Lezione 20 (ore 38,39,40) - 11 novembre 2025, 15:30 - 18:30 F3

(2.41) Esempio (continuazione).

Supponiamo che le costanti elastiche c_k siano note con incertezza. Assumiamo, ad esempio, che, per k = 1,2,3, sia:

$$c_k' = c_k + \delta c_k \qquad \text{con} \qquad |\delta c_k| < 1 \text{ N/m}$$

Il sistema A z = b si trasforma nel *sistema perturbato* $(A + \delta A)\, z = b + \delta b$ con:

$$\delta A = \begin{bmatrix} \delta c_1 + \delta c_2 & -\delta c_2 \\ -\delta c_2 & \delta c_2 + \delta c_3 \end{bmatrix} \quad , \quad \delta b = \begin{bmatrix} 0 \\ h\, \delta c_3 \end{bmatrix}$$

Per le perturbazioni dei dati si ha:

$$\varepsilon_A = N_1(\delta A)/N_1(A) < 10^{-2} \quad , \quad \varepsilon_b = N_1(\delta b)/N_1(b) < 10^{-2}$$

Inoltre:

$$c_1(A)\, \varepsilon_A \leqslant 3 \times 10^{-2}$$

In base al Teorema di condizionamento, per lo scostamento della soluzione $\hat{z}$ del sistema perturbato dalla soluzione z^* si ha la limitazione:

$$\varepsilon_x \leqslant \frac{c_1(A)}{1 - c_1(A)\, \varepsilon_A} (\varepsilon_A + \varepsilon_b) \approx 6.2 \times 10^{-2}$$

Per quanto riguarda lo scostamento delle componenti si ha, questa volta:

$$\varepsilon_{x,1} \leqslant 0.19 \ (19\,\%) \quad , \quad \varepsilon_{x,2} \leqslant 0.09 \ (9\,\%)$$

(2.42) Esempio (continuazione).

Sia adesso $\hat{z}$ una colonna (ad esempio ottenuta dal calcolatore utilizzando una procedura per la soluzione del sistema A z = b) da usare come *approssimazione* di z^*. Per ottenere una limitazione dell'errore commesso si procede come nell'Osservazione (2.39) della Lezione 19.

Il vettore residuo è:

$$r = A\, \hat{z} - b$$

(1) Si interpreta $\hat{z}$ come soluzione del sistema perturbato A z = b + r. Per il Teorema di condizionamento:

$$\frac{N_1(\hat{z} - z^*)}{N_1(z^*)} \leqslant c_1(A)\, \frac{N_1(r)}{N_1(b)}$$

Domanda: esistono perturbazioni dei parametri δg, δc_k, δm_k, δh che generano perturbazioni dei dati $\delta A = 0$ e $\delta b = r$ (ovvero: si riesce ad 'interpretare fisicamente' il sistema perturbato A z = b + r) ?

Osservazione: la limitazione trovata è valida *indipendentemente dalla risposta* alla domanda: il sistema perturbato *non deve necessariamente essere fisicamente significativo*.

Risposta: si. Ad esempio: $\delta g = 0$, $\delta c_k = 0$, $\delta h = 0$ e $\delta m_1 = r_1/g$, $\delta m_2 = r_2/g$.

(2) Si cerca $M \in R^{2\times 2}$ tale che $M\hat{z} = -r$, e si interpreta $\hat{z}$ come soluzione del sistema perturbato $(A + M)\, z = b$. Per il Teorema di condizionamento, posto $\varepsilon_A = \| M \|_1 / \| A \|_1$:

$$\text{se} \quad c_1(A)\varepsilon_A \leqslant 1 \text{ allora } \frac{N_1(\hat{z}-z^*)}{N_1(z^*)} \leqslant \frac{c_1(A)\varepsilon_A}{1-c_1(A)\varepsilon_A}$$

Domanda: esistono perturbazioni dei parametri δg, δc_k, δm_k, δh che generano perturbazioni dei dati $\delta A = M$ e $\delta b = 0$ (ovvero: si riesce ad 'interpretare fisicamente' il sistema perturbato $A\, z = b + r$) ?

(2.43) Esempio.

Sia:

$$\hat{z} = \begin{bmatrix} 1.8 \\ 3.4 \end{bmatrix} \text{ m}$$

Allora:

$$r = A\,\hat{z} - b = \begin{bmatrix} 10.19 \\ -9.81 \end{bmatrix} \text{ N}$$

Cerchiamo α e β in modo che, posto:

$$M = \begin{bmatrix} \alpha+\beta & -\beta \\ -\beta & \beta \end{bmatrix}$$

si abbia:

$$M\hat{z} = -r$$

Si ottiene un sistema di due equazioni nelle incognite α e β la cui unica soluzione è:

$$\alpha = -(r_1 + r_2)/\hat{z}_1 \approx -0.21 \text{ N/m} \quad \text{e} \quad \beta = -r_2/(\hat{z}_2 - \hat{z}_1) \approx -6.13 \text{ N/m}$$

Si ottiene allora:

$$\varepsilon_A \approx 4.1 \times 10^{-2} \quad \text{e} \quad c_1(A)\,\varepsilon_A \approx 0.12 < 1$$

da cui, per il Teorema di condizionamento:

$$\varepsilon_x \leqslant \text{circa } 0.14$$

Infine, la risposta è si: $\delta g = 0$, $\delta m_k = 0$, $\delta h = 0$ e $\delta c_1 = \alpha$ N/m, $\delta c_2 = \beta$ N/m, $\delta c_3 = 0$.

(2.2) STUDIO DI UN SISTEMA DI EQUAZIONI LINEARI IN F(β,m)

(2.44) Osservazione (studio con EGP).

Siano $A \in R^{n \times n}$ e $b \in R^n$. Il procedimento per lo studio del sistema $A\,x = b$ che usa la procedura EGP è:

```
(S,D,P) = EGP(A);
      se esiste k tale che d_kk = 0 allora STOP;
        altrimenti
             c = SA(S,Pb);
             x* = SI(D,c)
```

in R

Quando si utilizza un calcolatore, con insieme di numeri di macchina F(β,m), la procedura si trasforma in:

```
(Ŝ,D̂,P̂) = EGP_M(Â);
      se esiste k tale che d̂_kk = 0 allora STOP;
        altrimenti
             ĉ = SA_M(S,Pb̂);
             x̂ = SI_M(D,c)
```

in F(β,m)

dove:

- EGP_M, SA_M e SI_M sono, rispettivamente, la procedura EGP, SA ed SI in cui ciascuna operazione aritmetica è sostituita dalla corrispondente funzione predefinita,
- $\hat{A}$ e $\hat{b}$ sono, rispettivamente, la matrice rd(A) e la colonna rd(b) di elementi gli arrotondati in F(β,m) dei corrispondenti elementi di A e b.

(2.45) Esempio.

Ricordando il Teorema (1.38) della Lezione 6, per ciascuna componente della matrice $\hat{A} = rd(A)$ e della colonna $\hat{b} = rd(b)$ si ha:

$$\hat{a}_{ij} = rd(a_{ij}) = (1 + \varepsilon_{ij})\, a_{ij} \quad , \quad \hat{b}_i = rd(b_i) = (1 + \varepsilon_i)\, b_i$$

con $|\varepsilon_{ij}| \leqslant u$ e $|\varepsilon_i| \leqslant u$ per ogni i e j. Ne segue che, utilizzando ad esempio la norma uno in R^n, per le misure assolute delle perturbazioni si ha:

$$\| \delta A \|_1 \leqslant u \| A \|_1 \quad , \quad N_1(\delta b) \leqslant u\, N_1(\delta b)$$

e quindi, per le misure relative:

$$\varepsilon_A \leqslant u \quad \text{e} \quad \varepsilon_b \leqslant u$$

Se $c_1(A)\, u \leqslant 1$ allora $c_1(A)\, \varepsilon_A \leqslant 1$ e, per il Teorema di condizionamento (Teorema (2.36) della Lezione 18) si ha:

$$\varepsilon_x \leqslant 2\,\frac{c_1(A)\,u}{1-c_1(A)\,u} \equiv \Lambda$$

Quando il calcolatore *legge i dati* A e b, li cambia (salvo il caso in cui le componenti dei dati siano in $F(\beta,m)$) e il sistema $A\,x = b$ è sostituito dal sistema $\hat{A}\,x = \hat{b}$. Questa sostituzione, nel caso migliore possibile in cui sia trascurabile l'effetto delle sostituzioni di EGP, SA ed SI con EGP_M, SA_M e SI_M, *può generare* uno scostamento della soluzione x^* di misura relativa Λ. Dunque, nel caso usuale in cui l'effetto delle sostituzioni di EGP, SA ed SI con EGP_M, SA_M e SI_M non è trascurabile, *non è ragionevole* aspettarsi uno scostamento tra x^* e l'approssimazione $\hat{x}$ ottenuta dal calcolatore *minore di* Λ.

(2.46) Esempio.

Si consideri la seguente situazione 'quasi ideale':

- $\hat{A} = A$, $\hat{b} = b$ - i dati hanno componenti in $F(\beta,m)$);
- $EGP_M(A) = EGP(A) = (S,D,P)$ - la fattorizzazione EGP_M è esatta, con D invertibile;
- $SA_M(S,Pb) = \hat{c} = rd(c)$ - il risultato di SA_M è 'quasi ideale';
- $SI_M(D,\hat{c}) = SI(D,\hat{c})$ - il risultato di SI_M è esatto.

Sotto queste ipotesi si ha: $x^* = SI(D,c)$ è la soluzione del sistema $D\,x = c$, $\hat{x} = SI(D,\hat{c})$ è la soluzione del sistema $D\,x = \hat{c}$. Introdotta la perturbazione $\delta c = \hat{c} - c$ si ha, utilizzando la norma uno (vedi l'esempio precedente):

$$N_1(\delta c) \leqslant u\,N_1(c) \qquad \text{e quindi} \qquad \varepsilon_c \leqslant u$$

Per il Teorema di condizionamento si ha allora:

$$\varepsilon_x \leqslant c_1(D)\,\varepsilon_c \leqslant c_1(D)\,u$$

La limitazione della misura relativa dello scostamento dipende da $c_1(D)$ ovvero, posto:

$$c_1(D) = c_1(A)\,\frac{c_1(D)}{c_1(A)}$$

dal *fattore di amplificazione del numero di condizionamento* $c_1(D)/c_1(A)$.

(2.47) Esempio.

Siano $\gamma \in (0,1)$ e $A = \begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix}$. Si ha:

- $\| A \|_1 = 1 + \gamma < 2$
- $A^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\gamma \end{bmatrix}$ da cui $\| A^{-1} \|_1 = 1 + \gamma$ e $c_1(A) = (1 + \gamma)^2 < 4$
- $EGP(A) = (S,D,P) = \left(\begin{bmatrix} 1 & 0 \\ 1/\gamma & 1 \end{bmatrix}, \begin{bmatrix} \gamma & 1 \\ 0 & -1/\gamma \end{bmatrix}, I \right)$ e $\| D \|_1 = 1 + 1/\gamma$
- $D^{-1} = \begin{bmatrix} 1/\gamma & 1 \\ 0 & -\gamma \end{bmatrix}$ da cui $\| D^{-1} \|_1 = \max\{ 1/\gamma, 1 + \gamma \}$ e $c_1(D) = (1 + 1/\gamma)\max\{ 1/\gamma, 1 + \gamma \}$

Per il fattore di amplificazione del numero di condizionamento si ha allora:

$$\lim_{\gamma \to 0} \frac{c_1(D)}{c_1(A)} = +\infty$$

Dunque: scegliendo γ *sufficientemente piccolo* è possibile ottenere un fattore di amplificazione del numero di condizionamento *grande quanto si vuole*: il procedimento di soluzione del sistema di equazioni che usa EGP trasforma il sistema A x = b nel sistema *equivalente* D x = c ma mentre le proprietà di condizionamento di A sono *buone* ($c_1(A) < 4$) quelle di D, scelto γ opportunamente piccolo, sono *pessime* ($c_1(D)$ enorme).

Mentre il procedimento di soluzione del sistema di equazioni che usa EGP è *soddisfacente* quando si opera in R (si veda (2.16) della Lezione 17), il procedimento può risultare *non soddisfacente* quando si opera in F(β,m).

(2.48) Definizione (procedura EGPP).

Per ovviare al potenziale pericolo evidenziato nell'esempio precedente, si ricorre ad una modifica della procedura EGP che porta alla definizione della procedura EGPP (Eliminazione di Gauss con Pivoting Parziale). La differenza con EGP consiste *solo* nella *scelta della matrice di permutazione* P_k. Nella procedura EGP si ha:

se $A_k(k,k) \neq 0$ allora $P_k = I$ altrimenti
se esiste $i > k$ tale che $A_k(i,k) \neq 0$ allora $P_k = P_{k,i}$ altrimenti $P_k = I$

Nella procedura EGPP si pone:

se per ogni $i \geqslant k$ si ha $A_k(i,k) = 0$ allora $P_k = I$ altrimenti
scelto i tale che $|A_k(i,k)| = \max \{ |A_k(j,k)| , j \geqslant k \}$ si pone $P_k = P_{k,i}$

La scelta nella procedura EGP ha lo scopo di assicurarsi che il pivot sia *diverso da zero*, nella procedura EGPP lo scopo è quello di avere come pivot *l'elemento della colonna* k*-esima di modulo massimo possibile* tra tutti quelli con indice di riga $j \geqslant k$.

(2.49) Esempio.

Calcolo di EGPP(A) con:

```
A = [ 1, 0, 1;
      2, 1,-1;
      1, 2, 1 ]
```

(*) $A_1 = A$;

(*) $k = 1$; $|A_1(2,1)| = \max \{ |A_1(j,1)| , j \geqslant 1 \} \Rightarrow P_1 = P_{1,2}$;

$$T_1 = P_1 A_1 = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} \quad , \quad H_1 = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_2 & 1 & 0 \\ \lambda_3 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix}$$

I valori λ_2, λ_3 sono determinati come nella procedura EGP.

Infine:

$$H_1 T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & -1/2 & 3/2 \\ 0 & 3/2 & 3/2 \end{bmatrix} = A_2$$

(*) $k = 2$; $|A_2(3,2)| = \max \{ |A_2(j,2)| \ , \ j \geqslant 2 \} \Rightarrow P_2 = P_{2,3}$;

$$T_2 = P_2 A_2 = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} \quad , \quad H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix}$$

Il valore λ_3 è determinato come nella procedura EGP.

Infine:

$$H_2 T_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & 0 & 2 \end{bmatrix} = A_3$$

(*) $D = A_3$; $P = P_2 P_1$; $S = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & -1/3 & 1 \end{bmatrix}$ (ricavata come in EGP).

(2.50) Osservazione.

Per ogni $A \in R^{n \times n}$ invertibile, posto (S,D,P) = EGPP(A), si ha: $c_1(D)/c_1(A) \leqslant F(n)$. La funzione F dipende *solo* dalla dimensione n della matrice e dalla norma scelta, in particolare *non dipende* da A. Dunque, il fattore di crescita del numero di condizionamento è *limitato*.

Tornando all'Esempio (2.47) si ha:

$$\text{EGPP}\left(\begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix} \right) = (S,D,P) \text{ con } D = I \Rightarrow c_1(D) = 1$$

(2.51) Osservazione (studio con qr).

Siano $A \in R^{n \times n}$ e $b \in R^n$. Il procedimento per lo studio del sistema $Ax = b$ che usa la procedura qr è:

(U,T) = qr(A);
se esiste k tale che $t_{kk} = 0$ allora STOP;
altrimenti
$c = U^t b$;
$x^* = SI(D,c)$

in R

Quando si utilizza un calcolatore, con insieme di numeri di macchina $F(\beta,m)$, la procedura si trasforma in:

$(\hat{U},\hat{T}) = qr_M(\hat{A})$;
  se esiste k tale che $\hat{t}_{kk} = 0$ allora STOP;
    altrimenti
      $\hat{c} = \hat{U}^t \otimes \hat{b}$;
      $\hat{x} = SI_M(\hat{T},\hat{c})$

in $F(\beta,m)$

dove $\hat{U}^t \otimes b$ è la colonna che si ottiene sostituendo in $U^t b$ le operazioni aritmetiche con le corrispondenti funzioni predefinite in $F(\beta,m)$.

(2.52) Esempio.

Analogamente a quanto fatto per il procedimento che usa EGP, si consideri la seguente situazione 'quasi ideale':

- $\hat{A} = A$, $\hat{b} = b$ - i dati hanno componenti in $F(\beta,m)$);
- $qr_M(A) = qr(A) = (U,T)$ - la fattorizzazione qr_M è esatta, con T invertibile;
- $U^t \otimes b = \hat{c} = rd(c)$ - il risultato di $U^t \otimes b$ è 'quasi ideale';
- $SI_M(T,\hat{c}) = SI(T,\hat{c})$ - il risultato di SI_M è esatto.

Sotto queste ipotesi si ha: $x^* = SI(T,c)$ è la soluzione del sistema $T\,x = c$, $\hat{x} = SI(T,\hat{c})$ è la soluzione del sistema $T\,x = \hat{c}$. Introdotta la perturbazione $\delta c = \hat{c} - c$ si ha, utilizzando la norma due (la norma 'naturale' da utilizzare in R^n quando si utilizza la fattorizzazione QR che fa entrare in gioco la nozione di ortogonalità, dunque il prodotto scalare in R^n, è la norma due: quella indotta dal prodotto scalare):

$$N_2(\delta c) \leqslant u\, N_2(c) \qquad \text{e quindi} \qquad \varepsilon_c \leqslant u$$

Per il Teorema di condizionamento si ha ancora:

$$\varepsilon_x \leqslant c_2(T)\, \varepsilon_c \leqslant c_2(T)\, u$$

e la limitazione della misura relativa dello scostamento dipende dal fattore di amplificazione del numero di condizionamento $c_2(T)/c_2(A)$.

Però in questo caso si ha:

- $A = U\,T \Rightarrow \| A \|_2 = \| U\,T \|_2 = \max \{ N_2(U\,T\,v),\ N_2(v) = 1 \} =^1 \max \{ N_2(T\,v),\ N_2(v) = 1 \} = \| T \|_2$
- $T^{-1} = A^{-1}\,U \Rightarrow \| T^{-1} \|_2 = \| A^{-1}\,U \|_2 = \max \{ N_2(\,A^{-1}\,U\,v),\ N_2(v) = 1 \} =^2 \max \{ N_2(A^{-1}\,w),\ N_2(U^t\,w) = 1 \} = \max \{ N_2(A^{-1}\,w),\ N_2(w) = 1 \} = \| A^{-1} \|_2$

Ne segue che $c_2(T) = c_2(A)$, ovvero il fattore di amplificazione del numero di condizionamento è $c_2(T)/c_2(A) = 1$.

Il procedimento di soluzione del sistema di equazioni che usa qr è soddisfacente *anche* quando si opera in $F(\beta,m)$.

---

1 Poiché U è ortogonale si ha: $N_2(U\,T\,v) = \text{sqrt}(\ v^t\,T^t\,U^t\,U\,T\,v\ ) = \text{sqrt}(\ v^t\,T^t\,T\,v\ ) = N_2(T\,v)$.

2 Cambio di variabile: $w = U\,v$. Essendo U ortogonale si ha poi $v = U^t\,w$ e $N_2(U^t\,w) = N_2(w)$.

(2.3) COSTO DELLA SOLUZIONE DI UN SISTEMA DI EQUAZIONI LINEARI

(2.53) Definizione (costo aritmetico).

Un metodo per confrontare i due procedimenti descritti per ottenere un'approssimazione della soluzione di un sistema di equazioni lineari (quello che usa EGPP e quello che usa qr) è di considerare il tempo necessario per il calcolo dell'approssimazione.

Nel contesto della risoluzione dei sistemi di equazioni lineari, si introduce la seguente nozione di *costo* del calcolo di $\varphi(x)$, $C(\varphi)$, dove φ è l'*algoritmo ingenuo* (si veda Definizione (1.32), Lezione 6) per f:

$C(\varphi)$ = il numero di operazioni aritmetiche necessario per calcolare f

(2.54) Osservazione (ragionevolezza della definizione di costo).

Perché $C(\varphi)$ sia indicativo del *tempo* necessario per il calcolo di $\varphi(x)$ è necessario che siano soddisfatte le seguenti due condizioni:

(1) Durante il calcolo di $\varphi(x)$, il tempo impiegato in attività che *non* siano l'esecuzione di operazioni aritmetiche (ovvero: nel calcolo di funzioni predefinite corrispondenti a funzioni elementari o confronti) *deve essere trascurabile* (un esempio di algoritmo in cui questa condizione *non è* verificata è quello che calcola la norma infinito di un vettore: in questo caso l'algoritmo esegue solo *confronti* tra le componenti del vettore);
(2) Il tempo di calcolo di ciascuna delle funzioni predefinite corrispondenti ad operazioni aritmetiche deve essere *indipendente dagli operandi.*

La seconda condizione *non è verificata*, ad esempio, nel caso della moltiplicazione tra due elementi di $F(\beta,m)$: per calcolare $\xi_1 \otimes \xi_2$ occorre *moltiplicare le frazioni* - e questo avviene in un tempo indipendente dai fattori perché le frazioni hanno sempre *lo stesso numero di cifre* - e *sommare gli esponenti*; è quest'ultima operazione che non può essere ritenuta indipendente dai fattori perché gli esponenti sono numeri interi qualsiasi che hanno *un numero di cifre che dipende da quali elementi di* $F(\beta,m)$ *si considerano.* In particolare, perché la nozione di costo aritmetico sia indicativa del tempo necessario per il calcolo occorre che *l'insieme dei numeri di macchina del calcolatore non sia* $F(\beta,m)$.