{\underline{X}} = \underline{\underline{A}}_{k+1} \end{array}\right\} \quad k = 1, \ldots, n-2,$$

$$\underline{\underline{A}} = \underline{\underline{A}}_{n-1}$$

onde a relação final $\underline{\underline{A}} = \underline{\underline{A}}_{n-1}$ substitui a matriz $\underline{\underline{A}}$ na forma original pela sua equivalente na forma de Hessenberg.

b) Método QR

Este método baseia-se na construção iterativa de uma seqüência de matrizes $\underline{\underline{A}}_k$, $k = 0, 1, \ldots$ que obedecem as seguintes regras:

- $\underline{\underline{A}}_0 = \underline{\underline{A}}$,

  onde $\underline{\underline{A}}$ é a matriz da qual se desejam os autovalores.

- Para $k = 0, 1, \ldots$, calculam-se duas matrizes, uma unitária $\underline{\underline{Q}}_k$ e outra triangular superior, satisfazendo as relações:

  $$\underline{\underline{A}}_k - \alpha_k \underline{\underline{I}} = \underline{\underline{Q}}_k \underline{\underline{R}}_k ,$$

  onde os α_k constituem uma seqüência numérica.

- O processo de recorrência é estabelecido pela relação:

  $$\underline{\underline{A}}_{k+1} = \underline{\underline{R}}_k \underline{\underline{Q}}_k + \alpha_k \underline{\underline{I}} .$$

Como conseqüência, segue-se que a matriz $\underline{\underline{A}}_{k+1}$ é similar à matriz $\underline{\underline{A}}_k$. De fato, pelas relações acima, deduz-se que:

$$\underline{\underline{A}}_{k+1} = \underline{\underline{Q}}_k^H \, \underline{\underline{A}}_k \, \underline{\underline{Q}}_k$$

e, por recorrência, conclui-se que o processo gera uma seqüência de matrizes similares.

A existência da decomposição QR, para cada matriz $\underline{\underline{X}}$ da seqüência, é demonstrada pelo processo de construção apresentado a seguir.

c) Implementação do Método QR

Primeiramente, devem-se fazer algumas considerações:

- A matriz $\underline{\underline{A}}_0$ da seqüência está na forma superior de Hessenberg.

- Supõe-se que a forma de Hessenberg é preservada para cada matriz $\underline{\underline{X}}$ da seqüência (o que será justificado posteriormente).

Considere-se agora a matriz unitária $\underline{\underline{S}}_{pq}$ de rotação plana definida por:

$$\underline{\underline{S}}_{pq} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & x_{pp}/\delta & \cdots & x_{qp}/\delta & & & \leftarrow p \\ & & & & 1 & & & & \\ & & & \vdots & & \vdots & & & \\ & & & & & 1 & & & \\ & & & -x_{qp}/\delta & \cdots & x_{pp}/\delta & & & \leftarrow q \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \\ & & & \uparrow & & \uparrow & & & \\ & & & p & & q & & & \end{bmatrix}$$

onde $\delta = (|x_{pp}|^2 + |x_{qp}|^2)^{1/2}$.

O produto $\underline{\underline{S}}_{pq}\ \underline{\underline{X}}$ foi projetado de forma a anular o elemento p da linha q. Para verificar isto, considerem-se os elementos y_{ij} da matriz produto, dados por:

$$y_{ij} = x_{ij} , \qquad i \neq p, q ,$$

$$y_{pj} = x_{pj}\ x_{pp}/\delta + x_{qj}\ x_{qp}/\delta ,$$

$$y_{qj} = -x_{pj}\ x_{qp}/\delta + x_{qj}\ x_{pp}/\delta ,$$

para j = 1, ..., n. Pode-se notar que $y_{qp} = 0$.

Como a matriz $\underline{\underline{X}}$ está na forma de Hessenberg, pode-se verificar que, se os elementos y_{qp} de produtos sucessivos forem anulados na ordem (2,1), (3,2), ...,(n, n - 1), obtém-se uma matriz triangular, pois os elementos anulados num produto permanecem nulos nos produtos seguintes. Chamando $\underline{\underline{C}}$ esta matriz triangular, pode-se escrever que:

$$\underline{\underline{C}} = (\underline{\underline{S}}_{n-1,\ n} \cdots \underline{\underline{S}}_{2,3}\ \underline{\underline{S}}_{1,2})\ \underline{\underline{X}} = \underline{\underline{S}}^H\ \underline{\underline{X}} ,$$

onde $\underline{\underline{S}}^H$, sendo o produto de matrizes unitárias, é também uma matriz unitária. Assim, tem-se uma decomposição do tipo QR para matriz $\underline{\underline{X}}$ que se escreve:

$$\underline{\underline{X}} = \underline{\underline{S}}\ \underline{\underline{C}} .$$

Fazendo $\underline{\underline{X}} = \underline{\underline{A}}_k - \alpha_k\ \underline{\underline{I}}$ obtém-se $\underline{\underline{Q}}_k = \underline{\underline{S}}$ e $\underline{\underline{R}}_k = \underline{\underline{C}}$.

Resta mostrar que o produto $\underline{\underline{C}}\ \underline{\underline{S}}$, gerador do elemento seguinte da seqüência, está na forma de Hessenberg. Este produto pode ser escrito como:

$$\underline{\underline{C}}\ \underline{\underline{S}} = \underline{\underline{C}}(\underline{\underline{S}}^{H}_{1,2} \cdots \underline{\underline{S}}^{H}_{n-1,n}).$$

Considerando a forma triangular da matriz $\underline{\underline{C}}$, é possível demonstrar por indução finita que o produto acima gera uma matriz na forma de Hessenberg. Esta demonstração baseia-se no fato de que o produto de uma matriz $\underline{\underline{M}}$ por $\underline{\underline{S}}_{k,k+1}$ altera apenas as colunas k e k+1 da matriz $\underline{\underline{M}}$, e as novas colunas são combinações lineares das colunas originais k e k+1.

d) Convergência do Método QR

Viu-se anteriormente a relação de recorrência:

$$\underline{\underline{A}}_{k+1} = \underline{\underline{Q}}^{H}_{k}\ \underline{\underline{A}}_{k}\ \underline{\underline{Q}}_{k}.$$

Outro resultado importante é que a forma de Hessenberg não é destruída.

Para provar que a seqüência $\underline{\underline{A}}_{k}$, k = 1, 2, 3 ... converge para uma matriz triangular superior, basta mostrar que os elementos abaixo da diagonal principal, $(a_{i+1,i})_{k+1}$, tendem a zero quando $k \to \infty$, o que é feito a seguir.

Considere-se a relação:

$$\underline{\underline{A}}_{k+1} = \underline{\underline{Q}}^{H}_{k}\ \underline{\underline{A}}_{k}\ \underline{\underline{Q}}_{k}$$

que pode ser reescrita como:

$$\underline{\underline{A}}_{k+1} = (\underline{\underline{S}}_{n-1,n} \cdots \underline{\underline{S}}_{2,3}\ \underline{\underline{S}}_{1,2})\ \underline{\underline{A}}_{k}(\underline{\underline{S}}^{H}_{1,2}\ \underline{\underline{S}}^{H}_{2,3} \cdots \underline{\underline{S}}^{H}_{n-1,n}),$$

e construa-se a seqüência:

$$\underline{\underline{T}}_{1} = \underline{\underline{A}}_{k},$$

$$\underline{\underline{T}}_m = \underline{\underline{S}}_{m-1,m}\, \underline{\underline{T}}_{m-1}\, \underline{\underline{S}}^H_{m-1,m} , \qquad m = 2, \ldots, n$$

geradora dos elementos de $\underline{\underline{A}}_{k+1} = \underline{\underline{T}}_n$. As relações que permitem determinar os elementos de cada matriz $\underline{\underline{T}}_m$ resultam em:

$$(t_{ij})_m = (t_{ij})_{m-1} ,$$

$$(t_{i,m-1})_m = [t_{i,m-1}\, t^*_{m-1,m-1}/\delta + t_{i,m}\, t^*_{m,m-1}/\delta]_{m-1} ,$$

$$(t_{i,m})_m = [t_{i,m}\, t^*_{m-1,m-1}/\delta - t_{i,m-1}\, t^*_{m,m-1}/\delta]_{m-1} ,$$

$$(t_{m-i,j})_m = [t_{m-1,j}\, t_{m-1,m-1}/\delta + t_{mj}\, t_{m,m-1}/\delta]_{m-1} ,$$

$$(t_{mj})_m = [t_{mj}\, t_{m-1,m-1}/\delta - t_{m-1,j}\, t_{m,m-1}/\delta]_{m-1} ,$$

para $i, j \neq m-1, m$. Adicionalmente

$$(t_{m,m-1})_m = [t_{m,m-1}|t_{m-1,m-1}|^2/\delta^2 + t^*_{m,m-1}\, t_{m,m}\, t_{m-1,m-1}/\delta^2 - t_{m,m-1}|t_{m-1,m-1}|^2/\delta^2 - t_{m-1,m}|t_{m,m-1}|^2/\delta^2]_{m-1} ,$$

$$(t_{m-1,m})_m = [t_{m-1,m}|t_{m-1,m-1}|^2/\delta^2 - t^*_{m,m-1}\, t^2_{m-1,m-1}/\delta^2 + t_{m,m-1}\, t_{m,m}\, t^*_{m-1,m-1}/\delta^2 - t_{m,m-1}|t_{m,m-1}|^2/\delta^2]_{m-1} ,$$

$$(t_{m-1,m-1})_m = [t_{m-1,m-1}|t_{m-1,m-1}|^2/\delta^2 + t_{m-1,m-1}\, t_{m-1,m}\, t^*_{m,m-1}/\delta^2 + t^2_{m,m-1}\, t^*_{m-1,m-1}/\delta^2 + t_{m,m}|t_{m,m-1}|^2/\delta^2]_{m-1} ,$$

$$(t_{m,m})_m = [t_{m,m}|t_{m-1,m-1}|^2/\delta^2 - t_{m-1,m-1}|t_{m,m-1}|^2/\delta^2$$

$$+ t_{m-1,m-1}|t_{m,m-1}|^2/\delta^2 - t_{m-1,m}\, t_{m,m-1}\, t^*_{m-1,m-1}/\delta^2]_{m-1},$$

onde:

$$\delta_m = (|t_{m,m}|^2 + |t_{m+1,m}|^2)^{1/2},\ m = 1,\ \ldots,\ n-1\ .$$

Fazendo m = M - 1, M, M + 1, onde M inteiro é um valor particular escolhido, obtêm-se todas as transformações que afetam o elemento $t_{m,m-1}$. Para m = M - 1 esse elemento é alterado por:

$$(t_{M,M-1})_{M-1} = [t_{M,M-1}\, t^*_{M-2,M-2}/\delta - t_{M,M-2}\, t^*_{M-1,M-2}/\delta]_{M-2}\ .$$

Para m = M a alteração é expressa por:

$$(t_{M,M-1})_M = [t_{M,M-1}|t_{M-1,M-1}|^2/\delta^2 + t^*_{M,M-1}\, t_{MM}\, t_{M-1,M-1}/\delta^2 -$$

$$- t_{M,M-1}|t_{M-1,M-1}|^2/\delta^2 - t_{M-1,M}|t_{M,M-1}|^2/\delta^2]_{M-1}\ .$$

Para m = M + 1 ocorre a última alteração através de:

$$(t_{M,M-1})_{M+1} = [t_{M,M-1}\, t_{MM}/\delta + t_{M+1,M-2}\, t_{M+1,M}/\delta]_M\ .$$

É interessante analisar o caso em que os elementos abaixo da diagonal principal tendem a zero. Para isto, os elementos da primeira diagonal abaixo da principal são considerados perturbações de primeira ordem no sistema. Desprezando as perturbações de segunda ordem, dadas pelo produto de dois elementos abaixo da diagonal principal, têm-se as alterações simplificadas:

$$(t_{M,M-1})_{M-1} = [t_{M,M-1}(t^*_{M-2,M-2}/t_{M-2,M-2})^{1/2}]_{M-2}\ ,$$

$$(t_{M,M-1})_M = [t^*_{M,M-1}\, t_{MM}/t^*_{M-1,M-1}]_{M-1} ,$$

$$(t_{M,M-1})_{M+1} = [t_{M,M-1}(t_{MM}/t^*_{MM})^{1/2}]_M .$$

Essencialmente, a única transformação que pode alterar o valor absoluto de $t_{M,M-1}$ é dada pela segunda relação acima. Para que $t_{M,M-1} \to 0$, deve-se mostrar que a situação $|t_{M-1,M-1}| < |t_{MM}|$ não pode subexistir. Isto pode ser entendido considerando o caso em que $|t_{M-1,M-1}| < |t_{MM}|$ no início do processo de recorrência (k = 1). Nestas condições é possível que o processo se afaste do equilíbrio até que:

$$|t_{M,M-1}| > |t_{M-1,M-1}| .$$

As equações para cálculo de $t_{M-1,M-1}$ e t_{MM}, longe do ponto de equilíbrio, são então simplificadas para:

$$(t_{M-1,M-1})_M = [t_{MM}|t_{M,M-1}|^2/\delta^2]_{M-1} ,$$

$$(t_{MM})_M = [-t^*_{M-1,M-1}\, t_{M-1,M}\, t_{M,M-1}/\delta^2]_{M-1} ,$$

e adicionalmente:

$$(t_{M,M-1})_M = [-t_{M-1,M}|t_{M,M-1}|^2/\delta^2]_{M-1} ,$$

$$(t_{M-1,M})_M = [-t_{M-1,M}|t_{M,M-1}|^2/\delta^2]_{M-1} .$$

As duas primeiras equações mostram que $(t_{M-1,M-1})_M$ é proporcional a $(t_{MM})_{M-1}$ e $(t_{MM})_M$ é proporcional a $(t^*_{M-1,M-1})_{M-1}$, havendo assim uma inversão relativa de posições. Concomitantemente, como mostram as duas últimas equações, $t_{M,M-1}$ troca de posição com $t_{M-1,M}$, atenuados pela relação $|t_{M,M-1}|^2/\delta^2$ que é sempre menor do que 1.

Consequentemente, a situação $|t_{M-1,M-1}| < |t_{MM}|$ é instável e retorna sistematicamente ao ponto de equilíbrio com:

$$\left.\begin{array}{l} |t_{MM}| < |t_{M-1,M-1}| \\ |t_{M,M-1}| \to 0 \end{array}\right\} \quad M = 2, \ldots, m \ .$$

Observa-se assim que o método QR é sempre convergente e estável, sendo por isto preferido a outros métodos.

Nessas condições, o sistema de transformações é consistente com a hipótese adotada de que $t_{M,M-1} \to 0$. Pode-se então concluir que o método QR conduz a matriz a uma forma triangular superior, em que os autovalores estão dispostos na diagonal principal em ordem decrescente de valor absoluto.

Do que foi exposto acima, somente quando dois autovalores possuirem o mesmo valor absoluto é que o método QR pode encontrar dificuldades. Este impasse entretanto é facilmente resolvido, lembrando-se que a relação completa é dada por:

$$(t_{M,M-1})_M = [t^*_{M,M-1} \ t_{MM} \ t_{M-1,M-1}/\delta^2]_{M-1} \ ,$$

e, como δ^2 é sistematicamente maior do que o valor absoluto de $t_{MM} \ t_{M-1,M-1}$ garante-se a convergência em todos os casos.

Viu-se que a relação $t_{MM}/t^*_{M-1,M-1}$ é a responsável (no caso simplificado) pela convergência do processo. Esta relação tende ao valor:

$$\rho_M = \frac{\lambda_M}{\lambda_{M-1}} \ , \ \lambda \text{ autovalor},$$

no limite quando $K \to \infty$. Este quociente entre autovalores consecutivos é chamado raio de convergência e o seu inverso mede a velocidade de convergência. Para aumentar a velocidade de convergência, trabalha-se com a matriz $\underline{\underline{A}} - \alpha \underline{\underline{I}}$ alterando assim a velocidade de convergência para:

$$v_n = \rho_n^{-1} = \frac{\lambda_{n-1} - \alpha}{\lambda_n - \alpha},$$

onde foi escolhida propositalmente a velocidade de convergência para o último autovalor, por ser ela a mais crítica no que se refere a isolar um autovalor (o menor). Assim, quanto mais próximo for α de λ_n, mais rapidamente este valor é isolado na forma triangular da matriz $\underline{\underline{A}}$. Considerando a matriz nesta forma e λ_n o último autovalor, o autovetor a ele correspondente é:

$$\underline{u}_n^t = (0, \ldots, 0, 1).$$

Pode-se usar para α a estimativa do autovalor λ_n, dada pelo quociente de Rayleigh:

$$\alpha_K = \frac{\underline{u}_n^t \underline{\underline{A}}_k \underline{u}_n}{\underline{u}_n^t \underline{u}_n} = \underline{u}_n^t \underline{\underline{A}}_k \underline{u}_n.$$

Uma vez isolado o autovalor λ_n, a matriz $\underline{\underline{A}}$ é modificada adquirindo a forma:

$$\underline{\underline{\bar{A}}} = \begin{bmatrix} \underline{\underline{\bar{A}}}_1 & \underline{v} \\ \underline{0}^t & \lambda_n \end{bmatrix}.$$

Para identificar os outros autovalores, aplica-se em seguida o método QR para a matriz reduzida $\underline{\underline{\bar{A}}}_1$ e assim sucessivamente.

21.3.4 - CÁLCULO DOS AUTOVETORES

Determinado um autovalor é possível encontrar o autovetor a ele correspondente, resolvendo o sistema homogêneo de equações mencionado na Seção 21.3.3. Como via de regra só se determina o valor aproximado do autovalor, este processo acarreta uma propagação de erros. Para

evitar este problema, foram desenvolvidos métodos mais convenientes para este propósito. A seguir discute-se um desses métodos.

Primeiramente, considere-se a transformação de similaridade:

$$\underline{\underline{A}} = \underline{\underline{M}}\ \underline{\underline{H}}\ \underline{\underline{M}}^{-1}$$

que relaciona a matriz $\underline{\underline{A}}$ com sua forma especial $\underline{\underline{H}}$ (tridiagonal ou de Hessenberg). Seja λ um autovalor e $\underline{v}$ o correspondente autovetor relativo à matriz $\underline{\underline{H}}$. Segue-se então que:

$$\lambda\ \underline{\underline{M}}\ \underline{v} = \underline{\underline{M}}\ \underline{\underline{H}}\ \underline{v} = \underline{\underline{M}}\ \underline{\underline{H}}\ \underline{\underline{M}}^{-1}(\underline{\underline{M}}\ \underline{v}) = \underline{\underline{A}}(\underline{\underline{M}}\ \underline{v})\ ,$$

de onde se conclui que o autovetor, correspondente a λ, relativo à matriz $\underline{\underline{A}}$ é dado por:

$$\underline{w} = \underline{\underline{M}}\ \underline{v}\ .$$

Considere-se agora o problema da determinação do autovetor $\underline{v}$. Para isto, construa-se a seqüência:

$$\left.\begin{aligned} (\underline{\underline{H}} - \alpha\ \underline{\underline{I}})\ \underline{v}_{i+1} &= \underline{z}_i \\ \underline{z}_i &= m_i^{-1}\ \underline{v}_i \end{aligned}\right\} \quad i = 1, 2, \ldots\ ,$$

onde α é um valor aproximado para o autovalor λ, $\underline{v}_0$ um vetor inicial escolhido arbitrariamente e m_i calculado por:

$$|m_i| = \max_k\ [(v_k)_i]\ ,$$

onde os valores $[(v_k)_i]$ indicam as componentes do vetor $\underline{v}_i$. Deve-se demonstrar que esta seqüência converge para o autovetor $\underline{v}/|\underline{v}|$.

Considerado um inteiro j, a expressão de recorrência da seqüência acima pode ser reescrita como:

$$\underline{v}_j = (\underline{\underline{H}} - \alpha\,\underline{\underline{I}})^{-j}\,(m^{-1}\,\underline{v}_0)\ ,$$

sendo:

$$m = \prod_{i=1}^{j} m_i .$$

Adimitindo-se que $\underline{v}_0$ é decomposto segundo a base constituída pelos autovetores normalizados $\underline{b}_k$, escreve-se:

$$\underline{v}_0 = \sum_{k=1}^{n} c_k\,\underline{b}_k\ ,$$

o que resulta em:

$$\underline{v}_j = \sum_{k=1}^{n} \frac{c_k}{(\lambda_k - \alpha)^j}\,\underline{b}_k .$$

Seja $\lambda_p = \lambda$ o autovalor considerado. Como $|\lambda-\alpha| < |\lambda_k-\alpha|$ para $k \neq p$, tem-se finalmente:

$$\underline{z}_j = \underline{b}_p + \ldots\ ,$$

onde os termos omitidos têm linha zero para $j \to \infty$. Este último resultado mostra que $\underline{z}_j$ tende a $\underline{b}_p = \underline{v}/|\underline{v}|$ que é o autovetor correspondente a $\lambda_p = \lambda$.

Este método também pode ser usado para calcular os autovalores, pois:

$$\underline{v}_{j+1} = \left(\frac{1}{\lambda - \alpha}\right) m_i^{-1}\,\underline{v}_j\ .$$

Assim o quociente entre as componentes dominantes de $\underline{v}_j$ e $\underline{v}_{j+1}$ fornecem uma estimativa da diferença $\lambda - \alpha$ e, consequentemente, λ.

No caso de autovalores múltiplos, um desenvolvimento análogo pode ser efetuado, mas cuidados adicionais devem ser tomados.

O fato de ser requerida a solução de um sistema de equações a cada passo do processo de recorrência, $(\underline{\underline{H}} - \alpha \underline{\underline{I}})\underline{v}_{j+1} = \underline{z}_j$, não apresenta inconvenientes, pois, dada a forma particular de $\underline{\underline{H}}$, o método de eliminação de Gauss fornece rapidamente a solução desejada.

21.4 - CÁLCULO DE PSEUDO-INVERSAS

A definição da pseudo-inversa $\underline{\underline{A}}$ de uma matriz $\underline{\underline{V}}$ foi apresentada no Capítulo 5 (Seção 5.7). Adicionalmente foi apresentado o problema de unicidade da pseudo-inversa (Exercícios 9 e 10 do Capítulo 5). Foi mostrado que as relações:

$$\underline{v} = \underline{\underline{V}}\,\underline{x}$$

e:

$$\underline{x} = \underline{\underline{A}}\,\underline{v}$$

constituiam o equacionamento básico para a definição da pseudo-inversa. Dessas relações resultaram:

$$\underline{\underline{A}}\,\underline{\underline{V}} = \underline{\underline{I}}_x \quad ,$$

$$\underline{\underline{V}}\,\underline{\underline{A}} = \underline{\underline{I}}_v$$

que, obedecidas simultaneamente, permitiam uma definição univoca da pseudo-inversa (Exercício 10, Capítulo 5). Esta definição é de E. H. Moore (Ben-Israel and Charnes, 1963).

Uma definição equivalente, a de Tseng (ver Ben-Israel and Charnes, 1963), consiste na resolução da equação:

$$\underline{\underline{V}}\,\underline{x} = \underline{v} \quad ,$$

onde $\underline{v}$ e $\underline{\underline{V}}$ são conhecidos, pelo método do mínimo dos quadrados (Capítulo 19; Pennington, 1970). Em seguida determina-se a matriz $\underline{\underline{A}}$ que permite executar a transformação:

$$\underline{x} = \underline{\underline{A}}\,\underline{v} \quad .$$

Esta definição permite o cálculo da pseudo-inversa $\underline{\underline{A}}$.

Uma terceira definição de Penrose (ver Ben-Israel and Charnes, 1963) considera a inversa $\underline{\underline{A}}$ como solução do sistema de equações:

$$\underline{\underline{V}}\,\underline{\underline{A}}\,\underline{\underline{V}} = \underline{\underline{V}} \quad ,$$

$$\underline{\underline{A}}\,\underline{\underline{V}}\,\underline{\underline{A}} = \underline{\underline{A}} \quad ,$$

$$(\underline{\underline{V}}\,\underline{\underline{A}})^H = \underline{\underline{V}}\,\underline{\underline{A}} \quad ,$$

$$(\underline{\underline{A}}\,\underline{\underline{V}})^H = \underline{\underline{A}}\,\underline{\underline{V}} \quad ,$$

onde $\underline{\underline{V}}$ é conhecido. Mostra-se facilmente que estas relações são derivadas das quatro primeiras igualdades apresentadas nesta seção. Existe assim uma equivalência desta definição com as precedentes.

Outras definições são apresentadas, envolvendo não só matrizes, mas também operadores lineares. Este caso mais geral não é discutido neste trabalho.

21.4.1 - MÉTODO DO MÍNIMO DOS QUADRADOS

Um método mais simples para determinação da pseudo-inversa (ou inversa generalizada) utiliza o método do mínimo dos quadrados. Este método pode ser implementado a partir da definição de Tseng ou usando a definição de Penrose. Utiliza-se esta última por permitir uma obtenção rápida do resultado.

Considerando o caso de um número de equações maior que o de incógnitas, utilizam-se as relações:

$$\underline{\underline{V}}\ \underline{\underline{A}}\ \underline{\underline{V}} = \underline{\underline{V}} \quad ,$$

$$(\underline{\underline{V}}\ \underline{\underline{A}})^H = \underline{\underline{V}}\ \underline{\underline{A}} \quad ,$$

de onde se obtém:

$$\underline{\underline{A}}^H = \underline{\underline{V}}\ (\underline{\underline{V}}^H\ \underline{\underline{V}})^{-1}$$

e finalmente:

$$\underline{\underline{A}} = (\underline{\underline{V}}^H\ \underline{\underline{V}})^{-1}\ \underline{\underline{V}}^H \quad .$$

Para o caso de um número de equações menor que o de incógnitas, recorre-se às relações:

$$\underline{\underline{V}}\ \underline{\underline{A}}\ \underline{\underline{V}} = \underline{\underline{V}} \quad ,$$

$$(\underline{\underline{A}}\ \underline{\underline{V}})^H = \underline{\underline{A}}\ \underline{\underline{V}} \quad ,$$

de onde resulta:

$$\underline{\underline{A}}^H = (\underline{\underline{V}}\ \underline{\underline{V}}^H)^{-1}\ \underline{\underline{V}}$$

que conduz a:

$$\underline{\underline{A}} = \underline{\underline{V}}^H (\underline{\underline{V}}\ \underline{\underline{V}}^H)^{-1} \quad .$$

Esta forma de obtenção da pseudo-inversa é considerada em Korganoff (1967), a qual dispensa a relação $\underline{\underline{A}}\ \underline{\underline{V}}\ \underline{\underline{A}} = \underline{\underline{A}}$ para a definição da inversa generalizada.

A vantagem deste procedimento é a obtenção direta da inversa generalizada de uma matriz retangular.

21.4.2 - MÉTODO ITERATIVO

Os esquemas iterativos são gerados pela equação $\underline{\underline{A}}\ \underline{\underline{V}}\ \underline{\underline{A}} = \underline{\underline{A}}$. Assim, conhecida uma aproximação inicial $\underline{\underline{A}}_0$, pode-se considerar a recorrência:

$$\underline{\underline{A}}_{i+1} = \underline{\underline{A}}_i\ \underline{\underline{V}}\ \underline{\underline{A}}_i \quad , \qquad i = 1, 2, \ldots,$$

que produz a sequência:

$$\underline{x}_{i+1} = \underline{\underline{A}}_i\ \underline{\underline{V}}\ \underline{x}_i \quad , \qquad i = 1, 2, \ldots,$$

cuja convergência ocorre quando:

$$|(\underline{\underline{A}}_i\ \underline{\underline{V}} - \underline{\underline{I}}_x)\underline{x}| < |\underline{\underline{I}}_x\ \underline{x}|$$

para qualquer $\underline{x}$. Esta condição pode ser testada no decorrer do processo iterativo.

Um método iterativo sempre convergente é discutido em Ben-Israel (1965) e baseia-se no esquema iterativo:

$$\underline{\underline{A}}_{i+1} = \underline{\underline{A}}_i\ (2\underline{\underline{I}}_V - \underline{\underline{V}}\ \underline{\underline{A}}_i) \quad , \qquad i = 1, 2, \ldots \quad .$$

A prova da convergência baseia-se na condição de convergência acima especificada, o que nos permite assegurar a convergência de qualquer processo fundamentado na equação $\underline{\underline{A}}\,\underline{\underline{V}}\,\underline{\underline{A}} = \underline{\underline{A}}$ (ver Ben-Israel, 1965).

21.4.3 - MÉTODO DA DECOMPOSIÇÃO DA MATRIZ A SER INVERTIDA

Um método alternativo para o cálculo da pseudo-inversa de uma matriz $\underline{\underline{V}}$ foi proposto por Golub e Kahan (1965) e tem por base a decomposição:

$$\underline{\underline{V}} = \underline{\underline{U}}_1\ \underline{\underline{W}}\ \underline{\underline{U}}_2^H \quad ,$$

onde $\underline{\underline{U}}_1$ e $\underline{\underline{U}}_2$ são matrizes unitárias e $\underline{\underline{W}}$ é uma matriz composta por uma matriz diagonal D e uma matriz nula complementar. No caso de um número de equações maior que o de incógnitas, tem-se:

$$\underline{\underline{W}} = \begin{bmatrix} \underline{\underline{D}} \\ \underline{\underline{0}} \end{bmatrix}$$

onde o número de elementos da matriz diagonal é igual ao número de incógnitas.

Pode-se utilizar o método de Gauss-Jordan na forma de pre-multiplicações e pós-multiplicações da matriz $\underline{\underline{V}}$ (composição dos Exercícios 3, 4 e 6 deste capítulo) para obtenção das matrizes $\underline{\underline{U}}_1$, $\underline{\underline{U}}_2$ e $\underline{\underline{W}}$.

A pseudo-inversa $\underline{\underline{A}}$ da matriz $\underline{\underline{V}}$ é calculada pelo produto:

$$\underline{\underline{A}} = \underline{\underline{U}}_1^H\ \underline{\underline{W}}'\ \underline{\underline{U}}_2 \quad ,$$

onde:

$$\underline{\underline{W}}' = [\ \underline{\underline{D}}^{-1} \quad \underline{\underline{0}}\ ]$$

(Golub e Kahan, 1965).

Apesar deste método ser bem conhecido, o número de operações requeridas para sua implementação pode tornar a conjunção dos dois métodos precedentes competitiva e de menor complexidade.

EXERCÍCIOS

1. Determine a solução dos sistemas triangulares de equações:

$$\sum_{j=i}^{N} c_{ij}\, x_j = y_i \,, \qquad i = 1, \ldots, N \,,$$

$$\sum_{j=1}^{i} b_{ij}\, y_j = v_i \,, \qquad i = 1, \ldots, N \,.$$

2. Mostre que a maximização de $(a_{ii})_{i-1}$ pode melhorar a qualidade dos resultados quando poucos dígitos significativos são usados (consulte Ralston, 1965).

3. Mostre que a eliminação da incógnita x_k é equivalente a multiplicar a matriz $\underline{\underline{A}}_{k-1}$ pela matriz $\underline{\underline{M}}_k$, obtendo a matriz $\underline{\underline{A}}_k = \underline{\underline{M}}_k \underline{\underline{A}}_{k-1}$. A matriz $\underline{\underline{M}}_k$ é dada por:

coluna k

$$\underline{\underline{M}}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ & 1 & & & & \\ 0 & & \ddots & & & \\ & & & 0 & \cdots & 0 \\ & & 0 & 1 & & \\ & & -(a_{k+1,k})_{k-1} & & 1 & \\ & & & & & 0 \\ 0 & \cdots & 0 & -(a_{Nk})_{k-1} & 0 \cdots 0 & 1 \end{bmatrix}$$

e a matriz $\underline{\underline{A}}_0$ é igual a matriz $\underline{\underline{A}}$ do sistema $\underline{\underline{A}}\,\underline{x} = \underline{v}$.

4. Mostre que o produto:

$$\underline{\underline{M}} = \underline{\underline{M}}_{N-1} \cdots \underline{\underline{M}}_2 \, \underline{\underline{M}}_1$$

é uma matriz triangular inferior com elementos unitários na diagonal.

5. Determine a forma e as características da matriz $\underline{\underline{B}} = \underline{\underline{M}}^{-1}$.

6. Para N = 3 mostre através de um exemplo que:

$$\underline{\underline{M}}\,\underline{\underline{A}} = \underline{\underline{C}} \ ,$$

onde $\underline{\underline{C}}$ é uma matriz triangular superior.

7. O método de eliminação de Gauss-Jordan consiste em eliminar a variável x_k não só das equações subseqüentes, mas também das equações precedentes à equação considerada. Escreva as equações do desenvolvimento recursivo deste método.

8. Simule um sistema de 3 equações a 3 incógnitas e, em seguida, resolva-o pelos métodos de:

   a) eliminação de Gauss,

   b) ortogonalização.

   Estude criticamente os resultados.

9. Escreva um programa para resolver sistemas de equações tridiagonais utilizando:

   a) o método de eliminação de Gauss-Jordan (Exercício 7),

   b) a variante do método de relaxação do item 1 da Seção 21.2.1.

   Compare criticamente os resultados.

10. Escreva um programa para resolver um sistema de equações usando o método de iteração simples item a do tópico 2 da Seção 21.2.2. Discuta a eficiência do método.

11. Escreva um programa para resolver um sistema de equações pelos métodos de Jacobi e Gauss-Seidel. Compare a eficiência dos métodos.

12. Implemente o método de sobre-relaxação utilizando as mesmas características utilizadas no exercício anterior e um valor $\omega = 1{,}5$. Compare a eficiência deste método com a do método de Gauss-Seidel.

13. Justifique as propriedades dos resíduos:

    1) $r_m(km) = 0$, k inteiro ,

    2) $r_m(x) = x$, $0<x<m$,

    3) $r_m[x \pm y] = r_m[r_m(x) \pm r_m(y)]$,

    4) $r_m[xy] = r_m[r_m(x)\ r_m(y)]$.

14. Mostre que uma das formas para encontrar uma coleção de inversos de y, com respeito a resíduo no módulo m, é procurar os valores inteiros de α, tais que $\alpha\, m + 1$ seja múltiplo de $r_m(y)$.

15. Utilizando o Teorema Chinês dos resíduos, calcule os resíduos do número 207 nas bases 2, 3, 5, 7, 11 e, a partir destes resultados, recupere o número 207.

16. Escreva um programa implementando o método dos resíduos para resolução de um sistema de equações. Compare a precisão dos resultados do programa com a do método de eliminação de Gauss.

17. Escreva um programa para inversão de matrizes utilizando o método de decomposição triangular.

18. Escreva um programa para inversão de matrizes utilizando o método de destruição de grau.

19. Estabeleça uma forma de recorrência para refinamento da inversa de uma matriz obtida aproximadamente pelos métodos apresentados no texto.

    (*Sugestão:* Considere as relações:

    $$\underline{\underline{A}}\,\underline{\underline{A}}^{-1} = \underline{\underline{I}} + \underline{\underline{E}} ,$$

    $$\underline{\underline{A}}^{-1} = \bar{\underline{\underline{A}}}^{-1}(\underline{\underline{I}} + \underline{\underline{E}}) = \bar{\underline{\underline{A}}}^{-1}\,\underline{\underline{A}}\,\bar{\underline{\underline{A}}}^{-1} ,$$

    onde $\bar{\underline{\underline{A}}}^{-1}$ é a inversa aproximada e $\underline{\underline{E}}$ é a matriz do erro cometido).

20. Escreva um programa para colocação de uma matriz simétrica na forma tridiagonal (método de Givens-Householder).

21. Complemente o programa do exercício anterior com a obtenção de estimativa dos autovalores. Utilize o teorema de Gershgorin.

22. Escreva um programa para redução de uma matriz não-hermitiana à forma de Hessenberg.

23. Implemente o método QR para complementar o programa do Exercício 21.

24. Implemente o método QR para complementar o programa do Exercício 22.

25. Complemente o programa dos Exercícios 23 e 24 com o cálculo de autovetores.

26. Escreva um programa para calcular os autovalores utilizando o método de iteração inversa (Seção 21.3.4).

27. Determine a pseudo-inversa da matriz:

$$\underline{\underline{V}} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}$$

Utilizando primeiramente a conjunção dos métodos das Seções 21.4.1 e 21.4.2 e, em seguida, o método da Seção 21.4.3. Faça uma análise comparativa dos seguintes aspectos:

a) qualidade do resultado para um trabalho computacional eqüivalente,

b) velocidade da obtenção do resultado.

BIBLIOGRAFIA

BAKHVALOV, N.S. *Numerical methods.* Moscou, Editorial Mir., 1977.

BEN-ISRAEL, A.; CHARNES, A. Contributions to the theory of generalized inverses. *J. Soc. Indust. Appl. Math.*, II(3):667-699, 1963.

BEN-ISRAREL, A. An iterative method for computing the generalized inverse of an arbitrary matrix. *Math. of Computation*, 19(91):452-456, 1965.

BEREZIN, I.S.; ZHIDKOV, N.P. *Computing methods.* Massachussets. Addison-Wesley, 1965. 2 v.

CARNAHAN, B.; LUTHER, H.A.; WILKES, J.O. *Applied numerical methods.* New York, John Wiley, 1969.

FRIEDMAN, B. *Principles and techniques of applied mathematics.* New York, John Wiley, 1956.

GANTMACHER, F.R. *The theory of matrices.* New York, Chelsea, 1959. 2 v.

GOLUB, G.; KAHAN, W. Calculating the singular values and pseudo-inverse of a matrix. *J. SIAM Numer. Anal.*, 2:205-223, 1965.

HOWELL, J.A. Algorithm 406 exact solution of linear equations using residue arithmetics. *Communications ACM*, 14:180-184, 1971.

HOWELL, J.A.; GREGORY, R.T. *An algorithm for solving linear algebraic equations using residue arithmetics I. Bit*, 9:200-224, 1969.

HOWELL, J.A.; GREGORY, R.T. *An algorithm for solving linear algebraic equations using residue arithmetics II. Bit*, 9:324-337, 1969.

JACOBS, D. *The state of the art in numerical analysis.* London, Academic Press, 1977.

JOHN, F. *Lecture in advanced numerical analysis.* New York, Gordon and Breach Science, 1967.

KORGANOFF, A.; PAVEL-PARVIM, M. *Elements de théore des matrices carrées et rectangles en analyse numérique Dunod.* Paris, 1967. Tome 2.

PENNINGTON, R.H. *Introductory computer methods and numerical analysis.* London, Collier MacMillan, 1970.

RALSTON, A. *A first course in numerical analysis.* McGraw-Hill, 1965.

RALSTON, A.; WILF, H.S. *Mathematical methods for digital computers.* New York, John Wiley, 1968. 2 v.

SZIDAROVSZKY, F.; YAKOVITZ, S. *Principles and procedures of numerical analysis.* New York, Plenum, 1978.

TODD, J. *A survey of numerical analysis.* New York, McGraw-Hill, 1962.

VINOGRADOV, I. *Fundamentos de la teoria de los números.* Editorial Mir., 1977.

WENDROFF, B. *Theoretical numerical analysis.* London, Academic Press, 1967.

YOUNG, D.M.; GREGORY, R.T. *A survey of numerical mathematics.* Reading, Addison-Wesley, 1973. v. 2.

ZAMLUTTI, C.J. *O problema da recuperação da informação na utilização da aritmética de resíduos.* São José dos Campos, (INPE-2485-PRE/169), 1982.

CAPÍTULO 22

SOLUÇÃO DE EQUAÇÕES DIFERENCIAIS ORDINÁRIAS

22.1 - INTRODUÇÃO

Um sistema de equações diferenciais ordinárias de primeira ordem pode ser colocado na forma genérica implícita:

$$\underline{\psi}\,[x,\ \underline{v}(x),\ \underline{v}'(x)] = 0\,,$$

ou então na forma explícita:

$$\underline{v}'(x) = \underline{f}\,[x,\ \underline{v}(x)],\ [x,\ \underline{v}(x)] \in \Omega\,,$$

onde Ω é um domínio dado, consistindo o problema na determinação das funções $v_i(x)$, $i = 1, \ldots, N$.

A formulação acima exige a determinação genérica de $\underline{v}(x)$. Para restringi-la são fornecidas condições adicionais, em Ω, que devem ser satisfeitas por $\underline{v}(x)$. Estas restrições permitem classificar três tipos distintos de problemas:

a) Problemas de valor inicial, para os quais é fornecida a informação adicional:

$$\underline{v}(x_0) = \underline{v}_0\ .$$

b) Problemas de valor de contorno, para os quais são especificadas restrições adicionais na forma:

$$\underline{g}\,[\underline{v}(\underline{x}_p),\ \underline{v}'(\underline{x}_p)] = \underline{0}\ .$$

c) Problemas unidimensionais de autovalores, para os quais a forma genérica é:

$$T\,[x,\ \underline{y}(x)] = \lambda v(x)\ ,$$

onde:

$$y_i(x) = v^{(i-1)}(x)\ .$$

O número λ é chamado autovalor, a função $v(x)$ autofunção e como informação adicional, fornecida pelas condições de contornos, tem-se:

$$\underline{g}\ [\underline{y}(x_p)] = \underline{0}\ .$$

Equações diferenciais de ordem superior são facilmente redutíveis a sistemas de equações diferenciais de primeira ordem. Isto pode ser verificado imediatamente considerando a equação de segunda ordem:

$$g''(x) = F\,[x,\ z(x),\ z'(x)]\ ,$$

que se reduz ao sistema:

$$\underline{v}'(x) = \underline{f}\ [x,\ \underline{v}(x)]\ ,$$

fazendo:

$$v_1(x) = z(x)\ ,$$

$$v_2(x) = z'(x)\ ,$$

de onde decorrem:

$$f_1\ [x,\ \underline{v}(x)] = v_2(x)\ ,$$

$$f_2\ [x,\ \underline{v}(x)] = F\,[x,\ v_1(x),\ v_2(x)]\ .$$

O tratamento de sistemas de equações diferenciais de primeira ordem engloba a maioria dos problemas práticos de equações diferenciais ordinárias.

Existem basicamente duas formas para obtenção da solução aproximada de uma equação diferencial:

a) representação de v(x) como soma de um número finito de funções independentes,

b) representação de v(x) por meio de seus valores para um conjunto discreto de pontos.

O segundo método é o mais apropriado para utilização em computadores.

Os problemas de valores de contorno podem eventualmente ser convertidos em problemas de valores iniciais, razão pela qual é dada maior ênfase a estes últimos.

Um tratamento completo para a solução numérica de equações diferenciais ordinárias envolve toda a teoria de aproximação já abordada, tornando-se impraticável em um único capítulo. Optou-se neste trabalho por uma apresentação que focalizasse mais a idéia geradora dos métodos do que propriamente o desenvolvimento destes. Esta forma tem por objetivo dar ao leitor os elementos essenciais para uma melhor compreensão de textos mais avançados no assunto.

Para que o leitor adquira familiaridade com os elementos envolvidos na solução de equações diferenciais ordinárias, prefere-se apresentar neste capítulo, com maior detalhe, o problema unidimensional. Os casos $N > 1$ constituem, via de regra, uma simples extensão do problema unidimensional, não oferecendo assim maiores dificuldades. Outra restrição é a consideração de funções $\underline{f}\,[x, \underline{v}(x)]$ que satisfazem a condição de Lipschitz:

$$\left|\underline{f}\ [x,\ \underline{v}(x+\Delta x)] - \underline{f}\ [x,\ \underline{v}(x)]\right| < M\ \left|\underline{v}(x+\Delta x) - \underline{v}(x)\right| ,$$

onde M é um número real.

22.2 - PROBLEMAS DE VALOR INICIAL

Os problemas de valor inicial são caracterizados pelas relações:

$$\left.\begin{aligned} v'(x) &= f\ [x,\ v(x)] \\ v(x_0) &= v_0 \end{aligned}\right\}\ x_0,\ x \in (a,b) ,$$

sendo $f\ [x,\ v(x)]$ e v_0 conhecidos. Deseja-se obter $v(x)$ para $x \in (a,b)$. Em geral $x_0 = a$, embora não necessariamente.

A seguir, apresentam-se alguns métodos para a obtenção de $v(x)$.

22.2.1 - APROXIMAÇÃO DA SOLUÇÃO POR SÉRIE DE FUNÇÕES

A primeira idéia que surge para a solução de um problema de valor inicial é considerar a possibilidade da solução ser representada por uma série de funções:

$$v(x) = \sum_{i=0}^{n} c_i\ \phi_i(x) + E(x) .$$

Como exemplo, na série de Taylor:

$$v(x) \cong v_0 + \sum_{i=1}^{n} \frac{(x - x_0)^i}{i!}\ v^{(i)}\ (x_0) ,$$

as derivadas $v^{(i)}(x)$ são obtidas a partir da relação:

$$v'(x) = f[x, v(x)] = f \ ,$$

que fornece:

$$v'' = f_x + f_v \, f \ ,$$

$$v^{(3)}(x) = f_{xx} + 2 \, f_{xv} \, f + f_{vv} \, f^2 + f_x \, f_v + f_v^2 \, f \ ,$$

etc.,

sendo os subescritos x e v usados para indicar a derivação parcial em relação a essas variáveis. O erro dessa série é facilmente determinado e vale:

$$E(x) = \frac{(x - x_0)^{N+1}}{(N+1)!} \; v^{(N+1)}(\eta) \ , \qquad x_0 < \eta < x \ .$$

Pela sua natureza, este método é mais apropriado para cálculos manuais do que para computadores. Também a complexidade crescente com o aumento de ordem da derivada restringe grandemente a aplicação do método.

22.2.2 - APROXIMAÇÃO DA SOLUÇÃO PARA UM CONJUNTO DISCRETO DE PONTOS

Quando se deseja determinar v(x) por meio de seus valores para um conjunto de pontos $\{x_j\}$, j=1, ..., m, a alternativa é calcular recursivamente os valores $v_j = v(x_j)$ a partir de outros valores v_i já obtidos. Os métodos são englobados em três tipos distintos de acordo com o fundamento que o originou, a saber:

a) Métodos baseados na integração de f[x, v(x)], ou seja, recorrência com um ponto.

b) Métodos baseados em interpolação, ou seja, recorrência com múltiplos pontos.

c) Métodos autocorretivos nos quais a equação diferencial é usada para refinamento da solução.

1) Métodos de recorrência com um ponto

Fundamentam-se na integração da equação diferencial no intervalo (x_k, x_{k+1}), que fornece:

$$v_{k+1} = v_k + \int_{x_k}^{x_{k+1}} f[x, v(x)]\, dx \ ,$$

começando o processo de recorrência com o valor inicial v_0 conhecido.

Para a integração numérica usando uma única variável, torna-se necessária uma forma explícita para v(x), que permite escrever:

$$\int_{x_k}^{x_{k+1}} f[x, v(x)]\, dx \cong \int_{x_k}^{x_{k+1}} g(x)\, dx \ .$$

De acordo com a aproximação usada para v(x) e o método de integração para g(x), decorrem as múltiplas variantes do método de recorrência com um ponto.

O método mais simples é o de Euler que utiliza:

$$f[x, v(x)] \cong f[x_k, v(x_k)] = \text{constante} \ ,$$

resultando em:

$$v_{k+1} = v_k + (x_{k+1} - x_k)\, f[x_k, v(x_k)] \ .$$

O interesse deste método é mais teórico. Detalhes de seu desenvolvimento podem ser encontrados em Henrici (1962).

A partir desse método e passando por vários estágios de desenvolvimento, chega-se finalmente aos métodos do tipo Runge-Kutta de real interesse prático.

Os métodos de Runge-Kutta utilizam integração numérica para a função g(x), podendo-se escrever:

$$\int_{x_k}^{x_{k+1}} g(x)\,dx = \sum_{i=1}^{p} c_i g(y_i) = \sum_{i=1}^{p} c_i g_i \ ,$$

onde os y_i são pontos pertencentes ao intervalo (x_k, x_{k+1}). Os valores g_i são obtidos por:

$$g_i = f[y_i, v(y_i)] = f[y_i, u_i] \ .$$

O valor p é chamado ordem do método Runge-Kutta.

Os valores desejados são então obtidos pela recorrência:

$$v_{k+1} = v_k + \sum_{i=1}^{p} c_i \, g_i \ .$$

A complexidade do método aumenta, quando se utiliza uma expressão análoga a esta última para o cálculo dos valores intermediários u_i. Neste caso escreve-se:

$$u_i = v_k + \sum_{j=1}^{p} a_j \, g_j \ .$$

O sistema de equações resultante é não-linear quando:

$$a_j \neq 0 \qquad \text{para } j \geq i \ ,$$

e os métodos decorrentes são denominados implícitos. Quando, entretanto, se utiliza: $a_j=0$ para $j \geq i$ não há necessidade de resolver um sistema de equações não-lineares. Estes últimos métodos são ditos explícitos, e o processo de integração da função g(x) para obtenção dos v_i envolve apenas valores já calculados.

Devido à complexidade dos cálculos, os métodos implícitos possuem aplicações reduzidas (Gear, 1971). Os métodos explícitos são, entretanto, de grande simplicidade e muito usados.

O método explícito de Runge-Kutta de primeira ordem coincide com o método de Euler. Exemplifica-se aqui o caso p=3 usando coeficientes de integração de Newton-Cotes. Assim, têm-se:

$$v_{k+1} = v_k + \frac{(x_{k+1} - x_k)}{6} [g_1 + 4g_2 + g_3] ,$$

$$y_1 = x_k, \; y_2 = x_k + (x_{k+1} - x_k)/2, \; y_3 = x_{k+1} ,$$

$$u_1 = v_k ,$$

$$g_1 = f(x_k, v_k) ,$$

$$u_2 = v_k + \frac{(x_{k+1} - x_k)}{2} g_1 ,$$

$$g_2 = f \left[x_k + \frac{(x_{k+1} - x_k)}{2}, u_2 \right] ,$$

$$u_3 = v_k + \frac{(x_{k+1} - x_k)}{2} (g_1 + g_2) ,$$

$$g_3 = f[x_{k+3}, u_3] .$$

O valor calculado acima para v_{k+1} pode ser considerado como uma aproximação do valor exato $v(x_{k+1})$. A diferença entre esses valores é dada pelo erro do método de integração que, no caso, vale:

$$v(x_{k+1}) - v_{k+1} = -\frac{t^5}{90}\, g^{(4)}(\xi) ,$$

onde $t = (x_{k+1} - x_k)/2$ e ξ é o ponto pertencente ao intervalo (x_k, x_{k+1}). Pode-se constatar que esta diferença tende a zero para $t \to 0$, de onde resulta a convergência do método.

A estabilidade do método pode ser facilmente estudada supondo que o valor v_k é afetado por um erro δ e observando a propagação desse erro através das expressões para obtenção de v_{k+1}. O que resulta em:

$$\Delta g_1 \cong \delta f_v ,$$

$$\Delta u_2 \cong \delta + t\delta f_v ,$$

$$\Delta g_2 \cong (\delta + t\delta f_v)\, f_v ,$$

$$\Delta u_3 \cong \delta + t(2\delta f_v + t\delta f_v^2) ,$$

$$\Delta g_3 \cong (\delta + 2t\delta f_v + \delta t^2 f_v^2)\, f_v ,$$

$$\Delta_{v_{k+1}} \cong \delta + \delta \frac{tf_v}{3}\, [6 + t(\ldots)],$$

de onde se conclui que o processo só será estável nos trechos em que $f_v < 0$.

De acordo com o método de integração obtêm-se diferentes variantes do tipo Runge-Kutta; eventualmente, pode-se obter maior convergência ou estabilidade para uma mesma ordem. Nos exercícios (no fi

nal deste capítulo) o leitor encontrará algumas alternativas muito usadas para o método de terceira ordem aqui discutido.

2) Métodos de recorrência com vários pontos

Estes métodos baseiam-se na função interpoladora de Hermite para aproximar a função v(x). Por hipótese, supõe-se conhecida v(x) para um conjunto discreto com q pontos, ordenados de 1 até q. Estima-se o valor para um ponto, x_{q+1}, consecutivo de x_q, usando a função interpoladora de Hermite:

$$v_{q+1} \cong H(x_{q+1}) \ .$$

Estabelecida a função v'(x), pode-se determinar a primeira derivada em todos os pontos desejados. Dessa forma, no polinômio de Hermite (Seção 15.11.2) impõe-se $m_i=2$ para todo valor de i. Esse polinômio toma então a forma:

$$H(x_{q+1}) = \sum_{i=1}^{q} v(x_i)\ \psi_{0i}(x) + \sum_{i=1}^{q} v'(x_i)\ \psi_{1i}(x) \ ,$$

denominada explicita por só envolver valores conhecidos da função e sua derivada. Pode-se, entretanto, aumentar de uma unidade a segunda somatória da expressão de $H(x_{q+1})$, criando assim uma relação não-linear para a determinação de v_{q+1}. Esta segunda forma é denominada implícita e escreve-se:

$$H(x_{q+1}) = \sum_{i=1}^{q} v(x_i)\ \psi_{0i}(x) + \sum_{i=1}^{q+1} v'(x_i)\ \psi_{1i}(x) \ .$$

Utilizando os valores deduzidos na Seção 15.11.2, tem-se:

$$\psi_{0i}(x) = \frac{\pi_m(x)}{(x-x_i)^2} \left\{ \left[\frac{(x-x_i)^2}{\pi_m(x)} \right]_{x=x_i} + \left[\frac{(x-x_i)^2}{\pi_m(x)} \right]^{(1)}_{x=x_i} (x-x_i) \right\} ,$$

$$\psi_{1i}(x) = \frac{\pi_m(x)}{(x-x_i)} \left[\frac{(x-x_i)^2}{\pi_m(x)} \right]_{x=x_i} ,$$

com

$$\frac{\pi_m(x)}{(x-x_i)^2} = \prod_{\substack{j=1 \\ j\neq i}}^{q} (x-x_j)^2 ,$$

de onde:

$$\left[\frac{(x-x_i)^2}{\pi_m(x)} \right]^{(1)}_{x=x_i} = - \left[\sum_{\substack{s=1 \\ s\neq i}}^{q} 2(x_i-x_s) \prod_{\substack{j=1 \\ j\neq s \\ j\neq i}}^{q} (x_i-x_j)^2 \right] \left[\prod_{\substack{j=1 \\ j\neq i}}^{q} (x_i-x_j)^2 \right]^{-2} .$$

Pela sua simplicidade, o caso em que os pontos são igualmente espaçados por um valor h oferece maior interesse prático. Nestas condições, resulta para a forma explícita:

$$\psi_{0i}(x_{q+1}) = \frac{(q!)^2}{(q+1-i)^2} \left[\frac{1}{D_i} - \frac{2(q-1)}{D_i} \sum_{\substack{s=1 \\ s\neq i}}^{q} \frac{1}{i-s} \right] ,$$

$$\psi_{1i}(x_{q+1}) = h \frac{(q!)^2}{(q+1-i)} \frac{1}{D_i} ,$$

onde:

$$D_i = [(i-1)!\ (q-i)!]^2 ,$$

e para a forma implícita:

$$\psi_{0i}(x_{q+1}) = \frac{(q!)^2}{(q+1-i)^2}\left[\frac{1}{D_i} - \frac{2(q-1)}{D_i}\sum_{\substack{s=1\\ s\neq i}}^{q}\frac{1}{i-s}\right],$$

$$\psi_{1i}(x_{q+1}) = h\,\frac{[(q+1)!]}{(q+2-i)}\,\frac{1}{D_i'},$$

onde:

$$D_i' = [(i-1)!\ (q+1-i)!]^2 .$$

Pode-se notar que para pontos igualmente espaçados os coeficientes podem ser escritos na forma:

$$\psi_{0i}(x_{q+1}) = \alpha_i ,$$

$$\psi_{1i}(x_{q+1}) = h\,\beta_i ,$$

onde α_i e β_i são constantes univocamente determinadas pelo número de pontos q e pelo valor i. Obtém-se assim a fórmula:

$$v_{q+1} = \sum_{i=1}^{q}\alpha_i\, v_i + h\sum_{i=1}^{q+1}\beta_i\, f(x_i, v_i) ,$$

que engloba genericamente os casos implícito e explícito. Neste último, faz-se $\beta_{q+1} = 0$, uma vez que o ponto x_{q+1} não é considerado.

Alguns autores preferem considerar a última expressão como ponto de partida, impondo condições diversas para a determinação dos α_i e β_i. Assim, resultam múltiplas variantes para este método.

No desenvolvimento aqui adotado, a diferença $v(x_{q+1})-v_{q+1}$ é facilmente determinada pelo termo corretivo do polinômio interpolador de Hermite e vale:

$$v(x_{q+1}) - v_{q+1} = K\,\pi_m(x_{q+1}) ,$$

onde:

$$K = \frac{1}{(p+1)!}\, v^{(p+1)}(\xi) ,$$

sendo ξ um ponto pertencente ao intervalo (x_1, x_{q+1}) e $p = 2q-1$ (caso explícito) ou $p = 2q$ (caso implícito) o grau do maior polinômio representado exatamente pela aproximação de Hermite.

O valor p é chamado ordem do método de vários pontos.

No caso de pontos igualmente espaçados, obtém-se:

$$v(x_{q+1}) - v_{q+1} = h^{p+1}\, K(q!)^2 ,$$

o que mostra que a diferença tende a zero quando $h \to 0$, de onde resulta a convergência do método.

Para estudar a estabilidade do método explícito, considera-se um erro δ no valor v_q. Esse erro propaga-se para v_{q+1} pela relação:

$$\Delta v_{q+1} = \delta\,\alpha_q + \beta_q \Delta f(x_q, v_q + \delta) ,$$

que desenvolvida conduz a:

$$\Delta v_{q+1} \cong \delta(\alpha_q + \beta_q\, f_v) ,$$

de onde resulta a condição para estabilidade:

$$|\alpha_q + \beta_q f_v| < 1 ,$$

que pode ser examinada em cada caso em função do número de pontos usados.

Eventualmente, no caso em que as derivadas de ordem superior da função f[x, v(x)] sejam facilmente determinadas, é possível utilizar $m_i > 2$ no polinômio interpolador de Hermite. Este fato não tem sido muito explorado em vista de sua grande complexidade e restrita aplicação.

3) Métodos autocorretivos

Como em todos os tipos de equações, os métodos autocorretivos também ocupam posição de destaque no refinamento de soluções de equações diferenciais ordinárias. Apresenta-se aqui o método de previsão-correção que decorre naturalmente dos métodos de múltiplos pontos.

A aplicação sucessiva dos métodos explícitos pode acarretar uma indesejável propagação de erros. Por outro lado, os métodos implícitos requerem a solução de equações não-lineares, apresentando o problema de uma estimativa inicial razoável para a obtenção rápida do resultado. Surgiu então a idéia de adoção de métodos explícitos para uma *previsão* do valor desejado, adotando métodos implícitos para *correção* do resultado. Assim, aparecem os métodos de previsão-correção de grande utilidade na solução de equações diferenciais ordinárias.

A implementação desses métodos é relativamente simples. Utilizando uma fórmula explícita, faz-se a previsão do valor v_{q+1}. Em seguida, emprega-se uma fórmula implícita que pode ser simbolicamente expressa por:

$$v_{q+1} = \phi\ (v_{q+1}) ,$$

cuja solução é obtida através de um esquema iterativo da forma:

$$(v_{q+1})_{k+1} = \phi[(v_{q+1})_k] \ ,$$

com o valor inicial fornecido pela previsão da fórmula explícita.

A convergência do método depende exclusivamente da convergência do esquema iterativo resultante da fórmula implícita.

A estabilidade do método depende da estabilidade da fórmula implícita.

Na literatura, o leitor encontrará os métodos explícitos estáveis relacionados com os nomes Adams-Bashforth, e os métodos implícitos estáveis com os nomes Adams-Moulton.

22.2.3 - EXTENSÃO PARA SISTEMAS DE EQUAÇÕES

Para a solução de um sistema de equações diferenciais ordinárias, pode-se considerar cada equação isoladamente, desde que se resolvam todas as equações simultaneamente.

A título de exemplo, considera-se a solução do sistema utilizando o método de Euler. Escreve-se o resultado, simbolicamente, em notação vetorial como:

$$\underline{v}_{k+1} = \underline{v}_k + h\ \underline{f}[x_k, \underline{v}_k] \ ,$$

e cada equação componente será resolvida recursivamente por:

$$(v_i)_{k+1} = (v_i)_k + h\ f_i[x_k, \underline{v}_k] \ .$$

A interligação das equações é estabelecida pelas dependências funcionais $f[x_k, \underline{v}_k]$. O sistema de equações dessa forma é denominado sistema de equações dependentes. Se entretanto $f_i[x_k, \underline{v}_k] =$

$= f_i[x_k, (v_i)_k]$, tem-se um sistema de equações independentes. Neste último caso não é necessária a resolução simultânea de todas as equações.

22.3 - PROBLEMAS DE VALOR DE CONTORNO

Estes problemas são caracterizados pela equação diferencial:

$$\underline{v}'(x) = \underline{f}[x, \underline{v}(x)] \ ,$$

sendo a solução particularizada por uma dependência implícita da forma:

$$\underline{g}[\underline{v}(\underline{x}_p), \underline{v}'(\underline{x}_p)] = \underline{0} \ ,$$

onde $\underline{x}_p$ é um vetor cujas componentes são valores particulares de x.

Os problemas de valor de contorno podem ser divididos em dois tipos:

i) problemas lineares para os quais pode-se escrever:

$$\underline{f}[x, \underline{v}(x)] = \underline{\underline{A}}(x)\underline{v}(x) + \underline{z}(x) \ ,$$

ii) problemas não-lineares para os quais a função f não pode ser decomposta no produto matricial acima.

Os métodos para solução destes problemas são englobados em três classes principais:

a) métodos baseados na aproximação local da solução,

b) métodos baseados na aproximação global da solução,

c) métodos baseados nas técnicas de tratamento dos problemas de valor inicial.

Os dois primeiros métodos transformam o sistema de equações diferenciais num sistema de equações algébricas, lineares ou não-lineares, que podem ser resolvidas com as técnicas dos Capítulos 20 e 21. O último método utiliza as formas de solução de problemas de valor inicial. As análises comparativas (Jacobs, 1977) têm apresentado ligeira vantagem deste último tipo sobre os demais.

22.3.1 - MÉTODOS DE APROXIMAÇÃO LOCAL

Existem dois tipos principais de métodos para aproximação local:

a) aproximação para um conjunto de pontos,

b) aproximação para um conjunto de subintervalos.

1) Aproximação para um conjunto de pontos

Este método utiliza o cálculo com diferenças finitas. Cada equação é calculada para um conjunto de pontos x_i, $i = 1, \ldots, n$. A forma mais simples para cada equação é:

$$(v_j)_{i+1} - (v_j)_i = h_i \, f_j \left\{ x_{i+\frac{1}{2}}, \frac{1}{2} [(\underline{v})_i + (\underline{v})_{i+1}] \right\},$$

onde j é o índice que designa a equação considerada, h_i é a amplitude do intervalo $x_{i+1} - x_i$ e $x_{i+\frac{1}{2}} = (x_i + x_{i+1})/2$.

O sistema acima fornece n-1 equações. Para complementá-lo usa-se a equação de contorno:

$$g_j \{\underline{v}(\underline{x}_p), \underline{f}[\underline{x}_p, \underline{v}(\underline{x}_p)]\} = 0 .$$

O espaçamento entre pontos h_i foi deixado propositadamente arbitrário para que o conjunto de pontos, cujos valores são as componentes de $\underline{x}_p$, estivesse contido no conjunto de pontos $\{x_i\}$.

A interligação entre as equações em i e j é estabelecida pela dependência vetorial em $\underline{v}(x)$ das funções f_j e g_j.

No caso geral não-linear, o sistema em i e j pode tornar-se extremamente complexo, devendo-se recorrer a métodos iterativos para sua solução, os quais envolvem Nn variáveis.

No caso em que f e g são lineares em v(x), o problema simplifica-se bastante, pois as condições de contorno reduzem-se a um sistema de equações algébricas lineares da forma:

$$\sum_{i=1}^{n} c_{ij}\, v_j(x_i) = 0\,, \qquad j = 1, \ldots, N\,,$$

que permite a determinação de N incógnitas: $v_j(x_i) = (v_j)_i$. Ainda neste caso restam N(n-1) incógnitas, o que pode acarretar problemas adicionais no caso de grandes sistemas.

Uma alternativa é a não-redução de equações de ordem superior a subsistemas de primeira ordem. Isto é conseguido utilizando diferenças finitas de ordem superior. Empregando os operadores lineares do Capítulo 14, mostra-se que para pontos igualmente espaçados:

$$2D \cong \Delta + \nabla\,,$$

$$D^2 \cong \Delta - \nabla\,,$$

e assim por diante. Desta forma, cada equação independente de ordem superior pode ser resolvida apenas por um sistema com n incógnitas. Na opção precedente, o número de incógnitas é n-1 vezes a ordem da equação diferencial.

Considerando uma equação diferencial linear de ordem superior, obtém-se um sistema de equações lineares algébricas da forma:

$$\underline{\underline{A}}\ (\underline{x}_I)\ \underline{v}_I = \underline{b}\ ,$$

onde:

$$\underline{x}_I = (x_1, \ldots, x_n)\ ,$$

$$\underline{v}_I = (v_1, \ldots, v_n)\ ,$$

$$v_i = v(x_i).$$

Quando as condições de contorno são fornecidas numa forma implícita, prefere-se, por vezes, optar por resolver esse sistema de equações, lineares ou não-lineares, obtendo N valores:

$$v_j(x_{pj}) = \alpha_j\ , \qquad j = 1, \ldots, N\ ,$$

x_{pj} valores particulares de x. Este processo acarreta um aumento de erros, mas simplifica consideravelmente a dificuldade de automação do processo para a solução de problemas com condições de contorno.

Outra forma de simplificação consiste na linearização das equações. Considerando o método de diferenças finitas aplicado ao sistema de equações diferenciais, tem-se:

$$\underline{v}_{i+1} - \underline{v}_i = h\ \underline{f}[x_{i+1/2},\ 1/2(\underline{v}_i + \underline{v}_{i+1})]\ ,$$

que, linearizado usando o desenvolvimento de Taylor em sua forma de diferenças finitas, conduz a:

$$v_{i+1} - \underline{v}_i \cong h(1+D/2)\ \underline{f}(x_i, \underline{v}_i)\ .$$

Como $D \cong (\Delta + \nabla)/2$ e

$$\underline{f}(x_i, \underline{v}_i) = \underline{v}_i' \cong \frac{1}{2h}\ (\underline{v}_{i+1} - \underline{v}_{i-1})\ ,$$

segue-se que:

$$\underline{v}_{i+1} - \underline{v}_i = (1/8)(4 + \Delta + \nabla)(\underline{v}_{i+1} - \underline{v}_{i-1}) ,$$

de onde:

$$\underline{v}_{i+2} - 4\underline{v}_{i+2} + 6\underline{v}_i - 4\underline{v}_{i-2} + \underline{v}_{i-2} = 0 .$$

O sistema assim formulado exige 4N restrições adicionais. Como as condições de contorno fornecem N incógnitas, 3N restrições ficam arbitrárias. Pode-se completar o sistema, por exemplo, com:

$$\underline{v}_{-1} = \underline{0}, \quad \underline{v}_0 = \underline{0} \quad \text{e} \quad \underline{v}_{n+2} = \underline{0} .$$

O processo de linearização também acarreta considerável aumento de erros se o espaçamento h não for bastante pequeno. A diminuição desse espaçamento, por outro lado, implica um aumento do número de equações, com os inconvenientes já discutidos no Capítulo 21.

2) Aproximação para um conjunto de subintervalos

Antes de passar ao desenvolvimento propriamente dito desta parte, é interessante tecer algumas considerações iniciais sobre os fundamentos que motivaram estes métodos.

Primeiramente, é necessário lembrar algumas conclusões importantes dos Capítulos 16, 17 e 18, tais como:

a) os processos de derivação numérica implicam necessariamente uma amplificação dos erros introduzidos pelo computador;

b) os métodos de integração numérica apresentam grande confiabilidade do ponto de vista computacional, não sendo tão drasticamente afetados por erros de arredondamento ou truncamento;

c) a teoria das aproximações fornece os elementos para minimização dos erros nos métodos numéricos diretos.

Com base nessas conclusões e tendo em vista o objetivo de reduzir ao máximo os erros computacionais, surgiu a idéia de converter a solução de equações diferenciais num problema dentro do campo dos métodos diretos. O problema assim convertido deve envolver técnicas de integração e otimização, evitando os inconvenientes da derivação numérica e dos métodos de diferenças finitas.

Destacam-se dois processos para converter a resolução de uma equação diferencial num problema de integração numérica:

i) Formulação do problema em termos do cálculo variacional, (método de Rayleigh-Ritz).

ii) Formulação do problema em termos de minimização de erros, utilizando o critério dos resíduos ponderados (e.g. método de Galerkin). Estes processos serão discutidos em mais detalhes no capítulo seguinte. Por ora, considera-se que o problema da resolução da equação diferencial é convertido no problema de minimização de integral:

$$\underline{I} = \int_a^b \underline{F}[x, \underline{v}(x), \underline{v}'(x)]\, dx$$

no método de Rayleigh-Ritz, ou no problema de ortogonalização de erro:

$$\int_a^b \underline{R}[x, \underline{v}(x)\ \underline{v}'(x)]\ \phi_i(x)\, dx = 0$$

no método de Galerkin.

Em quaisquer dos dois casos, cada componente do vetor $\underline{v}(x)$ é expressa como combinação linear de um conjunto de funções $\phi_i(x)$ com ϕ_i não-nula para $x \in (x_i, x_{i+1})$; então:

$$v_j(x) = \sum_{i=1}^{m} b_{ji}\, \phi_i(x) \ ,$$

onde m é o número de subintervalos considerados, e as funções $\phi_i(x)$ são definidas localmente, escolhidas de modo que todos os $v_j(x)$ satisfaçam as condições de contorno.

Para ambos os métodos obtém-se um sistema de Nm equações, cujas incógnitas são os b_{ji}. No método de Rayleigh-Ritz, essas equações são geradas pelas condições de minimização:

$$\frac{\partial I_j}{\partial b_{ji}} = 0$$

e, no método de Galerkin, pelas condições de ortogonalização do erro:

$$\int_a^b R_j\,[x, \underline{b}_j]\, \phi_i(x)dx = 0 \ ,$$

onde $\underline{b}_j$ é o vetor de componentes b_{ji}.

As integrais envolvidas na formação do sistema de equações são, em geral, resolvidas pelos métodos de integração numérica.

Pode-se notar a semelhança entre esta forma de solução e a precedente, no sentido de que o sistema de equações diferenciais é reduzido a um sistema de equações algébricas.

22.3.2 - MÉTODOS DE APROXIMAÇÃO GLOBAL

Os métodos de aproximação global fundamentam-se no mesmo princípio que deu origem aos métodos de elementos finitos. Eles também serão discutidos em detalhes no capítulo seguinte. Para o presente caso, a diferença essencial consiste em que as funções $\phi_i(x)$ não são definidas localmente, mas sim para todo o intervalo (a, b). De certa forma isto acarreta um aumento de complexidade dessas funções, que agora devem pertencer a um conjunto de polinômios ortogonais como os de Legendre, Tchebyshev, etc.

Os métodos para determinação dos coeficientes são essencialmente os mesmos expostos na última seção: de Rayleigh-Ritz e de Galerkin.

As características de convergência e estabilidade são também as mesmas dos métodos de elementos finitos.

A aproximação global pode fornecer bons resultados quando as $v_j(x)$ são funções contínuas e com derivadas contínuas. Quando isto não ocorre, em geral as funções $v_j(x)$ não podem ser desenvolvidas em séries como as de Taylor, Tchebyshev, etc. Neste caso, a hipótese fundamental deste tipo de abordagem não se verifica, devendo-se então optar por outro tipo de solução.

22.3.3 - MÉTODOS DE REDUÇÃO A PROBLEMAS DE VALOR INICIAL

No decorrer da Seção 22.3.1, viu-se que, resolvendo o sistema de equações das condições de contorno, os problemas deste tipo poderiam ser formulados pela equação:

$$\underline{v}'(x) = \underline{f}[x, \underline{v}(x)] ,$$

com as condições de contorno na forma:

$$v_j(x_{pj}) = \alpha_j , \qquad j = 1, \ldots, N .$$

Neste aspecto há uma estreita semelhança dos problemas de condições de contorno com os problemas de valor inicial. De fato, se todos os x_{p_j} coincidirem tem-se caracterizado um problema de valor inicial. No caso de x_{p_j} distintos, procura-se calcular o valor aproximado das condições de contorno para um ponto comum x_0 desejado.

Os métodos de solução por redução a problemas de valor inicial envolvem 3 etapas:

a) obtenção de aproximação das condições de contorno para um ponto x_0, o que resulta em:

$$v_j(x_0) \cong s_j, \qquad j = 1, \ldots, N;$$

b) refinamento do valor obtido $\underline{v}(x_0)$;

c) resolução do problema de valor inicial obtido.

Alguns autores dispensam a primeira etapa do processo, admitindo uma coleção de valores inteiramente arbitrários $\underline{s}$ ("shooting").

1) Obtenção do vetor $\underline{s}$

Utilizando o desenvolvimento de Taylor pode-se escrever:

$$v_j(x_0) \cong \sum_{k=0}^{m} \frac{(x_0 - x_{p_j})^k}{h!} \; v_j^{(k)}(x_{p_j}) = s_j \;,$$

onde m é a ordem da maior derivada considerada. Os valores $v_j^{(k)}(x_{p_j})$ são obtidos analiticamente a partir da equação diferencial.

Nos casos mais complexos, não é possível a utilização de derivadas de ordem superior. Utiliza-se então somente a primeira derivada, obtendo um valor inicial pouco preciso.

2) Refinamento do vetor $\underline{v}(x_0)$ e resolução do problema

O processo para refinamento do "valor inicial", $\underline{v}(x_0)$, requer que outras condições iniciais sejam adotadas para um ponto diferente de x_0. Seja x_f este ponto; então resulta em:

$$v_j(x_f) \cong \sum_{k=0}^{m} \frac{(x_f - x_{p_j})^k}{k!} \, v_j^{(k)}(x_{p_j}) = q_j \, .$$

Em seguida, procede-se à integração da equação diferencial com base em dois conjuntos diferentes de "condições iniciais". Desta forma, obtêm-se para qualquer ponto x, do intervalo (a, b), duas soluções:

$$\underline{v}(x,\underline{s}) \quad \text{para} \quad \underline{v}(x_0) = \underline{s}$$

e

$$\underline{v}(x,\underline{q}) \quad \text{para} \quad \underline{v}(x_f) = \underline{q}.$$

Em geral essas soluções são diferentes devido aos erros em $\underline{v}(x_0)$ e $\underline{v}(x_f)$. Impondo a condição de identidade das soluções, obtêm-se um sistema de equações que fornece a informação necessária para aprimorar os "valores iniciais". Considerando dois pontos distintos, x_1 e x_2, no intervalo (a,b), resulta o sistema com 2N equações e 2N incógnitas:

$$\underline{v}(x_1, \underline{s}) - \underline{v}(x_1, \underline{q}) = \underline{0} \, ,$$

$$v(x_2, \underline{s}') - v(x_2, \underline{q}') = \underline{0} \, ,$$

onde $\underline{s}' = \underline{s} + \Delta\underline{s}$ e $\underline{q}' = \underline{q} + \Delta\underline{q}$.

Utilizando o desenvolvimento de Taylor, pode-se escrever:

$$\underline{v}(x, \underline{s}') \cong \underline{v}(x, \underline{s}) + \underline{v}_{\underline{s}}(x, \underline{s}) \, \Delta\underline{s} \, ,$$

onde $\underline{v}_{\underline{s}}(x,\underline{s})$ é a matriz das derivadas parciais de $\underline{v}$ com relação a $\underline{s}$, calculada no ponto $(x,\underline{s})$. A mesma aproximação é usada para $\underline{v}(x,\underline{q}')$. Então, tem-se um sistema dependente em $\Delta\underline{s}$ e $\Delta\underline{q}$ que poderá fornecer essas correções necessárias para o refinamento de $\underline{v}(x_0)$ e $\underline{v}(x_f)$.

As matrizes das derivadas parciais $\underline{v}_{\underline{s}}(x,\underline{s})$ e $\underline{v}_{\underline{q}}(x,\underline{q})$ devem ser calculadas numericamente, uma vez que não se dispõe da expressão analítica de $\underline{v}(x)$. Usam-se para isto as aproximações:

$$\frac{\partial v_j(x,\underline{s})}{\partial s_i} \cong \frac{v_j(x,\underline{s}+\delta\, s_i) - v_j(x,\underline{s})}{\delta s_i}, \qquad i, j = 1, \ldots, N.$$

Em geral, os valores δs_i são escolhidos como uma percentagem fixa do valor s_i.

Consegue-se uma drástica redução no número de equações diferenciais de valor inicial, que devem ser resolvidas, pela escolha conveniente dos pontos x_1 e x_2. Uma possível escolha é $x_1 = x_0$ e $x_2 = x_f$. Por sua vez, x_0 e x_f são escolhidos de forma a coincidir com dois valores x_{pj} para reduzir o número de incógnitas.

A convergência e estabilidade do método dependem de dois fatores:

a) convergência e estabilidade do sistema de equações para refinamento do valor inicial;

b) convergência e estabilidade dos métodos para obtenção de $\underline{v}(x,\underline{s})$, $\underline{v}(x,\underline{q})$, $\underline{v}_{\underline{s}}(x,\underline{s})$ e $\underline{v}_{\underline{q}}(x,\underline{q})$.

O método aqui apresentado aplica-se a qualquer sistema de equações com condição de contorno. Para sistemas lineares, o leitor encontrará na bibliografia outras soluções alternativas.

22.3.4 - CONVERGÊNCIA E ESTABILIDADE

A convergência e estabilidade dos métodos de aproximação local podem ser facilmente analisadas. No caso de diferenças finitas, por exemplo, a equação básica é:

$$\underline{v}_{i+1} - \underline{v}_i = h\, f[x_{i+1/2},\ 1/2(\underline{v}_i + \underline{v}_{i+1})] \ ,$$

que mostra que para $h \to 0$, $\underline{v}_{i+1} \to \underline{v}_i$, o que implica na convergência do método. A condição de estabilidade, neste caso, é obtida impondo um erro $\underline{\delta}$ em $\underline{v}_i$ e calculando o erro em $\underline{v}_{i+1}$, o que resulta em:

$$\underline{v}_{i+1} = \underline{\delta} + \frac{h}{2}\, \underline{f}_{\underline{v}}\, \underline{\delta} \ ,$$

onde $\underline{f}_{\underline{v}}$ é a matriz das derivadas parciais da função $\underline{f}$ calculadas no ponto $(x_i, \underline{v}_i)$; então tem-se:

$$\left|\left(\underline{\underline{I}} + \frac{h}{2}\, \underline{f}_{\underline{v}}\right) \underline{\delta}\right| < \left|\underline{\underline{I}}\ \underline{\delta}\right|$$

como condição de estabilidade.

No caso de elementos finitos, a convergência também é garantida, pois as funções interpoladoras coincidem com o valor exato dos pontos conhecidos. A estabilidade depende de erros δ_{ji} nos valores b_{ji}. Para o método de Rayleigh-Ritz, a condição de insensibilização $\partial I_j / \partial b_{ji} = 0$ assegura a estabilidade. Para o método de Galerkin, a condição de ortogonalidade do erro faz com que o erro do elemento b_{ji} não se transfira para outros elementos, garantindo assim a estabilidade, desde que as $\phi_i(x)$ constituam um conjunto de funções ortogonais. Sob este aspecto, os métodos com elementos finitos apresentam pequena vantagem sobre os métodos de diferenças finitas.

22.4 - PROBLEMAS DE AUTOVALORES

Estes problemas consistem na necessidade de determinar os invariantes de uma particular transformação diferencial T, envolvendo derivadas até a ordem m.

As invariantes de uma transformação funcional são funções v(x) para as quais a transformação não altera a forma, mas atua apenas como um fator "peso" λ para essa função. Dentro do contexto de espaço vetorial abstrato (Capítulo 3), interpretada a função v(x) como um vetor $\underline{v}$, a transformação diferencial T como uma transformação vetorial $\underline{T}$ resulta em:

$$\underline{T}(\underline{v}) = \lambda \underline{v} \ ,$$

que caracteriza uma alteração no "módulo", mas não afeta a "direção" do vetor $\underline{v}$. O vetor $\underline{v}$ é chamado autovetor e, por analogia, a função v(x) autofunção. O número λ é denominado autovalor.

Em termos de relacionamento funcional pode-se escrever:

$$T[x, \underline{y}(x)] = \lambda \, v(x) \ ,$$

onde $y_i(x) = v^{(i-1)}(x)$, $\qquad i = 1, \ldots, m+1$.

O problema é particularizado pela informação adicional fornecida pelas condições de contorno:

$$\underline{g}[\underline{y}(\underline{x}_p)] = 0 \ .$$

A transformação diferencial T é um mapeamento de um espaço de infinitas dimensões em outro de mesma natureza e admite infinitos autovalores e autovetores.

O problema pode também ser colocado na forma:

$$\underline{z}'(x) = \underline{S}[x, \lambda, \underline{z}(x)]$$

por uma conveniente transformação, conforme exemplificado anteriormente.

22.4.1 - MÉTODO DE DIFERENÇAS FINITAS

Neste método, a função $v(x)$ é considerada para um conjunto discreto com n pontos distintos. Forma-se assim um sistema com n equações do tipo:

$$T[x_j, \underline{y}(x_j)] = \lambda\, v(x_j) \ ,$$

onde os $y_i(x_j)$ são calculados pelo método de diferenças finitas, envolvendo um total de $n+m$ pontos. Os m pontos complementares são fornecidos pelas m condições de contorno.

Fazendo $n = 1, 2, \ldots$ tem-se, respectivamente, aproximações para 1, 2, ... autovalores, desde que se verifique a condição de $v(x)$ não ser identicamente nula. Quanto maior o valor de n, tanto melhor será a aproximação.

No caso linear, obtém-se um sistema algébrico de equações lineares na forma:

$$\underline{\underline{A}}(\underline{x}_I)\ \underline{v}_I = \lambda\, \underline{v}_I \ ,$$

onde:

$$\underline{x}_I = (x_1, \ldots, x_n) \ ,$$

$$\underline{v}_I = (v_1, \ldots, v_n) \ ,$$

$$v_i = v(x_i) \ .$$

Neste caso, os autovalores procurados são dados aproximadamente pelos autovalores da matriz $\underline{\underline{A}}(\underline{x}_i)$ e os autovetores da transformação, pelos autovetores desta matriz.

Para problemas não-lineares, uma alternativa é a linearização, a partir da qual determina-se uma primeira aproximação doa autovalores. Em seguinda, utiliza-se o método variacional, que será visto na próxima seção, para refinamento da solução.

Para aplicação do método de diferenças finitas é aconselhável a prévia resolução das condições de contorno para obtenção de relações do tipo:

$$y_i(x_{p_i}) = \beta_i \quad , \qquad i = 1, \ldots, m \ ,$$

e, a partir destas, utilizando o método de diferenças finitas, chega-se finalmente a:

$$v(x_{q_i}) = \alpha_i \ , \qquad i = 1, \ldots, m \ .$$

Exemplos simples de aplicação deste método podem ser encontrados em Szidarovszky e Yakowitz (1978), Young e Gregory (1972), etc.

22.4.2 - MÉTODO VARIACIONAL

Este método deve-se a Rayleigh-Ritz (Friedman, 1966; Courant and Hilbert, 1966). Trata-se de um processo autocorretivo que permite o refinamento de autovalores e autovetores.

Partindo da equação básica:

$$\underline{T}(\underline{v}) = \lambda \underline{v} \ ,$$

onde $v_i = v(x_i)$, pode-se escrever:

$$\langle \underline{v}, \underline{T}(\underline{v}) \rangle = \lambda \langle \underline{v}, \underline{v} \rangle \ .$$

Em seguida, aproxima-se v(x) pela combinação linear de um conjunto de funções $\phi_i(x)$, definidas no intervalo (a,b) e satisfazendo as condições de contorno. Então tem-se:

$$v(x) \cong \sum_{i=1}^{n} c_i \, \phi_i(x) = \tilde{v}(x) \ .$$

Tomando uma aproximação λ_k para o autovalor, resulta um erro dado por:

$$E(\lambda_k, \underline{c}) = \langle \underline{\tilde{v}}, \underline{T}(\underline{\tilde{v}}) \rangle - \lambda_k \langle \underline{\tilde{v}}, \underline{\tilde{v}} \rangle \ ,$$

onde $\underline{c}$ é o vetor cujas componentes são os coeficientes c_i.

Os coeficientes c_i são determinados pela condição de mínimo erro:

$$\frac{\partial E(\lambda_k, \underline{c})}{\partial c_i} = 0 \ , \qquad i = 1, \ldots, n \ .$$

Pode-se então obter uma melhor aproximação, λ_{k+1}, para o autovalor impondo a condição de que o erro $E(\lambda_k, \underline{c})$ seja nulo. Assim resulta em:

$$\lambda_{k+1} = \frac{\langle \underline{\tilde{v}}, \underline{T}(\underline{\tilde{v}}) \rangle}{\langle \underline{\tilde{v}}, \underline{\tilde{v}} \rangle} \ .$$

Repetindo o processo, novos valores c_i serão obtidos e assim recursivamente até que a diferença $|\lambda_{k+1} - \lambda_k|$ esteja dentro de uma tolerância, ε, especificada.

Uma primeira aproximação, λ_0, para o autovalor pode ser obtida utilizando o método anterior.

Exemplos analíticos, ilustrativos do método, para o caso linear podem ser encontrados em Friedman (1966).

22.4.3 - CONVERGÊNCIA E ESTABILIDADE DOS MÉTODOS

Para o método de diferenças finitas, a precisão dos resultados depende diretamente do número n de pontos. Aumentando n há uma diminuição do espaçamento entre pontos, o que implica convergência do método. A estabilidade depende do sistema algébrico em que foi convertido o problema. Em geral, para $|\lambda| < 1$, é esperada a estabilidade do método, pois:

$$\Delta T = \lambda \Delta v$$

não implica amplificação do erro.

No método variacional, o erro $|v(x) - \tilde{v}(x)|$ também depende do valor de n. Aumentando n, $\tilde{v}(x) \to v(x)$, o que implica convergência do método. Eventualmente, o erro pode anular-se mesmo para n finito (ver Capítulo 18). A convergência do autovalor é mais rápida que a do autovetor (Friedman, 1966), pois:

$$|\Delta\lambda| \propto |\Delta v|^2 \ .$$

A estabilidade do método é, de certa forma, assegurada pela condição de insensibilidade do erro:

$$\frac{\partial E(\lambda_k, \underline{c})}{\partial c_i} = 0 \ , \qquad i = 1, \ldots, n \ .$$

EXERCÍCIOS

1. Discuta o método clássico de Runge-Kutta de terceira ordem, para o qual são usados:

$y_1 = x_k$, $y_2 = x_k + t$;

$y_3 = x_{k+1} = x_k + 2t$;

$c_1 = \frac{1}{3}$, $c_2 = \frac{2}{3}$, $c_3 = \frac{1}{3}$;

$a_1 = \frac{1}{2}$ para u_2 ;

$a_1 = -1$, $a_2 = 2$ para u_3 .

Compare os resultados com o exemplo do item 1 da Seção 22.2.2.

2. Discuta a variante de Heun para o método de Runge-Kutta de terceira ordem. Nessa variante são usados:

$y_1 = x_k$, $y_2 = x_k + 2t/3$;

$y_3 = x_k + 4t/3$;

$c_1 = \frac{1}{4}$, $c_2 = 0$, $c_3 = \frac{3}{4}$;

$a_1 = \frac{1}{3}$ para u_2 ;

$a_1 = 0$, $a_2 = \frac{2}{3}$ para u_3 .

Compare esta variante com a do exercício anterior e com a do texto.

3. Deduza as fórmulas de $\psi_{0i}(x_{q+1})$ e $\psi_{1i}(x_{q+1})$ do item 2 da Seção 22.2.2 para pontos igualmente espaçados nos casos explícito e implícito.

4. Determine os coeficientes α_i, β_i da expressão de v_{q+1} do item 2 da Seção 22.2.2 para o método explícito com q=3. Utilize o desenvolvimento de Taylor.

5. A equação para o cálculo de estabilidade do método implícito com múltiplos pontos escreve-se:

$$\Delta v_{q+1} = \delta\alpha_q + \beta_q \Delta f(x_q, v_q + \delta) + \beta_{q+1} \Delta f(x_{q+1}, v_q + \Delta v_{q+1}) .$$

Determine as condições para a estabilidade do método.

6. Utilizando o resultado do exercício anterior mostre que, se $\beta_q >> \alpha_q$ e $\beta_q >> 1$, a estabilidade do método implícito com múltiplos pontos independe da função f[x, v(x)].

7. Utilizando o resultado do exercício anterior sugira um método implícito absolutamente estável para qualquer função.
Sugestão: fixando um número de pontos e os valores convenientes de α_q, β_q e β_{q+1}, determine os outros coeficientes para que a fórmula:

$$v(x) = \sum_{i=1}^{q} \alpha_i \, v(x) + h \sum_{i=1}^{q+1} \beta_i \, f[x, v(x)]$$

seja exata nos pontos $x = x_i$, $i = 1, \ldots, q$, e sua derivada coincida no maior número de pontos possíveis próximos a x_{q+1}.

8. Determine as condições de estabilidade dos métodos implícitos obtidos no polinômio interpolador de Hermite para $q = 1, 2, 3, 4, 5$.

9. Estude o método de previsão-correção baseado na fórmula de previsão de Adams-Bashforth:

$v_3 = v_2 + (h/2)\ [3f(x_2, v_2) - f(x_1, v_1)]$

e na fórmula de correção de Adams Moulton:

$v_3 = v_2 + (h/2)\ [f(x_3, v_3) + f(x_2, v_2)]$.

10. Escreva um programa de computador para resolver numericamente a equação diferencial:

$v''(x) = -16\, v(x)$

no intervalo $(0, 2\pi)$, com as condições de contorno:

$v(0) = 0,\ v'(0) = 1.$

Utilize:

a) o método de diferenças finitas na forma de sistema de equações de primeira ordem com $h = \pi/8$,

b) o método de diferenças finitas com diferenças de ordem superior para o mesmo espaçamento do item a.

11. Analise os resultados do exercício anterior do ponto de vista de rapidez de processamento e precisão da solução.

12. Utilize o método de diferenças finitas e o processo de linearização para resolver numericamente a equação não-linear:

$v''(x) + \text{sen}\ [v(x)] = 0$,

com as condições de contorno:

$v(0) = 0,\ v(1) = 1$

e espaçamento $h=(1/8)$. Analise o resultado.

13. Use a equação de Euler-Lagrange (Chung, 1978) para mostrar que o mínimo da integral:

$$I = \int_0^{2\pi} [v'^2(x) - 16v^2(x)]\, dx$$

ocorre para:

$$v''(x) + 16v(x) = 0 .$$

Baseado nesta relação, empregue o método de Rayleigh-Ritz para resolver a equação diferencial do Exercício 10. Empregue intervalos de amplitude $\pi/4$ e funções:

$$\phi_i(x) = a_i\, x^2 + b_i\, x + c_i ,$$

válidas para cada intervalo e nulas fora dele.

14. No método de Galerkin, utilizando a aproximação:

$$v(x) = \sum_{i=1}^{8} \beta_i\, \phi_i(x)$$

e definindo o resíduo por:

$$R(x, \underline{\beta}) = \tilde{v}''(x) + 16\tilde{v}(x) ,$$

resolva a equação diferencial do Exercício 10. Empregue as mesmas funções $\phi_i(x)$ do exercício anterior.

15. Para o método de aproximação global considera-se um conjunto de funções:

$$\phi_i(x) = \text{sen } (i-1)\, x.$$

Fazendo:

$$\tilde{v}(x) = \sum_{i=1}^{8} \beta_i \, \phi_i(x)$$

e definindo o resíduo como no exercício anterior, resolva a equação diferencial do Exercício 10 pelo critério de Galerkin. Compare os resultados dos Exercícios 13, 14 e 15.

16. Reduza o problema de condições de contorno do Exercício 10 a um problema de valor inicial. Resolva este problema por qualquer dos métodos aplicáveis no caso. Utilize $h = \pi/8$.

17. No intervalo $(0,2\pi)$ caracteriza-se um problema de autovalor pela relação:

$$-v''(x) = \lambda \, v(x) \, ,$$

com as condições de contorno $v(0) = 0$ e $v(2\pi) = 0$. Encontre as aproximações para alguns autovalores pelo método de diferenças finitas. Em seguida, utilize o método variacional para aprimorar o resultado.

Sugestão: Empregue 8 pontos no método de diferenças finitas e funções periódicas de período 2π no método variacional. As funções periódicas são sugeridas pelas condições de contorno (ver Capítulo 15).

BIBLIOGRAFIA

BAKHVALOV, N.S. *Numerical methods*. Moscou, Mir, 1975.

CHUNG, T.J. *Finite element analysis in fluid dynamics*. New York, McGraw-Hill, 1978.

COURANT, R.; HILBERT, D. *Methods of mathematical physics*. New York, Interscience Publishers, 1966. v. 1.

FRIEDMAN, B. *Principles and techniques of applied mathematics*. New York, Wiley, 1966.

GEAR, C.W. *Numerical initial value problems in ordinary differential equations*. Englewood Cliffs, Prentice-Hall, 1971.

HENRICI, P. *Discrete variable methods in ordinary differential equations*. New York, John Wiley, 1962.

JACOBS, D. *The state of the art in numerical analysis*. London, Academic Press, 1977.

KELLER, H. *Numerical methods for two-point boundary - value problems*. Waltham, MA, Blaisdell, 1968.

MARCHOUK, G. *Methodes de calcul numérique*. Moscou, Mir, 1977.

SZIDAROVSZKY, F.; YAKOWITZ, S. *Principles and procedures of numerical analysis*. New York, Plenum Press, 1978.

YOUNG, D.; GREGORY, R. *A survey of numerical mathematics*. Massachusetts, Addison Wesley, Reading, 1972. v. 2.

CAPÍTULO 23

EQUAÇÕES DIFERENCIAIS PARCIAIS

23.1 - INTRODUÇÃO

A solução numérica de problemas com equações diferenciais parciais constitui certamente uma das mais difíceis tarefas em Análise Numérica. O assunto é muito vasto para admitir uma abordagem sintética satisfatória. O material aqui apresentado pretende unicamente fornecer ao leitor uns poucos elementos auxiliares para a compreensão de textos mais avançados.

Basicamente, tem-se uma equação parcial quando um operador diferencial parcial D aplicado a uma função f produz uma função v. Simbolicamente, este relacionamento é expresso por:

$$D\, f = v \ ,$$

onde f pertence a um particular domínio Ω. As funções envolvidas são, em geral, escalares de variável vetorial $\underline{x}$. Quando o espaço vetorial em que $\underline{x}$ é definido envolve a variável tempo, o problema é chamado não-estacionário. Quando a variável tempo não compõe o espaço vetorial, o problema é denominado estacionário.

O problema assim formalizado admite infinitas soluções f. Para torná-lo um problema unívoco, são impostas limitações sobre a função f.

Analogamente ao caso de equações diferenciais ordinárias, caracterizam-se no presente caso três tipos de problemas:

a) quando os valores da função e os de suas derivadas, até a ordem máxima presente no operador D, são conhecidos para um particular ponto $\underline{x}_0$, tem-se um problema de valor inicial;

b) quando são conhecidos apenas relacionamentos funcionais, que envolvem a função f e suas derivadas, válidos para uma região Γ, limite do domínio Ω, tem-se um problema de condições de contorno;

c) quando num problema de condições de contorno impõe-se a restrição adicional $v = \lambda f$, onde λ é um número, tem-se um problema de autovalor.

Considerando duas funções f e g, define-se como deficiência linear a diferença:

$$\Delta(f,g) = D(f+g) - [Df + Dg] .$$

Quando a deficiência linear for identicamente nula, o operador D é chamado linear e representado pela letra L. Diz-se que o problema resultante é linear. Caso contrário, tem-se um problema não-linear.

Oferecem especial interesse, principalmente no campo da Física, as equações diferenciais parciais lineares de segunda ordem. Elas podem ser colocadas na forma geral:

$$\sum_{i=1}^{n} A_i \frac{\partial^2 f}{\partial x_i^2} + \sum_{i=1}^{n} B_i \frac{\partial f}{\partial x_i} + Cf + G = 0 ,$$

na qual os coeficientes A_i dependem do ponto $\underline{x}$ considerado, podendo ser 1, -1, ou 0. A ausência de derivadas mistas é conseguida por uma conveniente transformação de variáveis. Destacam-se três casos de importância:

a) se todos A_i têm o mesmo sinal e nenhum deles é nulo, a equação é chamada elíptica;

b) se nenhum A_i se anula e, pelo menos, um deles possui sinal diferente, a equação é hiperbólica;

c) se ao menos um dos A_i se anula e o seu correspondente B_i não, e se todos os outros A_i têm o mesmo sinal, a equação é dita parabólica.

Esta classificação depende certamente do particular ponto $\underline{x}$ considerado.

Quando várias equações diferenciais parciais devem ser resolvidas simultaneamente, tem-se um sistema destas equações. Considera-se neste trabalho apenas o caso de uma única equação. Entretanto, a extensão para sistemas de equações pode ser facilmente efetuada.

O tratamento aqui apresentado não é particularizado para casos específicos. O objetivo é prover o leitor com uma forma generalizada de abordagem do problema. A título ilustrativo, dois exemplos são inseridos no texto.

A forma para a obtenção da solução é converter a equação diferencial parcial numa equação algébrica, cuja solução pode ser obtida pelos métodos apresentados no Capítulo 21. Dois métodos destacam-se neste particular: o de diferenças finitas e o de elementos finitos. Apresentam-se neste capítulo alguns aspectos do segundo método pelas razões que motivaram o seu aparecimento (ver item 2 da Seção 22.3.1 do capítulo anterior). O primeiro método é discutido em detalhes na literatura (e.g. Carnahan et alii, 1969) e pode, por vezes, ser considerado como caso particular do segundo método (Gallagher et alii, vol. 1, 1978). Este aspecto é apresentado, por exemplo, por Chung (1978).

23.2 - MÉTODOS DE ELEMENTOS FINITOS

As razões que motivaram o aparecimento deste método foram apresentadas no capítulo anterior. Basicamente, a idéia consiste em utilizar os métodos da teoria de aproximações (Capítulo 18) para converter a equação diferencial numa equação algébrica.

Duas formas de aproximação são possíveis: local e global. Na aproximação local uma função aproximadora é escolhida para cada região Ω_i, $i = 1, \ldots, N$, do domínio $\Omega(\Omega = \bigcup_i \Omega_i)$. A aproximação global exige uma função aproximadora válida para todo o domínio.

Os diferentes métodos dependem do critério utilizado para a otimização da função aproximadora. Assim tem-se:

a) métodos variacionais (Rayleigh-Ritz) - utilizam analogia com o cálculo de variações para otimização;

b) métodos de resíduos ponderados (Galerkin, mínimos quadrados) - utilizam o critério de resíduos ponderados para otimização.

23.3 - MÉTODOS VARIACIONAIS

Os métodos variacionais baseiam-se na minimização da integral:

$$I = \int_{\Omega} F d\Omega \ ,$$

onde F na condição de mínimo é uma função que depende de $\underline{x}$, de f e suas derivadas. Aqui f é a função incógnita que minimiza a integral.

A sequência para a obtenção da equação diferencial, cuja solução é f, pode ser facilmente encontrada na literatura (e.g. Courant and Hilbert, 1966, vol. I cap. IV). O método consiste em considerar uma aproximação $\tilde{f}$ para a função f:

$$\tilde{f}(\underline{x}) = f(\underline{x}) + \alpha\, g(\underline{x}) \ .$$

Esta aproximação deve satisfazer as mesmas condições da função f (pertencer ao domínio Ω com determinadas condições iniciais ou de contorno).

A presença da função $g(\underline{x})$ implica uma variação da integral, que será dada por:

$$\delta I(\alpha) = \frac{\partial I}{\partial \alpha} \delta \alpha$$

e deve ser anulada para a obtenção de f.

Desenvolve-se a derivada em relação a α, dentro do sinal de integração, considerando F como função de f e suas derivadas. Isto acarreta o aparecimento de derivadas parciais da função $g(\underline{x})$, que são removidas por integração por partes para a obtenção da forma:

$$\int_{\Omega} g(\underline{x}) \, [\mathcal{E}(F)] \, \delta \alpha \, d\Omega = 0 \ ,$$

onde $\mathcal{E}$ é o operador diferencial de Euler-Lagrange. A solução desta equação implica a relação:

$$\mathcal{E}(F) = 0 \ ,$$

chamada equação de Euler-Lagrange que é a equação diferencial, cuja solução é a função f procurada.

Dois resultados importantes, do exposto acima, devem ser enfatizados:

a) considerado o produto escalar de funções, a transformação diferencial de Euler-Lagrange de função F é ortogonal ao erro da função $\tilde{f}$ expresso por:

$$\delta \, \tilde{f}(\underline{x}) = \delta \, [\alpha \, g(\underline{x})] \ ;$$

b) uma equação diferencial que envolve uma função f pode ser considerada como a equação de Euler-Lagrange de uma função incógnita F.

Este último resultado sugere a utilização das técnicas variacionais para a solução de equações diferenciais. Assim, dada a equação:

$$D\ f - v = 0\ ,$$

associa-se a ela a equação de Euler-Lagrange de uma função F a ser determinada. Então tem-se:

$$\mathcal{E}(F) = D\ f - v = 0\ .$$

A seguir, com integração por partes da transformação diferencial D f-v, pode-se obter uma forma do tipo $\mathcal{E}(F)$, o que possibilita a identificação da função F desejada.

23.3.1 - MÉTODO DE RAYLEIGH-RITZ

Para este método, supõe-se que foi encontrada uma função F que relaciona a equação diferencial:

$$D\ f - v = 0$$

com o problema variacional de minimização da integral:

$$I(f) = \int_{\Omega} F d\Omega\ .$$

A função desejada f é então aproximada por meio de um conjunto de funções locais conhecidas $\phi_i(\underline{x})$, i = 1, ..., N, através da relação:

$$\tilde{f}(\underline{x}) = \sum_{i=1}^{N} c_i\ \phi_i(\underline{x})\ .$$

As funções $\phi_i(\underline{x})$ devem satisfazer as seguintes condições:

a) $\phi_i(\underline{x})$ é não-nula apenas em uma região Ω_i do domínio Ω;

b) quando a região Ω_i envolver uma zona de restrição (condição inicial ou de contorno) do problema, a correspondente função $\phi_i(\underline{x})$ deve obedecer também tal restrição.

c) a coleção $\{\phi_i(\underline{x})\}$ deve constituir um conjunto de funções ortonormais em Ω.

Esta última condição exige que se tenha para $i \neq j$

$$\int_\Omega \phi_i \, \phi_j \, d\Omega = 0 \ ,$$

o que nem sempre é possível. Aceita-se então uma condição de "quase-ortogonalidade" expressa por:

$$\int_\Omega \phi_i \, \phi_j \, d\Omega << 1 \ .$$

Com esta aproximação da função f tem-se o problema de minimização da integral:

$$I(\underline{c}) = \int_\Omega F(\underline{x},\underline{c}) \, d\Omega \ ,$$

onde $\underline{c}$ é o vetor cujas componentes são os coeficientes c_i. A solução é imediata, devendo-se ter:

$$\frac{\partial I(\underline{c})}{\partial c_i} = 0 \ , \qquad i = 1, \ldots, N \ ,$$

o que fornece um sistema com N equações algébricas, cuja solução são os coeficientes c_i desejados.

No caso de aproximação global, as funções $\phi_i(\underline{x})$ são definidas para todo o domínio Ω, não se restringindo a regiões particulares. Nestas circunstâncias, torna-se mais difícil satisfazer as condições de particularização do problema, além da dificuldade do estabelecimento de funções ortogonais multidimensionais.

Uma das sérias restrições a este método é que nem sempre é possível converter a transformação diferencial:

$$D\ f - v$$

em uma forma variacional do tipo $\mathcal{E}(F)$.

23.4 - MÉTODOS DE RESÍDUOS PONDERADOS

Estes métodos utilizam o critério de resíduos ponderados para seu desenvolvimento. Assim, o primeiro passo consiste em caracterizar o resíduo a ser considerado.

Dada uma equação diferencial:

$$D\ f - v = 0$$

e tomando uma função $\phi(\underline{x})$ como aproximação da solução $f(\underline{x})$, resulta em:

$$D\phi - v = R(\underline{x},\phi)\ ,$$

que é considerado resíduo da aproximação.

Em seguida, estabelece-se uma função peso $w(\underline{x})$ como medida da importância relativa do resíduo. Aplica-se o critério dos resíduos ponderados, minimizando o erro médio dado por:

$$\varepsilon = \langle \underset{\underline{x}}{U} R(\underline{x},\phi)\ \hat{\underline{x}},\ \underset{\underline{x}}{U}\ w(\underline{x})\ \hat{\underline{x}} \rangle = \langle \underline{R},\underline{w} \rangle = \int_{\Omega} R(\underline{x},\phi)\ w(\underline{x})\ d\Omega\ .$$

Os diferentes métodos resultam dos diferentes valores atribuídos à função peso $w(\underline{x})$. Quatro métodos principais merecem referência:

a) Método de Galerkin - utiliza $w(\underline{x}) = \phi(\underline{x})$, e o erro é minimizado pela condição de ortogonalidade $\varepsilon = 0$ (ver Capítulo 18).

b) Método do Mínimo dos Quadrados - utiliza $w(\underline{x}) = R(\underline{x},\phi)$, e a melhor aproximação é obtida pelo mínimo erro médio.

c) Método dos Momentos - utiliza um conjunto de funções peso $w_i(\underline{x}) = q_i(\underline{x})$, onde $q_i(\underline{x})$ pertence a um conjunto de funções linearmente independentes $\{q_i(\underline{x})\}$, $i = 1, \ldots, N$, e impõe as condições:

$$\langle \underline{R}, \underline{w}_i \rangle = 0 , \qquad i = 1, \ldots, N .$$

d) Método de Colocação ("Collocation Method") - utiliza um conjunto de funções peso:

$$w_i(\underline{x}) = \delta(\underline{x} - \underline{x}_i) , \qquad i = 1, \ldots, N$$

e impõe as condições:

$$\langle \underline{R}, \underline{w}_i \rangle = 0 , \qquad i = 1, \ldots, N .$$

23.4.1 - MÉTODO DE GALERKIN

Uma vez estabelecido o fundamento do método, é necessário formalizar o procedimento seqüencial para seu desenvolvimento.

Dada a equação diferencial:

$$D\, f - v = 0$$

e as condições restritivas para particularização do problema, admite-se uma solução aproximada:

$$\phi(\underline{x}) = \sum_{i=1}^{N} c_i \, \phi_i(\underline{x}) \ ,$$

onde as $\phi_i(\underline{x})$ satisfazem essencialmente os mesmos requisitos já estabelecidos para o método de Rayleigh-Ritz. Desta forma, obtém-se um resíduo dado por:

$$R(\underline{x}, \underline{c}) = D\phi - v \ ,$$

onde $\underline{c}$ é o vetor cujas componentes são os coeficientes c_i.

Escolhe-se a função peso $w=\phi$ e impõe-se a condição de ortogonalidade entre esta função e o resíduo. Então tem-se:

$$\int_{\Omega} R\,\phi \, d\Omega = 0 \ .$$

A condição suficiente para que a relação de ortogonalidade seja satisfeita é:

$$\int_{\Omega} R\,\phi_i \, d\Omega = 0 \ , \qquad i = 1, \ldots, N \ ,$$

que fornece um sistema de N equações algébricas, cuja solução são os coeficientes c_i desejados.

Como no método de Rayleigh-Ritz, a aproximação pode ser local ou global. A vantagem do método de Galerkin é o de ser sempre aplicável.

23.4.2 - MÉTODO DO MÍNIMO DOS QUADRADOS

O procedimento seqüencial do método do mínimo dos quadrados é idêntico ao de Galerkin no que se refere ao estabelecimento do resíduo.

Como função peso escolhe-se o próprio resíduo e estabele-se a condição de otimização pelo mínimo erro médio. Então têm-se:

$$\min \varepsilon = \min \int_{\Omega} R^2 \, d\Omega \ .$$

A condição de mínimo erro médio é satisfeita quando todas as derivadas parciais deste erro em relação aos coeficientes c_i forem nulas. Isto fornece um sistema de N equações algébricas:

$$\int_{\Omega} 2R \frac{\partial R}{\partial c_i} \, d\Omega = 0 \ , \qquad i = 1, \ldots, N \ ,$$

cuja solução são os coeficientes c_i desejados.

O método do mínimo dos quadrados é sempre aplicável, e a aproximação utilizada pode ser local ou global.

Alguns autores (e.g. Chung, 1978) atribuem a este método um melhor desempenho do ponto de vista de convergência, mas uma desvantagem no que se refere à complexidade da função peso.

23.4.3 - MÉTODO DOS MOMENTOS

O desenvolvimento seqüencial deste método é idêntico ao de Galerkin. Pretende-se com ele apenas uma simplificação no processamento requerido para a obtenção dos coeficientes c_i.

O fundamento para a simplificação do processamento baseia-se em duas considerações principais:

a) em virtude das condições restritivas do problema (inicial ou de contorno), nem sempre as funções ϕ_i são as mais simples para a identificação do espaço vetorial das aproximações;

b) para que a aproximação seja otimizada em relação à solução, é suficiente a ortogonalidade do resíduo com relação a um conjunto de funções linearmente independentes $\{q_i(\underline{x})\}$, $i = 1, \ldots, N$, que possa ser considerado como uma base do espaço vetorial das aproximações.

O exemplo mais clássico é o caso unidimensional com aproximação polinomial. Neste caso, qualquer polinômio $\phi_i(x)$ será uma combinação linear do conjunto de monômios:

$$1, x, x^2, x^3, \ldots = \{q_i(x)\} .$$

Assim, o conjunto de relações de ortogonalidade:

$$\langle \underline{R}, \underline{\phi}_i \rangle = 0 , \qquad i = 1, \ldots, N$$

torna-se equivalente ao conjunto de relações:

$$\langle \underline{R}, \underline{q}_i \rangle = 0 , \qquad i = 1, \ldots, N$$

do ponto de vista de otimização no espaço vetorial das aproximações.

As características deste método são pois essencialmente as mesmas do método de Galerkin.

23.4.4 - MÉTODO DE COLOCAÇÃO

O desenvolvimento sequencial deste método é idêntico ao de Galerkin no que diz respeito ao estabelecimento da função aproximadora $\phi(\underline{x})$ e à caracterização do resíduo $R(\underline{x}, \underline{c})$.

As funções peso, $w_i(\underline{x}) = \delta(\underline{x} - \underline{x}_i)$, e a condição de erro médio nulo para cada peso fornecem as relações:

$$\langle \underline{R}, \underline{w}_i \rangle = R(\underline{x}_i, \underline{c}) = 0 \ , \qquad i = 1, \ldots, N \ ,$$

que constituem um sistema de N equações algébricas, cujas soluções são os coeficientes c_i.

Basicamente, este método consiste numa discretização do problema e deve ser usado apenas quando se procura a solução nas vizinhanças dos pontos $\underline{x}_i$.

23.5 - CONSIDERAÇÕES GERAIS SOBRE OS MÉTODOS DE ELEMENTOS FINITOS

Apenas no caso de aproximação local, tem-se o método de elementos finitos propriamente dito. O caso de aproximação global constitui uma extensão deste método.

Os métodos apresentados requerem, como hipótese fundamental, que o espaço vetorial, onde f é definida, contenha v e o resultado da transformação Df. Isto acontece nos espaços chamados Sobolev. Nestas circunstâncias, pelo menos quando o operador D admitir uma decomposição espectral discreta (Friedmann, 1966), as relações de ortogonalidade consideradas são equivalentes à condição de aproximação ótima (ver Capítulo 18):

$$\langle \underline{f} - \underline{\phi}, \underline{\phi}_i \rangle = 0 \ , \qquad i = 1, \ldots, N \ .$$

As considerações do último parágrafo mostram que os métodos de elementos finitos constituem fundamentalmente dispositivos para aplicação da teoria das aproximações (Capítulo 18). Como tal, os erros envolvidos podem ser avaliados com as técnicas desenvolvidas no Capítulo 18.

Um outro aspecto do problema dos erros de processamento diz respeito aos arredondamentos numéricos envolvidos. Como basicamente:

$$\frac{\partial\, I(\underline{c})}{\partial\, c_i} = 0 \,, \qquad\qquad i = 1, \ldots, N \,,$$

no método de Rayleigh-Ritz há uma certa insensibilidade do método com respeito a pequenas imprecisões nos valores dos coeficientes. Em virtude da eqüivalência dos métodos, estabelecida quando D admite uma decomposição espectral, tal insensibilidade é preservada nos outros métodos.

23.6 - EXEMPLOS ILUSTRATIVOS

Uma aplicação simples de utilização dos métodos aqui expostos é a solução da equação de Poisson (elíptica):

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = v \,,$$

com as restrições:

v = constante ,

$f = 0$ para $x = 0$ e $x = a$,

$f = 0$ para $y = 0$ e $y = b$.

O domínio Ω é o retângulo, no plano xy, limitado pelas retas:

$$x=0, \quad x=a, \quad y=0, \quad y=b \ .$$

Resolve-se o problema usando os métodos de Rayleigh-Ritz e Galerkin. Em vista de sua simplicidade usa-se a aproximação global.

23.6.1 - SOLUÇÃO PELO MÉTODO DE GALERKIN

Escreve-se a função aproximadora mais simples para o problema satisfazendo as restrições especificadas, como:

$$\phi = c\, x(x-a)\, y(y-b) \ .$$

O resíduo dependente da constante c é:

$$R(c) = \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 2c\,[x(x-a) + y(y-b)] - v \ ,$$

e a condição de otimização que permite determinar c torna-se:

$$\int_0^a \int_0^b \{2c[x(x-a) + y(y-b)] - v\}$$

$$[cx(x-a)\, y(y-b)]dx\, dy = 0 \ .$$

Integrando o lado esquerdo da equação acima resulta em:

$$-\frac{2c}{90}\,[a^3\, b^3(a^2+b^2)] - \frac{v\, a^3\, b^3}{36} = 0 \ ,$$

que fornece o valor:

$$c = -(5/4)\,\frac{v}{a^2 + b^2}\ .$$

23.6.2 - SOLUÇÃO PELO MÉTODO DE RAYLEIGH-RITZ

No presente caso a determinação da função F é relativamente simples pela identificação direta da equação de Euler-Lagrange (ver Exercício 2 no final deste capítulo) com a equação diferencial (de Poisson). Assim, obtém-se:

$$\frac{\partial F}{\partial f} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial f_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial f_y}\right) = \frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} - v = 0\ ,$$

que identificada termo a termo fornece:

$$\frac{\partial F}{\partial f} = -v \Longrightarrow F(...f...) = -vf\ ,$$

$$\frac{\partial F}{\partial f_x} = -f_x \Longrightarrow F(...f_x...) = -(1/2)f_x^2\ ,$$

$$\frac{\partial F}{\partial f_x} = -f_y \Longrightarrow F(...f_y...) = -(1/2)_y^2\ ,$$

de onde se tem finalmente:

$$F(x,y,f,f_x,f_y) = -(1/2)\,[f_x^2 + f_y^2 + vf(x,y)]\ .$$

Considerando agora a função aproximadora ϕ que satisfaz as condições restritivas do problema, tem-se:

$$\phi = cx(x-a)\,y(y-b)\ ,$$

$$\phi_x = c(2x-a)\,y(y-b)\ ,$$

$$\phi_y = cx(x-a)\ (2y-b)\ .$$

A constante c é determinada pela condição de mínimo da integral:

$$I(c) = \int_0^a \int_0^b F\ dx\ dy\ ,$$

o que implica fazer:

$$\frac{\partial I(c)}{\partial c} = 0\ ,$$

que fornece a equação:

$$\frac{c}{45}\left[a^3\ b^3\ (a^2+b^2)\right] + \frac{v\ a^3\ b^3}{36} = 0\ ,$$

de onde se tem o valor:

$$c = -(5/4)\ \frac{v}{a^2\ +\ b^2}\ .$$

Pode-se notar que a mesma solução foi obtida pelos dois métodos. O último processo apresenta entretanto maior complexidade, pois a determinação do resíduo (método de Galerkin) é consideravelmente mais simples que a obtenção da função F (método de Rayleigh-Ritz).

23.7 - MÉTODO DE ELEMENTOS FINITOS PARA EQUAÇÕES INTEGRAIS

As equações integrais possuem um estreito relacionamento com as equações diferenciais parciais. Elas ocorrem quando a aplicação de um operador integral I, sobre uma função f, produz outra função v. O relacionamento é representado por:

$$If = v \ .$$

Em geral, a aplicação do operador integral pode ser representada por:

$$If = \int_{\Omega} F \, d\Omega \ ,$$

onde F é uma função de f, das variáveis independentes e, eventualmente, das derivadas da função f. Assim, tem-se:

$$\int_{\Omega} F \, d\Omega - v = 0 \ .$$

Esta última relação mostra a estreita analogia entre a resolução de equações integrais e a aplicação do método de elementos finitos às equações diferenciais. O procedimento para a resolução deste tipo de equações é similar ao apresentado anteriormente.

Primeiramente, toma-se uma função aproximadora ϕ para a função f. A utilização de ϕ no lugar de f provoca o aparecimento de um erro integrado, dado por:

$$\varepsilon(\phi) = \int_{\Omega} F(\phi) \, d\Omega - v \ ,$$

e a condição de otimização é minimizar este erro.

Como no caso de equações diferenciais, a aproximação pode ser local ou global. Em geral, ϕ é expressa como combinação linear de funções ϕ_i. Isto permite que se escreva:

$$\phi = \sum_{i=1}^{N} c_i \, \phi_i \ ,$$

o que resulta em:

$$\varepsilon(\underline{c}) = \int_{\Omega} F(c)\, d\Omega - v \quad ,$$

onde $\underline{c}$ é o vetor cujas componentes são os coeficientes c_i.

Para minimizar $(\underline{c})$ devem-se impor as condições:

$$\frac{\partial \varepsilon(\underline{c})}{\partial c_i} = 0 \quad , \qquad i = 1, \ldots, N \quad ,$$

as quais fornecem um sistema com N equações algébricas, cujas incógnitas são os coeficientes c_i procurados. A solução deste sistema é a resposta ao problema.

As mesmas restrições apresentadas para as funções ϕ_i devem ser respeitadas neste caso. Também são válidas as considerações da Seção 23.5 para o caso de equações integrais.

23.8 - CONSIDERAÇÕES ADICIONAIS SOBRE OS MÉTODOS DE ELEMENTOS FINITOS

Quando a função aproximativa é um polinômio idêntico ao obtido pelo desenvolvimento de Taylor em torno de um ponto $\underline{x}_0$, há uma equivalência entre os métodos de elementos finitos e o de diferenças finitas. Este último pode, por esta razão, ser considerado um caso particular dos métodos aqui abordados.

A eqüivalência entre os métodos de Rayleigh-Ritz e Galerkin nem sempre é verificada, como mostra um exemplo em Zamlutti (1984).

O caso de problemas não-estacionários não oferece maiores dificuldades do que se considerar um dos x_i como a variável tempo. Entretanto, recomenda-se previamente a tentativa de utilizar a técnica de separação de variáveis para isolar a dependência temporal da dependência espacial na solução. Este procedimento têm-se mostrado extremamente útil em muitos casos (veja Mathews, 1970; Morse and Feshbach, 1953; Sokolnikoff and Redheffer, 1958).

EXERCÍCIOS

1. Mostre que o operador de Euler-Lagrange para $F(x, f, f')$ é dado por:

$$\mathcal{E} = \frac{\partial}{\partial f} - \frac{d}{dx}\left(\frac{\partial}{\partial f'}\right) .$$

2. Mostre que o operador de Euler-Lagrange para $F(x, y, f, f_x, f_y)$ é dado por:

$$\mathcal{E} = \frac{\partial}{\partial f} - \frac{\partial}{\partial x}\left[\frac{\partial}{\partial(\partial f/\partial x)}\right] - \frac{\partial}{\partial y}\left[\frac{\partial}{\partial(\partial f/\partial y)}\right] .$$

3. Considerando o resultado do Exercício 2, identifique a função F que produz a equação diferencial:

$$-\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2} = v(x,y) .$$

4. Encontre a função F para a equação diferencial:

$$\frac{d^2 f}{dx^2} - \alpha^2 f = v(x), \qquad x \in (0,1) .$$

5. Aplique o método de Galerkin para solução da equação:

$$\frac{d^2 f}{dx^2} + f = 0 ,$$

com as condições restritivas:

$f = 0$ para $x = 0$,

$f = 0$ para $x = 1$.

Utilize $N = 2$ e a aproximação global. Escolha as funções ϕ_i que satisfazem as condições restritivas do problema.

6. Repita o exercício anterior utilizando o método do mínimo dos quadrados.

7. Repita o Exercício 5 usando o método dos momentos e ortogonalizando o erro com relação às funções:

$$q_1(x) = 1, \quad q_2(x) = x .$$

8. Repita o Exercício 5 usando o método de colocação e impondo a condição de erro nulo para os pontos:

$$x = 1/3, \quad x = 2/3 .$$

9. Resolva a equação integral:

$$\int_0^1 f(x)\, dx = 20 ,$$

utilizando o método de elementos finitos com aproximação global e duas funções:

$$\phi_1(x) = 1, \quad \phi_2(x) = x .$$

10. Resolva a equação diferencial parabólica:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial t} \qquad x \in (0, 1), \quad t \in (0, T)$$

com as condições de contorno:

$$f(x, 0) = a(x) , \qquad x \in (0, 1)$$

$$f(0, t) = b_1(t) , \qquad t \in (0, T)$$

$f(1, t) = b_2(t)$, $\quad t \in (0, T)$

usando o método de Galerkin e em seguida o de separação de variáveis. Compare os resultados.

BIBLIOGRAFIA

CARNAHAN, B.; LUTHER, H.A.; WILKES, J.O. *Applied numerical methods.* New York, John Wiley, 1969.

CHUNG, T.J. *Finite element analysis in fluid dynamics.* New York, McGraw-Hill, 1978.

CONNOR, J.J.; BREBBIA, C.A. *Finite element techniques for fluid flow.* London, Newnes Butterworths, 1976.

COURANT, R.; HILBERT, D. *Methods of mathematical physics.* New York, Interscience Publishers, 1966, vol. 1 e 2.

DAVIS, H.T. *Introduction to nonlinear differential and integral equations.* New York, Dover, 1962.

DESAI, C.S. *Elementary finite element method.* Englewood Cliffs, Prentice-Hall, 1979.

FRIEDMAN, B. *Principles and techniques of applied mathematics.* New York, John Wiley, 1966.

GALLAGHER, R.H.; ZIENKIEWICZ, O.C.; ODEN, J.T.; MORANDI CECCHI, M.; TAYLOR, C. *Finite elements in fluids.* Chichester, John Wiley, 1978, v. 1/3.

MARCHOUK, K.G. *Methodes de calcul numérique.* Moscou, Mir, 1980.

MATHEWS, J.; WALKER, R.L. *Mathematical methods of physics.* New York, W.A. Benjamin, 1970.

MORSE, P.M.; FESHBACH, H. *Methods of theoretical physics.* New York, McGraw-Hill, 1953, v. 1.

SOKOLNIKOFF, I.S.; REDHEFFER, R.M. *Mathematics of physics and modern engineering.* New York, McGraw-Hill, 1958.

STAKGOLD, I. *Green's functions and boundary value problems.* New York, John Wiley, 1979.

YOUNG, D.M.; GREGORY, R.T. *A survey of numerical mathematics.* Reading Addison-Wesley, 1972. v. 2.

ZAMLUTTI, C.J. *A few necessary considerations for the use of the finite element methods by non-specialists.* 1984.

COMENTÁRIOS GERAIS SOBRE A BIBLIOGRAFIA

O presente texto procura explorar ao máximo as idéias geradoras de métodos numéricos e não estes propriamente ditos. A razão é evitar a obsolescência observada em textos mais específicos (ver comentários em Ralston e Rabinowitz, 1978).

O leitor deve distinguir três níveis de aprendizado do assunto:

a) familiarização com o material,

b) estudo propriamente dito,

c) pesquisa de assuntos específicos.

Quanto ao item (a) devem ser citadas as obras de caráter introdutório:

CONTE, S.D. *Numerical analysis: An algorithmic approach.* New York, NY, McGraw-Hill, 1965. (McGraw-Hill Series in Information Processing and Computers)

RENNINGTON, R.H. *Introductory computer methods and numerical analysis.* London, MacMillan, 1970.

O item (b) engloba a maioria das referências citadas no texto, dentre as quais se destacam, por serem as mais citadas na literatura, as seguintes:

DAHLQUIST, C.; BJORCK, A. *Numerical Methods.* Englewood Cliffs, Prentice Hall, 1974.

HAMMING, R.W. *Numerical methods of scientists and engineers.* New York, McGraw-Hill, 1962.

HILDEBRAND, F.B. *Introduction to numerical analysis.* New York, McGraw-Hill, 1956.

RALSTON, A.; RABINOWITZ, P. *First course in numerical analysis.* New York, McGraw-Hill, 1978.

YOUNG, D.M.; GREGORY, R.T. *A survey of numerical mathematics.* Reading Addison-Wesley, 1972. v. 2.

Quanto ao item (c), o desenvolvimento deve começar pelas coletânias de artigos apresentados em conferências, entre as quais sobressaem:

RALSTON, A.; WILF, H.S. *Mathematical methods for digital computers.* New York, John Wiley, 1968. v. 1/2,

JACOBS, D. *The state of the arts in numerical analysis.* London, Academic Press, 1977;

e pelos textos específicos sobre cada assunto, entre os quais se incluem os trabalhos básicos de:

CHENEY, E.W. *Introduction to approximation theory.* New York, McGraw-Hill, 1966.

CHUNG, T.J. *Finite element analysis in fluid dynamics.* New York, McGraw-Hill, 1978.

GEAR, C.W. *Numerical initial value problems in ordinary differential equations.* Englewood Cliffs, Prentice-Hall, 1971.

HENRICI, P. *Discrete variable methods in ordinary differential equations.* New York, John Wiley, 1962.

TRAUB, J.F. *Iterative methods for the solution of equations.* New Jersey, Prentice-Hall, 1964.

O material aqui apresentado inclui informações complementares provenientes das seguintes revistas:

1) Journal of the Association for Computing Machinery. New York.
2) Journal of the Society for Industrial and Applied Mathematics. Philadelphia.
3) SIAM Journal on Numerical Analysis. Philadelphia.
4) SIAM Review. Philadelphia.

No tocante à implentação dos métodos ao nível elementar (linguagens Basic e Fortran), o leitor encontrará valiosos subsídios em:

CARNAHAN, B.; LUTHER, H.A.; WILKES, J.O. *Applied numerical methods.* New York, John Wiley, 1969.

FORSYTHE, G.E.; MALCOLM, M.A.; MOLU, C.B. *Computer methods for mathematical computations.* Englewood Cliffs, NJ, Prentice-Hall, 1977.

PENNINGTON, R.H. *Introductory computer methods and numerical analysis.* London, MacMillan, 1970.

Para um tratamento profissional do assunto não pode ser dispensada a utilização de linguagens algorítmicas (Algol e Ada por exemplo), bem como a consulta a textos especializados no assunto, dentre os quais se destacam:

KNUTH, D.E. *Seminumerical algorithms.* Reading Addison-Wesley, 1981. v. 2.

Collected Algorithms from CACM Association for Computing Machinery Inc. New York.