CURS 3
METODE NUMERICE PENTRU SISTEME DE ECUAȚII LINIARE

I. Metode directe: Gauss; LU; Cholesky; Cholesky – matrici bandă.

II. Metode iterative: Jacobi; Gauss-Seidel; SOR.

III. Stabilitatea soluției şi condiționarea sistemului.

0. INTRODUCERE

Sistemele de ecuații liniare apar în modelarea unor probleme ştiințifice, precum şi în metodele numerice de rezolvare a acestora. Exemple: sisteme de ecuații neliniare, aproximarea funcțiilor, probleme la limită pentru ecuații diferențiale, ecuații cu derivate parțiale, optimizare, etc.

Găsirea soluției unui sistem de ordin mare (sau inversarea unei matrici de ordin mare), poate fi o sarcină dificilă în practică, datorită:

(1) Numărului mare de operații aritmetice necesare pentru găsirea soluției.

(2) Erorilor propagate, într-un şir lung de operații cu numere reprezentate în virgulă flotantă.

Dificultatea (1) face nepractică rezolvarea manuală, iar (2) este inerentă în utilizarea calculatorului.

Analiza unei metode cuprinde problemele:

- Numărul de operații aritmetice necesare pentru a găsi soluția.
- Estimarea, înainte de calculare, a preciziei soluției (estimarea *a priori* a erorii).
- Verificarea preciziei soluției calculate (estimarea *a posteriori* a erorii).

Problemele practice se pot clasifica după ordinul sistemului (numărul de ecuații) şi tipul matricii sistemului. Matricea sistemului se zice *densă*, dacă majoritatea elementelor sale sunt diferite de zero, şi *rară* – dacă majoritatea elementelor sunt zero. Astfel, avem categoriile:

I. Sisteme de ordin moderat, cu matrice densă:

Acestea se rezolvă în memorie, şi ordinul sistemului este limitat numai de memoria disponibilă. Uzual, ordinul sistemului poate ajunge la sute de ecuații. Majoritatea algoritmilor se bazează pe eliminarea Gauss, cu variante pentru clase speciale de matrici.

II. Sisteme de ordin mare, cu matrice rară:

Asemenea sisteme pot ajunge la sute de mii de ecuații. Eliminarea Gauss nu mai este convenabilă, şi se utilizează *metode iterative*.

III. Sisteme de tip II – rezolvare prin utilizarea de fişiere disc

Memoria fiind insuficientă, se folosesc fişiere pe disc pentru generarea şi procesarea matricii. Lucrul cu matricea sistemului se face pe 'blocuri': de exemplu, matricea este generată pe disc; succesiv, blocurile sunt citite, procesate în memorie şi rescrise pe disc. Un exemplu este cel al sistemelor cu matrice bandă simetrică, sisteme care apar în problemele de analiză statică a structurilor ■

Un sistem de ordinul n se va nota prin:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ \quad \dots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n \end{cases}$$

Matriceal, sistemul se scrie:

$$\begin{bmatrix} a_{11} & \dots & \dots & a_{1n} \\ a_{21} & \dots & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \qquad (1)$$

Sau:

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \qquad (1')$$

în care $\mathbf{A}$ este matricea sistemului ($n \times n$), iar $\mathbf{x}$ şi $\mathbf{b}$ vectorul necunoscutelor şi, respectiv, vectorul termenilor liberi (matrici coloană cu n elemente):

$$\mathbf{A} = [a_{ij}]_{i,j=1,n} \qquad \mathbf{x} = [x_1 \ ... \ x_n]^T\,, \qquad \mathbf{b} = [b_1 \ ... \ b_n]^T$$

I. METODE DIRECTE

1 Metoda Gauss

1.1 Metoda

Sistemul dat se transformă în sistemul echivalent $\mathbf{Ux} = \mathbf{g}$, cu matricea $\mathbf{U}$ superior triunghiulară, cum urmează. Sistemul dat se notează

$\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$.

La pasul curent k ($k = 1, 2, \ldots, n-1$), sistemul este

$\mathbf{A}^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$.

Se lucrează cu liniile $i = k, \ldots, n$, din $\mathbf{A}^{(k)}$ şi $\mathbf{b}^{(k)}$. (Liniile $1, \ldots, k-1$, procesate anterior, rămân neschimbate). Concret, se operează cu sub-matricea $\mathbf{A}^{(k)}(k:n, k:n)$ şi sub-coloana $\mathbf{b}^{(k)}(k:n)$ a termenilor liberi.

$$\mathbf{A}^{(k)} = \left[\begin{array}{ccccccccc}
a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1k}^{(1)} & . & & . & & a_{1n}^{(1)} \\
0 & a_{22}^{(2)} & \ldots & a_{2k}^{(2)} & . & & . & & a_{2n}^{(2)} \\
. & 0 & \ddots & & . & & . & & \\
. & . & \ldots & a_{k-1,k-1}^{(k-1)} & . & & . & & a_{k-1,n}^{(k-1)} \\
\hline
0 & . & \ldots & 0 & a_{kk}^{(k)} & . & a_{kj}^{(k)} & . & a_{kn}^{(k)} \\
. & . & \ldots & . & . & . & . & . & . \\
. & . & \ldots & . & a_{ik}^{(k)} & . & a_{ij}^{(k)} & . & a_{in}^{(k)} \\
. & . & \ldots & . & . & . & . & . & . \\
0 & 0 & \ldots & 0 & a_{nk}^{(k)} & . & . & . & a_{nn}^{(k)}
\end{array}\right] \begin{array}{l} \times(-m_{ik}); \\ \downarrow \\ \oplus \hookleftarrow \end{array}$$

$$\mathbf{b}^{(k)} = \left[\begin{array}{c}
b_1^{(1)} \\
b_2^{(2)} \\
\\
\hline
b_{k-1,k-1}^{(k-1)} \\
\hline
b_k^{(k)} \\
. \\
b_i^{(k)} \\
. \\
b_n^{(k)}
\end{array}\right] \begin{array}{l} \times(-m_{ik}) \\ \downarrow \\ \oplus \hookleftarrow \end{array}$$

Elementul diagonal $a_{kk}^{(k)}$ se numeşte *pivot* (la pasul *k*). Presupunem $a_{kk}^{(k)} \neq 0$ – v. mai jos pivotare, şi definim multiplicatorii de la pasul *k*:

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}; \qquad i = k+1, \ldots, n$$

Pentru $i = k+1, \ldots, n$: linia *k* se înmulţeşte cu $-m_{ik}$, şi se adună la linia *i*. Rezultă elemente nule în coloana *k*, sub elementul diagonal $a_{kk}^{(k)}$. Linia *k* rămâne neschimbată. Noii coeficienţi şi termeni liberi vor fi:

$$\left.\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik} \cdot a_{kj}^{(k)}; \quad j = k+1, \ldots, n \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik} \cdot b_k^{(k)} \end{aligned}\right\} \quad i = k+1, \ldots, n$$

La pasul *n* se notează, pentru convenienţă, $\mathbf{U} = \mathbf{A}^{(n)}$, $\mathbf{g} = \mathbf{b}^{(n)}$, şi sistemul devine $\mathbf{Ux} = \mathbf{g}$. Explicit:

$$\begin{bmatrix} u_{11} & . & . & . & . & u_{1n} \\ 0 & u_{22} & . & . & . & u_{2n} \\ & & . & . & . & . \\ 0 & & 0 & u_{kk} & . & u_{kn} \\ & & & & . & . \\ 0 & & & & 0 & u_{nn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ . \\ x_k \\ . \\ x_n \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ . \\ g_k \\ . \\ g_n \end{bmatrix} \qquad (2)$$

Observaţi că $u_{kk} = a_{kk}^{(k)}$.

Acest sistem se rezolvă prin substituţie înapoi:

$$x_n = g_n / u_{nn}; \qquad x_k = (g_k - \sum_{j=k+1}^{n} u_{kj} \cdot x_j) / u_{kk}; \quad k = n-1, n-2, \ldots, 1$$

■

1.2 Factorizarea (descompunerea) LU

Multiplicatorii m_{ik} (unde $i = k+1, \ldots, n$) din eliminarea Gauss, se pot reţine pentru a rezolva $\mathbf{Ax} = \mathbf{b}$ cu diferiţi termeni liberi **b**. Introducem atunci, matricea inferior-triunghiulară (cu 1 pe diagonală, şi multiplicatorii m_{ik} în sub-coloanele $k+1 : n$):

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ m_{21} & 1 & 0 & & & 0 \\ m_{31} & m_{32} & 1 & 0 & & 0 \\ . & . & . & . & . & . \\ . & . & . & . & 1 & 0 \\ m_{n1} & m_{n2} & . & . & m_{n,n-1} & 1 \end{bmatrix} \qquad (3)$$

Avem următoarea propoziție:

Propoziția 1

Dacă, în eliminarea Gauss, *la fiecare pas k* pivotul $a_{kk}^{(k)} \neq 0$, atunci $\mathbf{A} = \mathbf{LU}$ ■

L este matricea inferior triunghiulară a multiplicatorilor (3), iar **U** matricea superior triunghiulară din (2) ■

***Consecință – Calculul determinantului*:**

Dacă la fiecare pas *k* pivotul $a_{kk}^{(k)} \neq 0$, atunci:

$$\det(\mathbf{A}) = u_{11} \cdot u_{22} \cdot \ldots \cdot u_{nn}$$

Observați că $u_{kk} = a_{kk}^{(k)}$, astfelcă valoarea determinantului este *produsul pivoților* ■

Într-adevăr: $\det(\mathbf{A}) = \det(\mathbf{L}) \cdot \det(\mathbf{U})$, iar $\det(\mathbf{L}) = 1$.

Pentru calculul determinatului în cazul când se schimbă liniii în matricea **A** (se pivotează), v. mai jos.

1.3 Număr de operații

Pentru *n* = ”mare”, numărul de operații în eliminarea Gauss este:

$$NOP_{Guass} \approx \frac{n^3}{3}$$

(Operație = adunare sau scădere; în calcule, acestea sunt urmate de obicei de o inmulțire sau împărțire.)

■

1.4 Pivotare

La fiecare pas *k*, *pivotul* $a_{kk}^{(k)} \neq 0$ apare la numitorul multiplicatorilor m_{ik}. Astfel, avem condițiile:

Pivotul nu trebuie să fie nul, şi nici "foarte mic":

- Dacă pivotul este nul, eliminarea Gauss nu poate continua. În acest caz, matricea este singulară.
- Dacă pivotul este "foarte mic": rezultă multiplicatori m_{ik} mari, iar erorile de rotunjire în valoarea lui m_{ik} conduc la erori mari în calculațiile următoare. Matricea este "aproape singulară".

Practic, se testează dacă pivotul este mai mic decăt un *prag* ales.
V. Exemplul de mai jos.

Procedeul de alegere a unui pivot ≠ 0 se zice *pivotare*. Aceasta constă în:

- Căutarea elementului de modul maxim, în sub-matricea cu liniile şi coloanele $k, \ldots, n$;
- Aducerea lui pe poziția de pivot – poziția (k,k) – prin schimbări de linii, sau schimbări de linii şi coloane.

  Schimbarea de linii din **A**, se face şi în vectorul **b**.

Dacă pivotul este mai mic decât pragul, eliminarea Gauss se opreşte.

Strategii de pivotare:

1) Pivotarea parțială:

Se caută elementul de modul maxim în sub-coloana $a(k:n,k)$. Acesta devine pivot, aducându-l în poziția (k,k). Fie pivotul găsit în linia $I \geq k$. Dacă pivotul este mai mare decât pragul şi $I > k$, atunci se schimbă liniile *k* şi *I* (în $\mathbf{A}^{(k)}$ şi în $\mathbf{b}^{(k)}$).

$$\left[\begin{array}{c|ccc} a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{Ik}^{(k)} & \dots & . \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{array}\right] \begin{matrix}(k)\\ \updownarrow \\ (I)\end{matrix} \qquad \left[\begin{array}{c} b_k^{(k)} \\ \vdots \\ b_I^{(k)} \\ \vdots \\ b_n^{(k)} \end{array}\right] \begin{matrix}(k)\\ \updownarrow \\ (I)\end{matrix}$$

2) Pivotarea completă:

Se caută elementul de modul maxim în sub-matricea $a(k:n, k:n)$. Dacă maximul este atins pentru linia $I > k$ şi coloana $J > k$, atunci se permută liniile k şi I şi coloanele k şi J - în **A**, şi liniile k şi I - în **b** ■

$$\left[\begin{array}{ccccc} a_{kk}^{(k)} & \dots & . & \dots & a_{kn}^{(k)} \\ \vdots & & & & \vdots \\ . & & a_{IJ}^{(k)} & & . \\ \vdots & & & & \vdots \\ a_{nk}^{(k)} & \dots & . & \dots & a_{nn}^{(k)} \end{array}\right] \begin{matrix}(k)\\ \updownarrow \\ (I)\end{matrix} \qquad \left[\begin{array}{c} b_k^{(k)} \\ \vdots \\ b_I^{(k)} \\ \vdots \\ b_n^{(k)} \end{array}\right] \begin{matrix}(k)\\ \updownarrow \\ (I)\end{matrix}$$

$$(k) \leftrightarrow (J)$$

Observaţii

- Prin pivotare rezultă, la fiecare pas, multiplicatori de modul ≤ 1:

  $|m_{ik}| \le 1, \quad i = k+1, \dots, n$.

  Acesta previne generarea în $\mathbf{A}^{(k+1)}$**,** de elemente de mărime foarte diferită, care ar putea duce la erori de pierdere de semnificaţie.

- La pivotarea completă, datorită permutării de coloane, ordinea necunoscutelor se schimbă. Ea trebuie refăcută la sfârşitul eliminării.

Teoretic, pivotarea completă duce la o precizie mai mare a soluției; în practică însă, diferența este mică față de pivotarea parțială. În plus, pivotarea completă cere mai mult timp-calculator. Din aceste motive, majoritatea algoritmilor utilizează pivotarea

parțială. În ceea ce urmează vom considera numai pivotarea parțială, numită pe scurt, pivotare ■

Observație

Dacă nu se pivotează, atunci avem **LU** = **A** (Propoziția 1).

Dacă se pivotează, atunci avem $\mathbf{LU} = \mathbf{A}'$, unde $\mathbf{A}' = \mathbf{PA}$, iar **P** este o matrice de permutare de linii. Matricea $\mathbf{A}'$ se obține din **A**, permutând liniile în aceeaşi ordine în care se permută la pivotare. Sistemul care se rezolvă este $\mathbf{A}'\mathbf{x} = \mathbf{b}'$, unde $\mathbf{b}'$ este **b** cu liniile permutate în ordinea de la pivotare ■

Calculul determinantului:

Determinantul matricii $\mathbf{A}' = \mathbf{LU}$ este $\det(\mathbf{A}') = u_{11}u_{22}\ \ldots u_{nn}$.

Determinantul matricii **A** va fi

$$\det(\mathbf{A}) = (-1)^{n\_l} \det(\mathbf{A}'),$$

unde *n_l* este numărul de schimbări de linii în cursul pivotării.

■

Exemplu – Pivot ”foarte mic”

Considerăm sistemul $\mathbf{Ax} = \mathbf{b}$, unde:

$$\mathbf{A} = \begin{bmatrix} 5 & 6 & 3 & 1 \\ -1 & 0 & -1 & 1 \\ 2 & 2 & 1 & 6 \\ 4 & 2 & 3 & 4 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$$

În calculul exact, **A** este singulară. Rezolvând în simplă precizie, cu programul Gauss 2007, cu un prag de $1E-7$, se obțin următoarele rezultate:

```
Permutare, Pivot:
1       5.000000
2<->4   -2.800000
3     -0.2857143
4      7.7486033E-07
```

```
 * Solutia pentru termenii liberi nr. 1
 x(1) =  -0.2193944E+08
 x(2) =    6452776.
 x(3) =   0.2322999E+08
 x(4) =    1290555.


* Proba
Diferente (A*x-b)
           1  -4.625000
           2  0.3750000
           3  -1.750000
           4  -4.500000
```

Adică, soluția nu verifică, şi este inutil a o calcula.

Erorile în soluție sunt produse de pivotul de la pasul 4, care este “foarte mic”.

Cu un prag de $1E-6$, programul se opreşte cu mesajul:

```
Linia 4: Pivot =  7.7486033E-07
* Pivot < 1.00E-06 !
* Sistemul nu se poate rezolva.
```

În calculul în dublă precizie, pivotul din linia 4 rezultă `0.000000000000000`.

■

3 Descompunerea LU

3.1 Calculul direct al factorilor L şi U

Problema este calculul direct al factorilor **L** şi **U**, astfel ca $\mathbf{A} = \mathbf{LU}$.

Elementele lui **L** şi **U** se găsesc din ecuațiile $a_{ij} = linia\ i_{\mathbf{L}} \times coloana\ j_{\mathbf{U}}$:

$$a_{ij} = \begin{bmatrix} l_{i1} & l_{i2} & \dots & l_{ii} & \vdots & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jj} \\ \cdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sum_{k=1}^{\min(i,j)} l_{ik} u_{kj}$$

În particular, înmulțind linia *i* din **L** cu coloana *i* din **U**, rezultă:

$$a_{ii} = \sum_{k=1}^{i-1} l_{ik} u_{ki} + l_{ii} u_{ii}$$

Există o neunicitate a descompunerii **LU**, care provine din alegerea arbitrară a lui l_{ii} sau u_{ii}, $i = 1, \dots, n$. Două metode se pot considera:

l_{ii} = arbitrar $\neq 0$;

u_{ii} = arbitrar $\neq 0$.

Prezentăm, pentru exemplificare, formulele generale pentru prima metodă.

l_{ii} = ales arbitrar (nenul), $i = 1, 2, .. , n$

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik} \cdot u_{kj}}{l_{ii}}, \qquad j = i,\ i+1, \dots, n \tag{a}$$

Dacă $u_{ii} \neq 0$, $i = \overline{1, n-1}$, rezultă:

$$l_{ji} = \frac{a_{ji} - \sum_{k=1}^{i-1} l_{jk} \cdot u_{ki}}{u_{ii}}, \qquad j = i+1, \dots, n \tag{b}$$

Elementele lui **L** si **U** se determină în următoarea secvență, conform schemei de mai jos (se calculează elementele din linia *i* din **U,** şi coloana *i* din **L**).

l_{ii}	$\rightarrow$ $u_{i,i}$	$u_{i,i+1}$	…	$u_{i,n}$
$\downarrow$				
$l_{i+1,i}$				
$\vdots$				
$l_{n,i}$				

În particular, la $i = n$, se calculează numai u_{nn} - din (a).

Numărul de operații este acelaşi ca în eliminarea Gauss ($\approx n^3/3$).

Două metode se utilizează în practică:

$l_{ii} = 1$ ... Metoda *Doolittle* (descompunerea **LU** revine la cea din eliminarea Gauss).

$u_{ii} = 1$ ... Metoda *Crout*.

Posibilitatea descompunerii LU

Propoziție

Descompunerea **LU** există dacă şi numai dacă toate sub-matricile *principale* $\mathbf{A}(1:i,1:i),\ i = \overline{1,n-1}$ sunt nesingulare ■

3.2 Pivotare la descompunerea LU

Pivotare parțială: A pivota în descompunerea **LU** inseamnă ca, la fiecare pas *i*, să căutm *pivotul* u_{ii} în coloana *i* a lui **A**, în liniile $i, i+1, \ldots, n$, şi apoi să permutăm două linii.

Pivotarea trebuie făcută înainte de calculul elementelor lui **U** şi **L**. Adică, pivotul u_{ii} trebuie calculat şi testat *înainte* de aplicarea formulelor (a) şi (b).

Pentru metoda Doolittle, pivotul este dat de (cf. a):

$$u_{ii} = a_{ii} - \sum_{k=1}^{i-1} l_{ik} \cdot u_{ki}$$

Aşa cum s-a arătat la metoda Gauss, trebuie pivotat şi dacă pivotul este foarte mic. Practic, se testează dacă pivotul este mai mic decât un *prag*.

Observaţia de la pivotarea în eliminarea Gauss se aplică).

Determinantul matricii **A** (în metoda Doolittle) este

$$\det(\mathbf{A}) = (-1)^{n\_l}\, u_{11}u_{22} \ldots u_{nn},$$

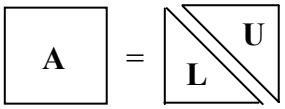
unde *n_l* este numărul de schimbări de linii în cursul pivotării.

■

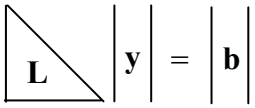
3.3 Metoda

Rezolvarea sistemului prin descompunerea **LU**, constă în paşii:

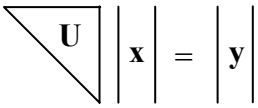
1. Factorizare $\mathbf{A} = \mathbf{LU}$, cu pivotare parţială.



2. Rezolvarea sistemului $\mathbf{Ly} = \mathbf{b}$, prin *substituţie înainte*; rezultă **y**.



3. Rezolvarea sistemului $\mathbf{Ux} = \mathbf{y}$, prin *substituţie înapoi*; rezultă **x**.



Observaţii

- Legătura cu eliminarea Gauss este **y** = **g**.
- Pentru noi termeni liberi **b**, se repetă numai paşii 2, 3 (n^2 operaţii suplimentare pentru fiecare **b**)
- Numărul de operaţii este acelaşi ca în eliminarea Gauss ($\approx n^3/3$).

■

4 Metoda Cholesky

Metoda Cholesky se aplică pentru un sistem cu matrice *simetrică şi pozitiv definită*.

Simetrie: $\mathbf{A}^T = \mathbf{A} \quad \Leftrightarrow \quad a_{ij} = a_{ji}, \quad i, j = \overline{1,n}$

Definire pozitivă: $\forall \mathbf{x} \neq \mathbf{0} \quad \mathbf{x}^T \mathbf{A}\mathbf{x} = \sum_i \sum_j a_{ij} x_i x_j > 0$

4.1 Proprietăți ale matricilor simetrice şi pozitiv definite

Pentru o matrice *pozitiv definită*, au loc proprietățile:

(1) Matricea este nesingulară.

(2) Elementele diagonale sunt pozitive: $a_{kk} > 0$. Mai mult, sub-matricile principale sunt pozitiv definite.

Urmează că a_{11} poate fi luat ca pivot în eliminarea Gauss, la pasul 1.

Pentru o matrice *simetrică şi pozitiv definită*, au loc şi proprietățile:

(3) Sub-matricile $\mathbf{A}^{(k)}(k:n, k:n), \quad k \geq 2$, rezultate din eliminarea Gauss sunt simetrice şi pozitiv definite.

De aici, rezultă că există pivoți $a_{kk}^{(k)} > 0$, la fiecare pas *k*.

(4) Valorile proprii ale lui **A** sunt reale şi pozitive.

Mai întâi, se arată că dacă **A** este simetrică, valorile proprii sunt reale – v. Cap. 5.
Apoi, fie λ o valoare proprie şi **x** vectorul propriu asociat, atunci $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.
Inmulțind la stânga cu $\mathbf{x}^T$ şi ținând cont de definirea pozitivă a matricii **A**, rezultă $\lambda \mathbf{x}^T \mathbf{x} > 0$, din care urmează că λ este pozitiv.

Observație

Reciproc, dacă **A** este simetrică şi valorile proprii sunt pozitive, atunci **A** este pozitiv definită. Demonstrația cere rezultate de algebră matriceală care sunt în afara scopului acesui paragraf. V. de exemplu, R. Bellman, "Introducere in analiza matriceală", E.T., Bucureşti, 1969.

(5) Determinatul matricii este pozitiv. Mai mult, toți determinații principali sunt pozitivi.

Întrucât determinatul matricii este $\det(\mathbf{A}) = \lambda_1\lambda_2 \ldots \lambda_n$ - v. Cap. 5, urmează că avem $\det(\mathbf{A}) > 0$. A doua afirmație decurge din faptul ca sub-matricile principale sunt şi ele pozitiv definite – conform Proprietății 2.

Observație

Reciproc, dacă matricea este simetrică şi toți minorii principali sunt pozitivi, matricea este pozitiv definită. Pentru demonstrație – v. Wilkinson-2.

Cu aceasta, rezultă:

Dacă **A** *este simetrică, o condiție necesară şi suficientă ca* **A** *să fie pozitiv definită, este ca toți determinații principali să fie pozitivi.*

■

4.2 Metoda Cholesky

Propoziție

Dacă **A** este simetrică şi pozitiv definită, atunci există descompunerea

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T, \qquad (1)$$

unde **L** este o matrice *inferior* triunghiulară, cu elemente diagonale *pozitive*.

■

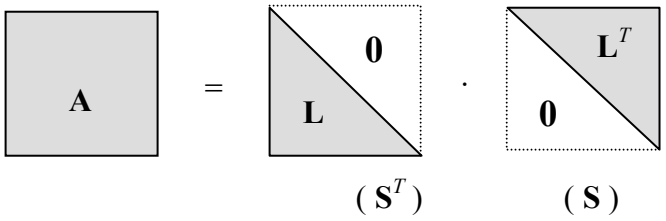
Cu alte cuvinte, în descompunerea **LU** se poate alege $\mathbf{U} = \mathbf{L}^T$. Notând

$\mathbf{S} = \mathbf{L}^T$,

unde **S** este *superior* triunghiulară, avem şi descompunerea

$$\mathbf{A} = \mathbf{S}^T\mathbf{S} \qquad (2)$$

Cele două forme (1) şi (2) reprezintă aceeaşi descompunere, luând ca referință una sau alta din cele două matrici triunghiulare. Ele sunt ilustrate mai jos. **0** desemnează elementele nule.



■

Exemplu – Evaluarea lui **L**, pentru $n = 3$:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \cdot \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

Ecuaţii: Lucrăm cu triunghiul *inferior* al lui **A**, luând elementele *pe coloane*:

$$a_{11} = l_{11}^{\,2}, \quad a_{21} = l_{21} \cdot l_{11}, \quad a_{31} = l_{31} \cdot l_{11}$$

$$a_{22} = l_{21}^{\,2} + l_{22}^{\,2}, \quad a_{32} = l_{31} \cdot l_{21} + l_{32} \cdot l_{22}$$

$$a_{33} = l_{31}^{\,2} + l_{32}^{\,2} + l_{33}^{\,2}$$

Rezultă elementele lui **L**, *pe coloane*:

$$l_{11} = \sqrt{a_{11}}; \qquad l_{21} = \frac{a_{21}}{l_{11}}, \qquad l_{31} = \frac{a_{31}}{l_{11}}.$$

$$l_{22} = (a_{22} - l_{21}^{\,2})^{1/2}; \qquad l_{32} = \frac{a_{32} - l_{31} \cdot l_{21}}{l_{22}}$$

$$l_{33} = (a_{33} - l_{31}^{\,2} - l_{32}^{\,2})^{1/2}$$

Analog, se poate lucra cu triunghiul superior al lui **A** şi determina elementele lui **L**, pe linii.

Exempu numeric:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 13 & 18 \\ 3 & 18 & 50 \end{bmatrix}; \qquad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 3 & 4 & 5 \end{bmatrix} \quad ■$$

Paşii rezolvării sunt aceeaşi ca la descompunerea **LU**, anume:

- Cu matricea **L** (unde $\mathbf{U} = \mathbf{L}^T$):

1) *Factorizare* (descompunere): $\mathbf{A} = \mathbf{L}\mathbf{L}^T$

2) S*ubstituţie înainte* (rezultă **y**): $\mathbf{L}\mathbf{y} = \mathbf{b}$

3) *Substituţie înapoi* (rezultă **x**): $\mathbf{L}^T\mathbf{x} = \mathbf{y}$.

- Cu matricea **S** (unde $\mathbf{U} = \mathbf{S}$; $\mathbf{L} = \mathbf{S}^T$):

1) $\mathbf{A} = \mathbf{S}^T\mathbf{S}$

2) $\mathbf{S}^T\mathbf{y} = \mathbf{b}$

3) $\mathbf{S}\mathbf{x} = \mathbf{y}$

■

4.3 Numărul de operaţii

Numărul de operaţii (adunări/scăderi), pentru *n* = ‘mare’, este

$$NOP_{Cholesky} \approx \frac{n^3}{6}$$

Acesta este cca. **½ din numărul de operaţii cerute eliminarea Gauss (sau LU)**.

(În afară de adunări/scăderi, se mai calculează *n* rădăcini pătrate.)

Alte avantaje, în raport cu descompunerea generală **LU**:

- Matricea **A** fiind simetrică, se poate stoca numai triunghiul inferior (sau superior), cerând numai $\frac{1}{2}n(n+1)$ locaţii de memorie. În aceste locaţii se stochează matricea **L** (sau **S**).
- Nu este necesară pivotarea – conform proprietăţilor (2, 3).

Observa□ie

Rădăcinile pătrate – care consumă timp-calculator – pot fi evitate printr-o uşoară modificare a descompunerii **LU**, şi anume: căutăm o matrice inferior triunghiulară $\widetilde{\mathbf{L}}$ cu 1 pe diagonală, şi o *matrice diagonală* **D**, astfel ca:

$$\mathbf{A} = \widetilde{\mathbf{L}}\mathbf{D}\widetilde{\mathbf{L}}^T$$

Într-adevăr: dacă avem $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, fie $\widetilde{\mathbf{L}}$ matricea obținută din **L** punînd pe diagonală $\widetilde{l}_{ii} = 1$ (în rest, $\widetilde{l}_{ij} = l_{ij}$), şi $\mathbf{D}' = diag(l_{ii})$. Avem $\mathbf{L} = \widetilde{\mathbf{L}}\mathbf{D}'$, şi $\mathbf{L}\mathbf{L}^T = \widetilde{\mathbf{L}}\mathbf{D}'\mathbf{D}'\widetilde{\mathbf{L}}^T$. Astfel, matricea $\mathbf{D} = \mathbf{D}'\mathbf{D}' = diag(l_{ii}^2)$.

Această factorizare se poate face cu aproximativ acelaşi număr de operații ca în metoda Cholesky, şi fără calcul de rădăcini pătrate ■

5 Matrici bandă – simetrice şi pozitiv definite

5.1 Matrice bandă simetrică

Presupunem că matricea sistemului este simetrică şi, în plus, are structura de *matrice bandă*, adică în fiecare linie elementele matricii sunt constitutite din:

- Elementul diagonal, un număr de *LIM*-1 elemente la stânga acestuia, şi *LIM*-1 elemente la dreapta.
- Celelalte elemente din linie sunt zero.

Numărul *LIM* va fi numit *semi-lățimea* de bandă. *LIM* reprezintă numărul elementelor din semi-bandă, inclusiv elementul diagonal.

Observație

Între elementele din bandă, pot fi şi elemente nule, dar caracterul de bandă este dat de faptul că toate elementele situate în afara benzii sunt nule. În acest sens *LIM* poate fi considerat semi-lățimea de bandă *maximă* ■

Exemplu: Matrice 6×6, $a_{ij} = a_{ji}$, şi *LIM* = 3:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & 0 \\ 0 & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ 0 & 0 & a_{53} & a_{54} & a_{55} & a_{56} \\ 0 & 0 & 0 & a_{64} & a_{65} & a_{66} \end{bmatrix}$$

5.2 Stocarea matricii bandă simetrică

Se stochează elementele din semi-banda superioară (sau inferioară). De exemplu, pentru matricea din Exemplul anterior şi semi-banda superioară, sunt de stocat elementele indicate mai jos:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & & & \\ & a_{22} & a_{23} & a_{24} & & \\ & & a_{33} & a_{34} & a_{35} & \\ & & & a_{44} & a_{45} & a_{46} \\ & & & & a_{55} & a_{56} \\ & & & & & a_{66} \end{bmatrix}$$

Stocarea se face într-un vector, în unul din modurile:

1) Pe linii sau coloane de *LIM* elemente. Exemplu, pe linii:

$\mathsf{a} = [a_{11}\ a_{12}\ a_{13} \mid a_{22}\ a_{23}\ a_{24} \mid a_{33}\ a_{34}\ a_{35} \mid a_{44}\ a_{45}\ a_{46} \mid a_{55}\ a_{56}\ 0 \mid a_{66}\ 0\ 0]$

2) Pe linii sau coloane. Exemplu, pe coloane:

$\mathsf{a} = [a_{11} \mid a_{12}\ a_{22} \mid a_{13}\ a_{23}\ a_{33} \mid a_{24}\ a_{34}\ a_{44} \mid a_{35}\ a_{45}\ a_{55} \mid a_{46}\ a_{56}\ a_{66}]$

3) Pe diagonale (diagonala principală şi paralele la aceasta):

$\mathsf{a} = [a_{11} \quad \ldots \quad a_{66} \mid a_{12} \quad \ldots \quad a_{56} \mid a_{13} \quad \ldots \quad a_{46}]$

Adresa unui element a_{ij}, în vectorul a, este dată de o funcție întreagă $Loca(i, j)$.

■

Stocarea se mai poate face pe linii de *LIM* elemente, într-un tablou `B(n, LIM)`. Liniile care conţin mai puţin de *LIM* elemente se completează cu elemene nule (exemplu: liniile 5 şi 6).

5.3 Metoda Cholesky

Proprietatea esenţială este următoarea:

> Dacă matricea bandă este *simetrică şi pozitiv definită*, atunci descompunerea $\mathbf{L}\mathbf{L}^T$ sau $\mathbf{S}^T\mathbf{S}$ poate fi făcută lucrând *exclusiv* în bandă ■

În ceea ce urmează vom considera descompunerea $\mathbf{A} = \mathbf{S}^T\mathbf{S}$, lucrând cu semi-banda superioară.

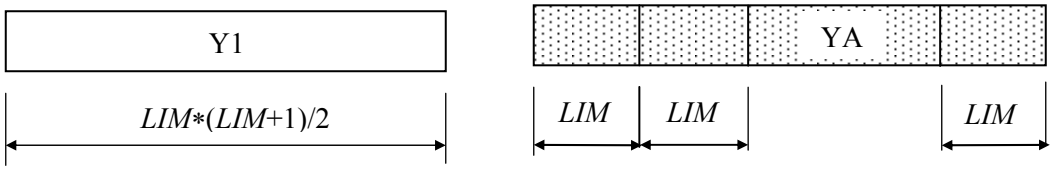
Paşii sunt cei de la Cholesky, lucrul cu matricea S.

Elementele active la un pas al descompunerii, sunt conţinute într-un triunghi cu laturile *LIM* – numit "triunghiul activ". În cursul procesului, triunghiul activ coboară cu câte o linie în bandă.

Metoda implementată în ANA\Choloesky_Band utilizează un vector de lucru Y, de dimensiune NY, unde:

$NY0 = LIM * (LIM + 1)/2$; $NY = NY0 + n * LIM$.

Acest vector este constituit din două părţi:



Tablourile Y1 şi YA

- Sub-vectorul $Y1(1:NY0)$, de dimensiune $LIM * (LIM + 1)/2$: serveşte ca front de lucru, pentru descompunerea Cholesky; în acestea se generează elementele triunghiului activ la un pas (triunghi cu laturile *LIM*).

- Sub-vectorul YA($NY0+1:NY$), de dimensiune $n*LIM$: inițial, stochează matricea **A**, pe linii. În cursul descompunerii Cholesky, stochează liniile procesate ale matricii A.

II. METODE ITERATIVE

1 Metoda JACOBI

Exemplu numeric

Fie sistemul:

$$8x_1 + x_2 - x_3 = 8$$

$$2x_1 + x_2 + 9x_3 = 12$$

$$x_1 - 7x_2 + 2x_3 = -4$$

Rezolvăm fiecare ecuație i = 1, 2, 3, în raport cu necunoscuta x_i, căutând ca aceasta să fie necunoscuta cu coeficientul cel mai mare din ecuație, eventual rearanjând ecuațiile. Intervertind ecuațiile 2 şi 3, avem:

$$x_1 = 1 - (1/8)x_2 + (1/8)x_3$$

$$x_2 = 4/7 + (1/7)x_1 + (2/7)x_3$$

$$x_3 = 12/9 - (2/9)x_1 - (1/9)x_2$$

Sau matriceal:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4/7 \\ 12/9 \end{bmatrix} + \begin{bmatrix} 0 & -1/8 & 1/8 \\ 1/7 & 0 & 2/7 \\ -2/9 & -1/9 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad (1')$$

care este de forma

$$\mathbf{x}^{(m+1)} = \mathbf{g} + \mathbf{M}\mathbf{x}^{(m)}$$

Se iterează, cu aproximația inițială $\mathbf{x}^{(0)} = [1 \quad 4/7 \quad 12/9]^T$; testul de oprire a iterației este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \le 1E-6$. La iterația 12 rezultă soluția $(1.0, 1.0, 1.0)$. Coordonata de modul maxim în $\mathbf{Ax} - \mathbf{b}$ (rezidualul maxim) este 1.43E-06 (ecuația 3).

■

Metoda

Fie sistemul dat $\mathbf{Ax} = \mathbf{b}$. Se rezolvă fiecare ecuație *i* în raport cu necunoscuta x_i . Explicitând x_i se obține:

$$x_i = (b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij} x_j) / a_{ii}, \quad i = 1, 2, \ldots, n$$

S-a presupus $a_{ii} \neq 0$. Iterația Jacobi va fi:

$$\begin{cases} \mathbf{x}^{(0)} = arbitrar \\ x_i^{(m+1)} = (b_i - \sum_{j=1,\, j\neq i}^{n} a_{ij} x_j^{(m)}) / a_{ii}, \quad i = 1, 2, ..., n; \quad m \ge 0 \end{cases}$$

Testul de oprire a iterației este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \le eps$.

Metoda Jacobi se mai zice metoda "iterațiilor simultane", pentru că, coordonatele x_i ale soluției **x** se calculează independent unele de altele. (Pentru alt mod de calcul – v. metoda Metoda Gauss_Seidel, mai jos.)

Condiție suficientă de convergență:

Matricea **A** este *diagonal dominantă*, adică avem:

$$| a_{ii} | > \sum_{j=1,\, j\neq i}^{n} | a_{ij} | , \quad i = \overline{1, n} .$$

■

2 Metoda Gauss_Seidel

Se modifică metoda Jacobi, astfel că, la calculul coordonatei $x_i^{(m)}$ se utilizează valorile $x_1^{(m)}, \ldots, x_{i-1}^{(m)}$ deja calculate, care în general, sunt aproximații mai bune ale soluției.

Pentru exemplul anterior:

$$x_1^{(1)} = 1 - (1/8)x_2^{(0)} + (1/8)x_3^{(0)}$$

$$x_2^{(1)} = 4/7 + (1/7)x_1^{(1)} + (2/7)x_3^{(0)}$$

$$x_3^{(1)} = 12/9 - (2/9)x_1^{(1)} - (1/9)x_2^{(1)}$$

Cu $\mathbf{x}^{(0)} = [0 \quad 0 \quad 0]^T$ şi aceeaşi toleranță $1E-6$, la iterația 9, se obține soluția $(1.0, 1.0, 1.0)$. Rezidualul maxim este 0.0.

■

Formulele generale ale metodei Gauss-Seidel sunt:

$$\begin{cases} \mathbf{x}^{(0)} = arbitrar \\ x_i^{(m+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)}) / a_{ii}, \quad i = 1,2,...,n; \quad m \geq 0 \end{cases}$$

Testul de oprire a iterației este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \leq eps$.

Metoda Gauss-Seidel se mai zice metoda "iterațiilor succesive", pentru că la un pas $m+1$, $m \geq 0$, de îndată ce o coordonată $x_j^{(m+1)}$ a soluției este calculată, ea se utilizează în ecuațiile pentru coordonatele următoare $x_i^{(m+1)}$, $i > j$.

Condiții suficiente de convergență:

- Matricea **A** este *diagonal dominantă*.
- Matricea **A** este *simetrică şi pozitiv definită*.

Când ambele metode Jacobi şi Gauss-Seidel converg, metoda Gauss-Seidel converge mai rapid. Pentru alte considerații privind convergența, v. Cap. 4-IV.

3 SOR - Metoda relaxării (Succesive Over-Relaxation).

Reluăm formula metodei Gauss-Seidel:

$$x_i^{(m+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)}) / a_{ii}, \quad i = 1,2,...,n; \quad m \geq 0$$

Formula se pune sub forma următoare, adunând şi scăzând $\mathbf{x}_i^{(m)}$ în membrul doi (observați că acum, în a doua sumă, indicele *j* ia valori de la *i* şi nu de la $i+1$):

$$x_i^{(m+1)} = x_i^{(m)} + (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(m)}) / a_{ii}, \quad i = \overline{1,n}$$

Termenul care se adună la $\mathbf{x}_i^{(m)}$ este diferența $(\mathbf{x}_i^{m+1} - \mathbf{x}_i^{m})$. Metoda SOR constă în înmulțirea acestei diferențe cu un factor de accelerare (sau relaxare) $\omega > 1$. Întrucât $\omega > 1$, metoda se zice *supra-relaxare*. Alegerea lui ω se discută mai jos. Formula metodei SOR este deci

$$x_i^{(m+1)} = x_i^{(m)} + \omega(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(m)}) / a_{ii}, \quad i = \overline{1,n}$$

Explicitând $\mathbf{x}_i^{(m)}$ din a doua sumă, rezultă:

$$x_i^{(m+1)} = \omega(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)}) / a_{ii} + (1-\omega) x_i^{(m)}, \quad i = \overline{1,n}$$

Se notează expresia din prima paranteză cu $\mathbf{z}_i^{(m+1)}$ (aceasta este coordonata *i* a iteratei $(m+1)$ din metoda Gauss-Seidel.) .

Formula de iterare este echivalentă cu următoarele ecuații considerate pentru $i = \overline{1,n}$:

$$z_i^{(m+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(m)}) / a_{ii},$$

$$x_i^{(m+1)} = \omega\, z_i^{(m+1)} + (1-\omega) x_i^{(m)}$$

În cele mai multe cazuri, valoarea optimă a lui ω satisface relația

$1 < \omega < 2$.

În practică, se poate alege ω astfel: se utilizează valori ω de test, pe un număr limitat de iterații; se alege ca valoare optimă acel ω, pentru care convergența este cea mai rapidă.

III. STABILITATEA SOLUȚIEI şi CONDIȚIONAREA SISTEMULUI

Se consideră sistemul de *n* ecuații liniare

$$\mathbf{Ax} = \mathbf{b} \qquad (1)$$

cu **A** = nesingulară. Se va analiza stabilitatea soluției sistemului, în raport cu o mică variație (perturbare), în:

- Termenii liberi **b**
- Matricea **A** a sistemului

Prima problemă conduce la numărul de condiție al matricii **A**. A doua problemă va conduce la o evaluare a variației soluției, în care intervine numărul de condiție.

1 Perturbare în b. Număr de condiție

Fie sistemul (1), în care termenii liberi **b** suferă o perturbare **r**, devenind $\tilde{\mathbf{b}}$, unde

$\mathbf{r} = \tilde{\mathbf{b}} - \mathbf{b}$, şi $\| \mathbf{r} \| << \| \mathbf{b} \|$. Sistemul perturbat va fi:

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} \qquad (2)$$

Notând perturbarea soluției prin $\mathbf{e} = \tilde{\mathbf{x}} - \mathbf{x}$, şi scăzând (1) din (2) rezultă:

$$\mathbf{Ae} = \mathbf{r}, \qquad (3)$$

din care avem şi:

$$\mathbf{e} = \mathbf{A}^{-1}\mathbf{r} \qquad (3').$$

Pentru a examina stabilitatea soluției **x**, căutăm o evaluare a raportului

$$\frac{\| \mathbf{e} \| / \| \mathbf{x} \|}{\| \mathbf{r} \| / \| \mathbf{b} \|} = \frac{\textit{perturbarea relativã în } \mathbf{x}}{\textit{perturbarea relativã în } \mathbf{b}}.$$

Acest raport exprimă efectul perturbării în **b** asupra soluției **x**, şi anume: După cum raportul este ≈ 1, respectiv >> 1, perturbarea relativă a soluției este de acelaşi ordin, respectiv de ordin mult mai mare, în raport cu perturbarea relativă în **b**.

Luând norma în (3), (3′) avem:

$$\| \mathbf{r} \| \le \| \mathbf{A} \| \cdot \| \mathbf{e} \|, \qquad \| \mathbf{e} \| \le \| \mathbf{A}^{-1} \| \cdot \| \mathbf{r} \|,$$

din care, rezultă:

$$\frac{1}{\| \mathbf{A} \|} \le \frac{\| \mathbf{e} \|}{\| \mathbf{r} \|} \le \| \mathbf{A}^{-1} \| \tag{4}$$

Înmulțind (4) cu $\| \mathbf{b} \| / \| \mathbf{x} \|$, rezultă:

$$\frac{1}{\| \mathbf{A} \|} \cdot \frac{\| \mathbf{b} \|}{\| \mathbf{x} \|} \le \frac{\| \mathbf{e} \| / \| \mathbf{x} \|}{\| \mathbf{r} \| / \| \mathbf{b} \|} \le \| \mathbf{A}^{-1} \| \cdot \frac{\| \mathbf{b} \|}{\| \mathbf{x} \|} \tag{5}$$

Analog, luând norma în (1) şi în $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$, rezultă:

$$\frac{1}{\| \mathbf{A}^{-1} \|} \le \frac{\| \mathbf{b} \|}{\| \mathbf{x} \|} \le \| \mathbf{A} \| \tag{6}$$

Din (5) şi (6) rezultă:

$$\frac{1}{\| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \|} \le \frac{\| \mathbf{e} \| / \| \mathbf{x} \|}{\| \mathbf{r} \| / \| \mathbf{b} \|} \le \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \tag{7}$$

Introducem următoarea

Definiție

Numărul de condiție al matricii **A**, este:

$$Cond(\mathbf{A}) = \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \qquad \blacksquare \tag{8}$$

Numărul de condiție depinde de normă, dar este *mărginit inferior* de 1:

$$Cond(\mathbf{A}) \ge 1 \tag{9}$$

Într-adevăr, din $\mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$ rezultă: $1 \le \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| = Cond(\mathbf{A})$ ■

Cu definiția (8), din (7) rezultă:

$$\frac{1}{Cond(\mathbf{A})} \le \frac{\| \mathbf{e} \| / \| \mathbf{x} \|}{\| \mathbf{r} \| / \| \mathbf{b} \|} \le Cond(\mathbf{A}) \qquad (10)$$

Sau:

$$\frac{1}{Cond(\mathbf{A})} \cdot \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} \le \frac{\| \mathbf{e} \|}{\| \mathbf{x} \|} \le Cond(\mathbf{A}) \cdot \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} \qquad (10')$$

Din această relație rezultă că, pentru $\| \mathbf{r} \| / \| \mathbf{b} \|$ = 'mic', avem:

- Dacă $Cond(\mathbf{A}) \sim 1$: $\| \mathbf{e} \| / \| \mathbf{x} \|$ este 'mic', şi **A** este *bine-condiționată*.
- Dacă $Cond(\mathbf{A}) >> 1$: $\| \mathbf{e} \| / \| \mathbf{x} \|$ *poate fi* 'mare'; **A** este *rău-condiționată*.

Întrucât în Definiția (8) *Cond*(**A**) depinde de normă, se utilizează şi un alt număr de condiție care este independent de normă. Avem $\| \mathbf{A} \| \ge \rho(\mathbf{A})$, unde $\rho(\mathbf{A})$ este raza spectrală a matricii **A**; urmează:

$$Cond(\mathbf{A}) \ge \rho(\mathbf{A}) \cdot \rho(\mathbf{A}^{-1})$$

Sau, definind

$$Cond(\mathbf{A})_* = \rho(\mathbf{A}) \cdot \rho(\mathbf{A}^{-1}) \qquad (11)$$

avem

$$Cond(\mathbf{A}) \ge Cond(\mathbf{A})_*$$

Cum valorile proprii ale matricii $\mathbf{A}^{-1}$ sunt inversele valorilor proprii ale lui **A**, notând cu $\sigma(\mathbf{A})$ spectrul matricii **A**, rezultă formula de calcul:

$$Cond(\mathbf{A})_* = \frac{\max\limits_{\lambda \in \sigma(\mathbf{A})} |\lambda|}{\min\limits_{\lambda \in \sigma(\mathbf{A})} |\lambda|} \qquad (11')$$

Observații asupra calculului lui $Cond(\mathbf{A})_2$

După definiție, avem

$$Cond(\mathbf{A})_2 = \| \mathbf{A} \|_2 \cdot \| \mathbf{A}^{-1} \|_2$$

Cu $\mathbf{B} = \mathbf{A}^T\mathbf{A}$, valorile proprii ale lui **B** sunt reale şi pozitive, şi avem

$\| \mathbf{A} \|_2 = (\lambda_{B,\max})^{1/2}$ (v. 4-I, 3.3 – Observaţii şi Cap. 5).

Analog, cu $\mathbf{D} = \mathbf{A}^{-T}\mathbf{A}^{-1}$, avem $\| \mathbf{A}^{-1} \|_2 = (\lambda_{D,\max})^{1/2}$, şi rezultă

$Cond(\mathbf{A})_2 = (\lambda_{B,\max})^{1/2}(\lambda_{D,\max})^{1/2}$.

Putem evita inversarea matricii A, conform următoarelor proprietăţi:

- Valorile proprii ale lui **D** sunt inversele valorilor proprii ale lui $\mathbf{D}^{-1} = \mathbf{A}\mathbf{A}^T$.

  Punem $\mathbf{C} = \mathbf{A}\mathbf{A}^T$, şi astfel, $\lambda_{D,\max} = 1/\lambda_{C,\min}$. Expresia anterioară devine:

$$Cond(\mathbf{A})_2 = \frac{(\lambda_{B,\max})^{1/2}}{(\lambda_{C,\min})^{1/2}}$$

- Apoi, se arată uşor că $\mathbf{B} = \mathbf{A}^T\mathbf{A}$ şi $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ au *acelaşi valori proprii* (Exerciţiu). Urmează că, în final avem:

$$Cond(\mathbf{A})_2 = \left(\frac{\lambda_{B,\max}}{\lambda_{B,\min}} \right)^{1/2}$$

  Aceasta se mai scrie: $Cond(\mathbf{A})_2 = (Cond(\mathbf{B})_*)^{1/2}$ ■

Exemplu – 1

Fie sistemul liniar:

$$8x_1 + 9x_2 = b_1$$
$$7x_1 + 8x_2 = b_2$$

Avem:

$$\mathbf{A} = \begin{bmatrix} 8 & 9 \\ 7 & 8 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} 8 & -9 \\ -7 & 8 \end{bmatrix}$$

Numerele de condiţie, după diferite norme, sunt:

$Cond(\mathbf{A})_\infty = Cond(\mathbf{A})_1 = 289$,

$Cond(\mathbf{A})_2 \approx 258$,

$Cond(\mathbf{A})_* \approx 254$

a) Detalii pentru calculul lui $Cond(\mathbf{A})_*$:

Polinomul caracteristic al lui **A**, este $p(\lambda) = \lambda^2 - 16\lambda + 1$, de unde $\lambda_{1,2} = 8 \pm \sqrt{63}$.

Urmează $Cond(\mathbf{A})_* = \lambda_1 / \lambda_2 \approx 254$.

b) Detalii pentru calculul lui $Cond(\mathbf{A})_2$:

$\mathbf{B} = \mathbf{A}^T\mathbf{A} = \begin{bmatrix} 8 & 7 \\ 9 & 8 \end{bmatrix}\begin{bmatrix} 8 & 9 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 113 & 128 \\ 128 & 145 \end{bmatrix}$. Polinomul caracteristic al lui **B**, este

$p(\lambda) = \lambda^2 - 258\lambda + 1$. Avem $\lambda_{1,2} = 129 \pm \sqrt{129^2 - 1}$, $\lambda_1 / \lambda_2 = (\lambda_1)^2$. Rezultă

$Cond(\mathbf{A})_2 = \lambda_1 \approx 258$.

Se verifică, în particular, observația anterioară: $\mathbf{C} = \mathbf{A}\mathbf{A}^T = \begin{bmatrix} 145 & 128 \\ 128 & 113 \end{bmatrix}$, are acelaşi

polinom caracteristic ca şi **B**.

Exercițiu: Calculați $Cond(\mathbf{A})_2$, direct după definiția (8).

■

Întrucât numărul de condiție este mare, aceasta arată că sistemul este sensibil la mici schimbări în **b**. Într-adevăr, fie de exemplu, termenii liberi dați şi perturbați, şi soluțiile respective, cum urmează:

$\mathbf{b} = \begin{bmatrix} 17 \\ 15 \end{bmatrix}$; $\quad \widetilde{\mathbf{b}} = \begin{bmatrix} 16.9 \\ 15.1 \end{bmatrix}$; $\quad \mathbf{r} = \widetilde{\mathbf{b}} - \mathbf{b} = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix}$.

$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; $\quad \widetilde{\mathbf{x}} = \begin{bmatrix} -0.7 \\ 2.5 \end{bmatrix}$; $\quad \mathbf{e} = \widetilde{\mathbf{x}} - \mathbf{x} = \begin{bmatrix} -1.7 \\ 1.5 \end{bmatrix}$.

Cu acestea, rezultă:

$\frac{\| \mathbf{r} \|_\infty}{\| \mathbf{b} \|_\infty} = \frac{0.1}{17}$, $\quad \frac{\| \mathbf{e} \|_\infty}{\| \mathbf{x} \|_\infty} = \frac{1.7}{1} = 1.7$, $\quad \frac{\| \mathbf{e} \|_\infty / \| \mathbf{x} \|_\infty}{\| \mathbf{r} \|_\infty / \| \mathbf{b} \|_\infty} = \frac{1.7}{0.1/17} = 289$.

Se observă că schimbări "mici" în **b** produc schimbări "mari" în **x**: perturbarea relativă în **x** este de 289 ori mai mare decât perturbarea relativă în **b**.

În particular, în acest exemplu, marginea superioară din (10) este atinsă – conform $Cond(\mathbf{A})_\infty = 289$.

Un asemenea sistem se zice **rău-condiționat**. Se observă că numerele $Cond(\mathbf{A})$ şi $Cond(\mathbf{A})_*$ sunt o măsură bună pentru condiționarea sistemului.

Exemplu – 2

Există sisteme care nu sunt rău-condiționate, deşi *Cond*(**A**) este mare. De exemplu, să considerăm următoarea matrice, în care *m* este un întreg ≥ 1:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-m} \end{bmatrix}; \quad \mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{m} \end{bmatrix}.$$

Avem $Cond(\mathbf{A})_{1,\infty} = Cond(\mathbf{A})_* = 10^m$. Totuşi sistemul este bine-condiționat.

Valoarea mare a numărului de condiție se elimină prin scalarea matricii **A**.

Exemplu – 3: Interpretare geometrică a condiționării sistemului

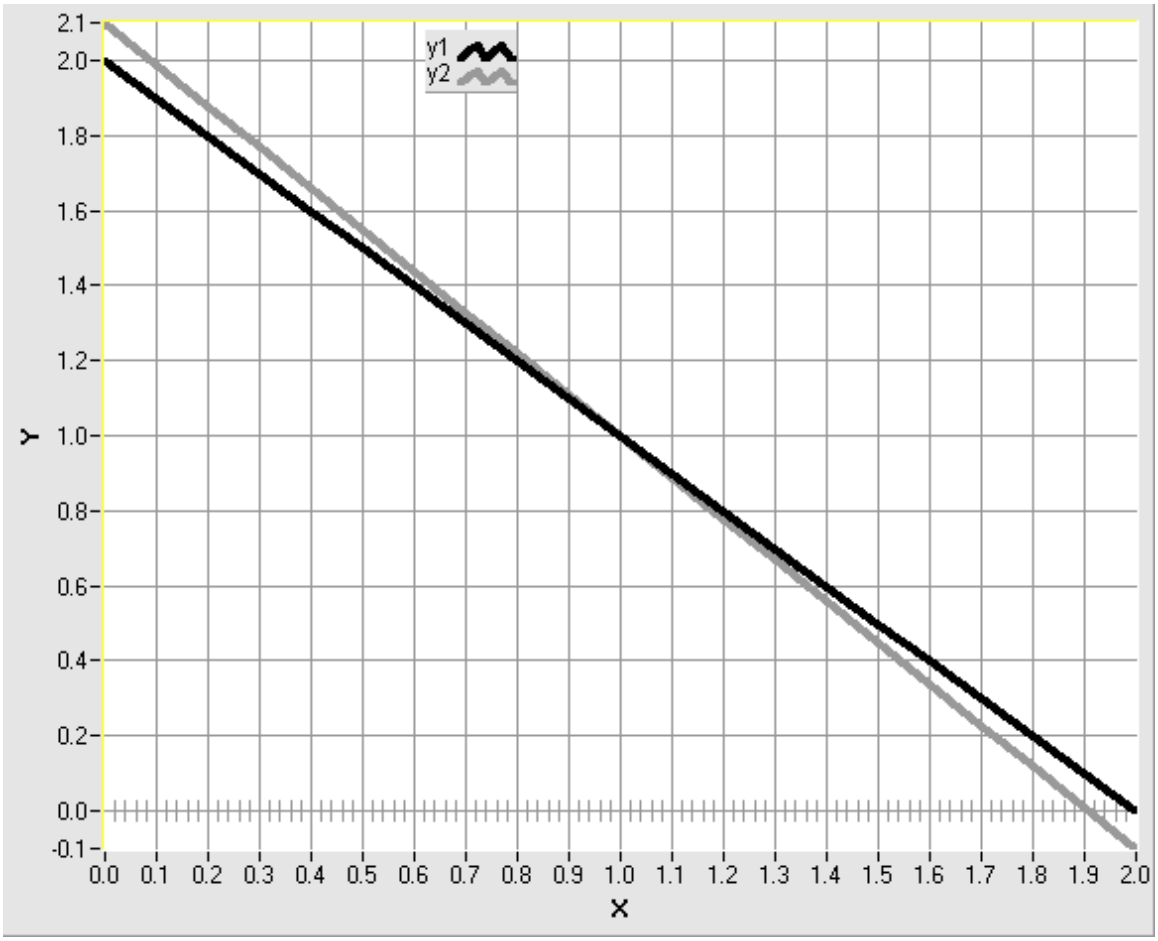
Considerăm sistemul următor:

$$x + y = 2$$
$$1.1x + y = 2.1$$

care are soluția $x = y = 1.0$. Numerele de condiție sunt:

$$Cond(\mathbf{A})_1 = Cond(\mathbf{A})_\infty = 44.10,\ Cond(\mathbf{A})_2 \approx 42.08,\quad Cond(\mathbf{A})_* \approx 41.98,$$

astfel că sistemul este rău-condiționat. Rezolvarea sistemului revine la găsirea intersecției dreptelor reprezentate de cele două ecuații.



Sistem rău condiționat: Interpretare geometrică

Pantele celor două drepte sunt respectiv -1.0 şi -1.1, adică apropiate. Dacă considerăm o incertitudine în coeficienții sistemului, banda de incertitudine a rădăcinii va fi 'largă' – şi cu atât mai largă cu cât pantele sunt mai apropiate. Vezi Figura de mai sus: incertitudinea în valorile $y(x)$ este reprezentată de linia 'groasă' a graficului. În schimb, dacă pantele sunt diferite, intersecția dreptelor este netă şi banda de incertitudine este 'îngustă'.

Exemplu – 4: Matricea Hilbert

Un exemplu de matrice rău-condiționată este următoarea matrice numită matricea Hilbert:

$$\mathbf{H}_n = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & & & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix}$$

Inversa matricii $\mathbf{H}_n$ este cunoscută analitic, şi anume: punând $\mathbf{H}_n^{-1} = [\alpha_{ij}^{(n)}]$, avem – v. Atkinson (1978), Ralston & Rabinowitz (1978):

$$\alpha_{ij}^{(n)} = (-1)^{i+j} \frac{(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2 (n-i)!(n-j)!}, \quad \overline{i,j} = \overline{1,n}$$

Numărul de condiție al matricii $\mathbf{H}_n$ creşte cu *n*, matricea fiind cu atât mai rău condiționată cu cât *n* este mai mare. Exemple:

n	Cond($\mathbf{H}_n$)$_1$
3	7.48 E+02
4	2.84 E+04
5	9.44 E+05
6	2.91 E+07
7	9.85 E+08

Ca exemplu, calculând inversa matricii pentru *n* = 4, prin eliminare Gauss, cu elementele lui $\mathbf{H}_4$ reprezentate în simplă precizie, se obține:

$$\hat{\mathbf{H}}_4^{-1} = \begin{bmatrix} 15.99979 & -119.9980 & 239.9954 & -139.9971 \\ -119.9979 & 1199.981 & -2699.957 & 1679.974 \\ 239.9954 & -2699.958 & 6479.907 & -4199.943 \\ -139.9972 & 1679.974 & -4199.943 & 2799.965 \end{bmatrix}$$

Inversa calculată analitic $\mathbf{H}_4^{-1}$, are elemente intregi: $\alpha_{11} = 16, \ldots, \alpha_{44} = 2800$.

Calculând numărul de condiție al matricii $\mathbf{H}_4$, cu 1-norma, se obține:

$$Cond(\mathbf{H}_4)_1 = \| \mathbf{H}_4 \|_1 \cdot \| \mathbf{H}_4^{-1} \|_1 = (25/12) \cdot 13260 = 28375$$

2 Perturbare în A şi b

Să presupunem că atât matricea **A** a sistemului, cât şi termenul liber **b**, suferă mici schimbări $\delta\mathbf{A}$, respectiv, $\delta\mathbf{b}$, iar soluția devine $\mathbf{x} + \delta\mathbf{x}$:

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

Avem următoarea

Teoremă

Presupunem **A** nesingulară şi fie perturbarea $\delta\mathbf{A}$ satisfăcând condiția

$$\| \delta\mathbf{A} \| < 1 / \| \mathbf{A}^{-1} \|.$$

Atunci:

$$\frac{\| \delta\mathbf{x} \|}{\| \mathbf{x} \|} \leq \frac{Cond(\mathbf{A})}{1 - Cond(\mathbf{A}) \cdot \| \delta\mathbf{A} \| / \| \mathbf{A} \|} \left(\frac{\| \delta\mathbf{A} \|}{\| \mathbf{A} \|} + \frac{\| \delta\mathbf{b} \|}{\| \mathbf{b} \|} \right) \qquad (12)$$

■

Pentru demonstrație – v. Atkinson (1978), Ralston & Rabinowitz (1978).

În particular, dacă $\delta\mathbf{b} = 0$, atunci efectul unei perturbări în **A** este dat de:

$$\frac{\| \delta\mathbf{x} \|}{\| \mathbf{x} \|} \leq \frac{Cond(\mathbf{A})}{1 - Cond(\mathbf{A}) \cdot \| \delta\mathbf{A} \| / \| \mathbf{A} \|} \cdot \frac{\| \delta\mathbf{A} \|}{\| \mathbf{A} \|} \qquad (13)$$

■

În ceea ce urmează vom da rezultate privind efectul erorilor de rotunjire asupra soluției calculate prin eliminarea Gauss.

Rezultatul (13) va fi utilizat în estimarea *a priori* a erorii în eliminarea Gauss.