CURS 3

METODE NUMERICE PENTRU SISTEME DE ECUAŢII LINIARE

Partea II

4 Matrici bandă – simetrice şi pozitiv definite

4.1 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-lăţimea* de bandă. *LIM* reprezintă numărul elementelor din semi-bandă, inclusiv elementul diagonal.

Observaţie

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-lăţimea de bandă *maximă* ■

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

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

4.2 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:

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

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

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

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

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

Adresa unui element a_{ij}, în vectorul a, este dată de o funcţie 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).

4.3 Metoda Cholesky

Proprietatea esenţială este următoarea:

> Dacă matricea bandă este *simetrică şi pozitiv definită*, atunci descompunerea $\mathbf{LL}^T$ 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.

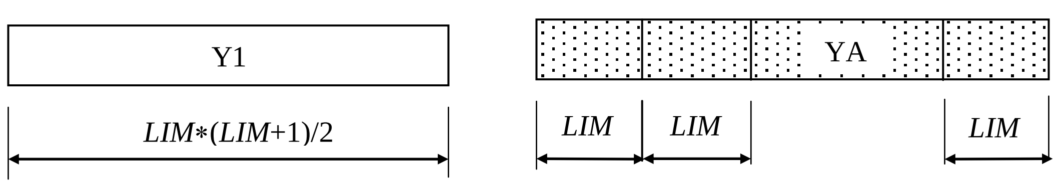
Paşii sunt cei de la Cholesky, lucrul cu matricea S.

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 $n*LIM$: iniţial, stochează matricea **A**, pe linii. În cursul descompunerii Cholesky, stochează liniile procesate ale matricii A.

5. METODE ITERATIVE

5.1 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 ecuaţie i = 1, 2, 3, în raport cu necunoscuta x_i, căutând ca aceasta să fie necunoscuta cu coeficientul cel mai mare din ecuaţie, eventual rearanjând ecuaţiile.
Intervertind ecuaţiile 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 aproximaţia iniţială $\mathbf{x}^{(0)} = [1 \quad 4/7 \quad 12/9]^T$; testul de oprire a iteraţiei este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \le 1E-6$. La iteraţia 12 rezultă soluţia $(1.0, 1.0, 1.0)$. Coordonata de modul maxim în $\mathbf{Ax} - \mathbf{b}$ (rezidualul maxim) este 1.43E-06 (ecuaţia 3).

■

Metoda

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

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

S-a presupus $a_{ii} \neq 0$. Iteraţia 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 iteraţiei este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \le eps$.

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

Condiţie suficientă de convergenţă:

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

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

■

5.2 Metoda Gauss_Seidel

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

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 toleranţă $1E-6$, la iteraţia 9, se obţine soluţia $(1.0, 1.0, 1.0)$. 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 iteraţiei este $\| \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} \|_\infty \leq eps$.

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

Condiţii suficiente de convergenţă:

- 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 consideraţii privind convergenţa, v. Cap. 4-IV.

5.3 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 (observaţi 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 diferenţa $(\mathbf{x}_i^{m+1} - \mathbf{x}_i^{m})$. Metoda SOR constă în înmulţirea acestei diferenţe 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 ecuaţii 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 relaţia

$1 < \omega < 2$.

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

6. STABILITATEA SOLUŢIEI şi CONDIȚIONAREA SISTEMULUI

Se consideră sistemul de n ecuaţii liniare

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

cu $\mathbf{A}$ = nesingulară. Se va analiza stabilitatea soluţiei sistemului, în raport cu o mică variaţie (perturbare), în:

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

Prima problemă conduce la numărul de condiţie al matricii **A**. A doua problemă va conduce la o evaluare a variaţei soluţiei, în care intervine numărul de condiţie.

6.1 Perturbare în b. Număr de condiţie

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:

$$\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} \tag{2}$$

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

$$\mathbf{Ae} = \mathbf{r}, \tag{3}$$

din care avem şi:

$$\mathbf{e} = \mathbf{A}^{-1}\mathbf{r} \tag{3'}.$$

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

$$\frac{\| \mathbf{e} \| / \| \mathbf{x} \|}{\| \mathbf{r} \| / \| \mathbf{b} \|} = \frac{perturbarea\ relativ\tilde{a}\ \hat{i}n\ \mathbf{x}}{perturbarea\ relativ\tilde{a}\ \hat{i}n\ \mathbf{b}}.$$

Acest raport exprimă efectul perturbării în **b** asupra soluţiei **x**, şi anume: După cum raportul este ≈ 1, respectiv >> 1, perturbarea relativă a soluţiei 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:

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

Înmulţind (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

Definiţie

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

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

Numărul de condiţie 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 definiţia (8), din (7) rezultă:

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

Sau:

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

Din această relaţie 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-condiţionată*.
- Dacă $Cond(\mathbf{A}) >> 1$: $\|\mathbf{e}\| / \|\mathbf{x}\|$ *poate fi* 'mare'; **A** este *rău-condiţionată*.

Întrucât în Definiţia (8) *Cond*(**A**) depinde de normă, se utilizează şi un alt număr de condiţie 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

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

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:

$$Cond(\mathbf{A})_* = \frac{\max_{\lambda \in \sigma(\mathbf{A})} |\lambda|}{\min_{\lambda \in \sigma(\mathbf{A})} |\lambda|} \tag{11'}$$

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

După definiţie, 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 *aceleaşi 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:

$$\begin{aligned} 8x_1 + 9x_2 &= b_1 \\ 7x_1 + 8x_2 &= b_2 \end{aligned}$$

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 calculului 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 calculului 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, observaţia anterioară: $\mathbf{C} = \mathbf{A}\mathbf{A}^T = \begin{bmatrix} 145 & 128 \\ 128 & 113 \end{bmatrix}$, are acelaşi polinom caracteristic ca şi **B**.

Exerciţiu: Calculaţi $Cond(\mathbf{A})_2$, direct după definiţia (8).

■

Întrucât numărul de condiţie este mare, aceasta arată că sistemul este sensibil la mici schimbări în **b**. Într-adevăr, fie de exemplu, termenii liberi daţi şi perturbaţi, şi soluţiile respective, cum urmează:

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

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

Cu acestea, rezultă:

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

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-condiţionat**. Se observă că numerele $Cond(\mathbf{A})$ şi $Cond(\mathbf{A})_*$ sunt o măsură bună pentru condiţionarea sistemului.

Exemplu – 2

Există sisteme care nu sunt rău-condiţionate, 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-condiţionat. Valoarea mare a numărului de condiţie se elimină prin scalarea matricii **A**.

Exemplu – 3: Interpretare geometrică a condiţionării sistemului

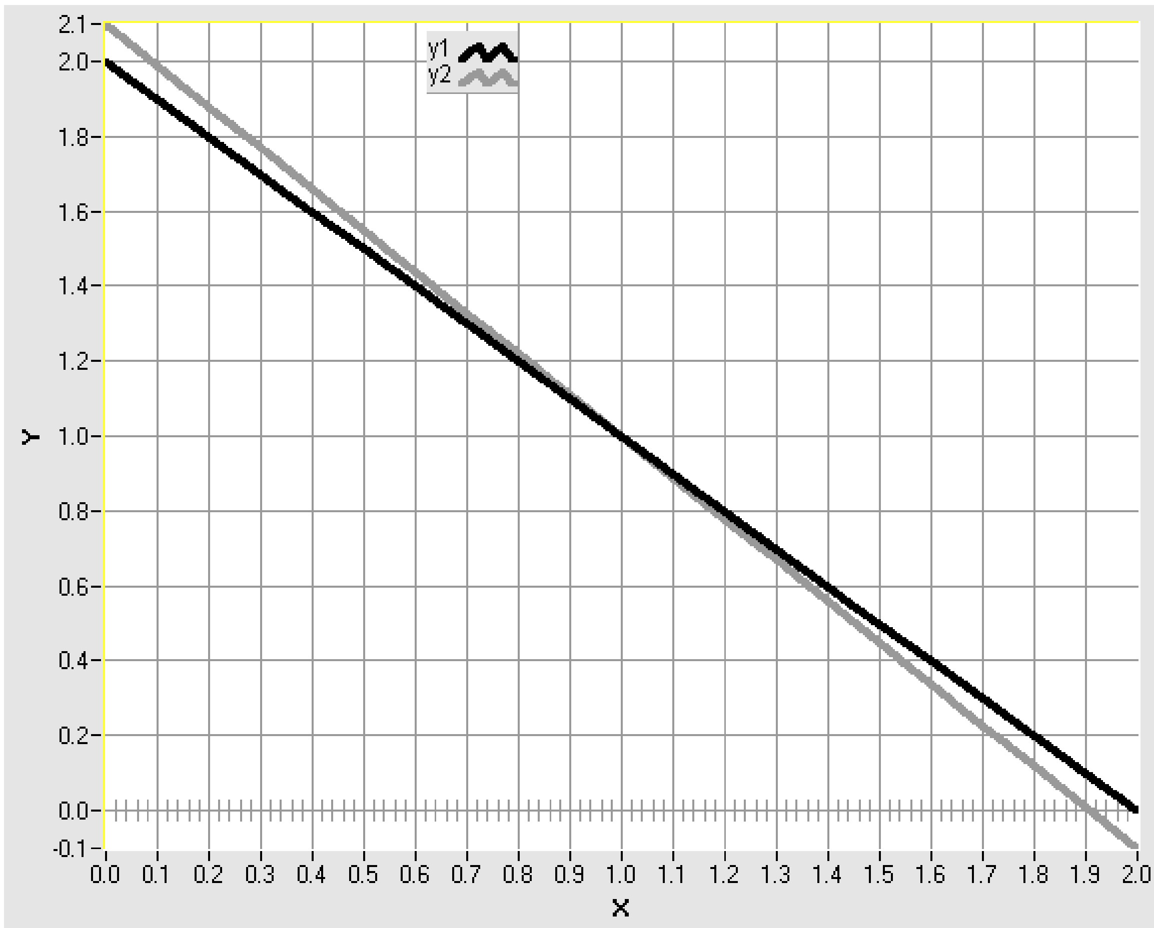
Considerăm sistemul următor:

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

care are soluţia $x = y = 1.0$. Numerele de condiţie 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-condiţionat. Rezolvarea sistemului revine la găsirea intersecţiei dreptelor reprezentate de cele două ecuaţii.



Sistem rău condiţionat: Interpretare geometrică

Pantele celor două drepte sunt respectiv -1.0 şi -1.1, adică apropiate. Dacă considerăm o incertitudine în coeficienţii 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, intersecţia dreptelor este netă şi banda de incertitudine este 'îngustă'.

Exemplu – 4: Matricea Hilbert

Un exemplu de matrice rău-condiţionată 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 condiţie al matricii $\mathbf{H}_n$ creşte cu *n*, matricea fiind cu atât mai rău condiţionată 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 obţine:

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

Inversa calculată analitic $\mathbf{H}_4^{-1}$, are elemente intregi: $\alpha_{11} = 16, \ldots, \alpha_{44} = 2800$. Calculând numărul de condiţie al matricii $\mathbf{H}_4$, cu 1-norma, se obţine:

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

6.2 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 soluţia 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 condiţia

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

Atunci:

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

■

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

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

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

■

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

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