

I numerisk analys har man dessutom varianter av en faktorisering beroende på matrisens egenskaper. Detta för att spara datorminne och beräkningstid.

Några exempel för LU-faktoriseringen.

Om A är symmetrisk, $A^T = A$, så behövs halva minnet och beräkningstiden. Om A dessutom är positivt definit (positiva egenvärden) kan man förenkla metoden ytterligare.

Det finns motsvarande varianter om matrisen är komplex.

- $A^H = A$, A är Hermitsk ($A^H = \bar{A}^T$)
- $A^T = A$, A är komplexsymmetrisk

Många element i en matris kan vara noll, en gles matris. Detta kan man utnyttja för att spara minne och beräkningstid. Det är viktigt eftersom glesa matriser brukar vara stora, med en dimension om $10^4 - 10^6$ kanske. Ett specialfall av en gles matris är en sk bandmatris, här två små exempel:

$$\begin{bmatrix} \times & \times & \times & 0 & 0 \\ \times & \times & \times & \times & 0 \\ 0 & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix} \quad \begin{bmatrix} \times & \times & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 \\ 0 & \times & \times & \times & 0 \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times \end{bmatrix}$$

Den högra matrisen är ett exempel på en sk tridiagonal matris. Den kan vara osymmetrisk, symmetrisk, symmetrisk positivt definit, Hermitsk, komplexsymmetrisk etc. Dessutom brukar det finnas stöd för enkel- och dubbel precision liksom för komplex och dubbel komplex.

Man inser att det kan bli många varianter av LU-faktoriseringen (många olika rutiner i programbiblioteken, t.ex. Lapack).

LU-faktorisering

Now [let] there be

hemp 1 dōu, wheat 3 dōu, beans 2 dōu, peas 8 dōu, millet 5 dōu, worth 95 coins.

hemp 2 dōu, wheat 5 dōu, beans 3 dōu, peas 9 dōu, millet 4 dōu, worth 112 coins;

hemp 3 dōu, wheat 5 dōu, beans 7 dōu, peas 6 dōu, millet 4 dōu, worth 116 coins;

hemp 7 dōu, wheat 6 dōu, beans 4 dōu, peas 5 dōu, millet 3 dōu, worth 128 coins;

hemp 9 dōu, wheat 7 dōu, beans 3 dōu, peas 2 dōu, millet 5 dōu, worth 140 coins;

Question: how much is 1 dōu [of each] worth?

The answer says:

hemp 1 dōu 7 coins,

wheat 1 dōu 4 coins,

beans 1 dōu 3 coins,

peas 1 dōu 5 coins,

millet 1 dōu 6 coins.

En dōu \approx 2 liter.

Detta exempel är hämtat ur kapitel 8 i den cirka 2000 år gamla boken "The Nine Chapters on the Mathematical Art (Jiuzhang Suanshu)" Boken behandlar 246 problem i 9 kapitel. Boken är den mest betydelsefulla kinesiska matematiska klassikern.

$$\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix} \begin{bmatrix} h \\ w \\ b \\ p \\ m \end{bmatrix} = \begin{bmatrix} 95 \\ 112 \\ 116 \\ 128 \\ 140 \end{bmatrix}$$

I Matlab:

```
>> x = A \ b
x =
    7.0000
    4.0000
    3.0000
    5.0000
    6.0000
```

Kineserna utnyttjar det som vi kallar Gausselimination (\approx 1800). Kombinerar rader så att vi till slut har en triangulär matris.

$$\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix} x = \begin{bmatrix} 95 \\ 112 \\ 116 \\ 128 \\ 140 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & -4 & 1 & -18 & -11 \\ 0 & -15 & -10 & -51 & -32 \\ 0 & -20 & -15 & -70 & -40 \end{bmatrix} x = \begin{bmatrix} 95 \\ -78 \\ -169 \\ -537 \\ -715 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 5 & 54 & 58 \\ 0 & 0 & 5 & 70 & 80 \end{bmatrix} x = \begin{bmatrix} 95 \\ -78 \\ 143 \\ 633 \\ 845 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 0 & 44 & 45 \\ 0 & 0 & 0 & 60 & 67 \end{bmatrix} x = \begin{bmatrix} 95 \\ -78 \\ 143 \\ 490 \\ 702 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 0 & 44 & 45 \\ 0 & 0 & 0 & 0 & \frac{62}{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 95 \\ -78 \\ 143 \\ 490 \\ 372/11 \end{bmatrix}$$

Vi löser det triangulära problemet med så kallad bakåtsubstitution:

$$\begin{aligned} x_5 &= (372/11)/(62/11) = 6 \\ x_4 &= (490 - 45x_5)/44 = 5 \\ x_3 &= (143 - 10x_4 - 13x_5)/5 = 3 \\ x_2 &= (-78 - (-1)x_3 - (-7)x_4 - (-6)x_5)/(-1) = 4 \\ x_1 &= (95 - 3x_2 - 2x_3 - 8x_4 - 5x_5)/1 = 7 \end{aligned}$$

Kan formulera eliminationen som en serie matrismultiplikationer.

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ -3 & 0 & 1 & 0 & 0 \\ -7 & 0 & 0 & 1 & 0 \\ -9 & 0 & 0 & 0 & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix}}_A = \begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & -4 & 1 & -18 & -11 \\ 0 & -15 & -10 & -51 & -32 \\ 0 & -20 & -15 & -70 & -40 \end{bmatrix}$$

Elementen $-L_1(2 : 5, 1)$ (dvs. 2, 3, 7, 9) kallas multiplikatorer.

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & -4 & 1 & 0 & 0 \\ 0 & -15 & -10 & 1 & 0 \\ 0 & -20 & -15 & 0 & 1 \end{bmatrix}}_{L_2} \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix}}_A = \begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 5 & 54 & 58 \\ 0 & 0 & 5 & 70 & 80 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}}_{L_3} \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix}}_A = \begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 0 & 44 & 45 \\ 0 & 0 & 0 & 60 & 67 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{-15}{11} & 1 \end{bmatrix}}_{L_4} \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 0 & 44 & 45 \\ 0 & 0 & 0 & 0 & \frac{62}{11} \end{bmatrix}}_U$$

Vi får

$$L_4 L_3 L_2 L_1 A = U$$

eller

$$A = (L_4 L_3 L_2 L_1)^{-1} U$$

Vi noterar (se övningarna):

- L_k är inverterbar
- L_k^{-1} ser nästan ut som L_k

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \alpha & 1 & 0 \\ 0 & \beta & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\alpha & 1 & 0 \\ 0 & -\beta & 0 & 1 \end{bmatrix}$$

- det är enkelt att multiplicera L_k -matriser:

L1 =

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 4 & 0 & 0 & 1 \end{bmatrix}$$

L2 =

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 20 & 1 & 0 \\ 0 & 30 & 0 & 1 \end{bmatrix}$$

>> L1*L2

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 20 & 1 & 0 \\ 4 & 30 & 0 & 1 \end{bmatrix}$$

>> L2*L1

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 43 & 20 & 1 & 0 \\ 64 & 30 & 0 & 1 \end{bmatrix}$$

37

Om vi sammanställer allt detta får vi med andra ord

$$\underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 2 & 5 & 3 & 9 & 4 \\ 3 & 5 & 7 & 6 & 4 \\ 7 & 6 & 4 & 5 & 3 \\ 9 & 7 & 3 & 2 & 5 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 4 & 1 & 0 & 0 \\ 7 & 15 & 1 & 1 & 0 \\ 9 & 20 & 1 & \frac{15}{11} & 1 \end{bmatrix}}_{L^{-1}L_2^{-1}L_3^{-1}L_4^{-1}} \underbrace{\begin{bmatrix} 1 & 3 & 2 & 8 & 5 \\ 0 & -1 & -1 & -7 & -6 \\ 0 & 0 & 5 & 10 & 13 \\ 0 & 0 & 0 & 44 & 45 \\ 0 & 0 & 0 & 0 & \frac{62}{11} \end{bmatrix}}_U$$

Detta kallas LU-uppdelning. L är undertriangulär och U övertriangulär.

För att konstruera L plockar vi alltså in multiplikatorerna från de olika L_k på respektive plats i L .

Vad är det för fördel med detta jämfört med vanlig GE (Gausselimination)? Svar: enklare att hantera vid teoretiskt arbete. Gör det möjligt att lösa problem av typen $Ax_k = b_k$ där b_{k+1} beror av x_k . (Om alla högerleden är kända på en gång kan givetvis vanlig GE utnyttjas).

Så normalt löses $Ax = b$ i de tre stegen

1. beräkna L och U så att $A = LU$
för att lösa $LUx = b$ inför vi beteckningen $z = Ux$ och får då problemet $Lz = b$
2. lös $Lz = b$ (framåtsubstitution)
3. lös $Ux = z$ (bakåtsubstitution)

Framåtsubstitution går till på samma vis som bakåtsubstitutionen fast man tar raderna i omvänd ordning.

Kostnad?

$A = LU$ tar ungefär $n^3/3$ vardera av + och *
 $Lz = b$ kostar $n^2/2$ vardera av + och *
($Ux = z$ kostar lika mycket).

38

Ett numeriskt exempel

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 4 & -2 & 12 \\ 3 & -7 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}$$

Vi startar med en "tom" $L = I$ och fyller i multiplikatorerna eftersom under diagonalen.

$$A = \begin{bmatrix} \boxed{1} & -1 & 2 \\ \boxed{4} & -2 & 12 \\ \boxed{3} & -7 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 2 \\ 0 & \boxed{2} & 4 \\ 0 & \boxed{-4} & -5 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & 2 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} = U$$

Så, det som skall stå i L är symboliskt $\frac{\square}{\square}$ och U är det som blir kvar av A efter trianguleringen.

Alltså:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \boxed{4} & 1 & 0 \\ \boxed{3} & \boxed{-2} & 1 \end{bmatrix} \quad \text{och} \quad U = \begin{bmatrix} 1 & -1 & 2 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix}$$

Lös $Lz = b$, ger:

$$z = \begin{bmatrix} 3 \\ -6 \\ -12 \end{bmatrix}$$

Lös $Ux = z$, ger:

$$x = \begin{bmatrix} 16 \\ 5 \\ -4 \end{bmatrix}$$

Kontroll:

$$Ax = \begin{bmatrix} 1 & -1 & 2 \\ 4 & -2 & 12 \\ 3 & -7 & 1 \end{bmatrix} \begin{bmatrix} 16 \\ 5 \\ -4 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} = b, \quad \text{OK!}$$

39

Är detta en stabil algoritm?

Här följer en grov skiss som visar vad som kan gå fel.

Låt ϵ stå för ett litet tal. a_1, a_2 och a_3 markerar "medelstora" tal. LU-faktorisering blir då:

$$\underbrace{\begin{bmatrix} \epsilon & a_2 \\ a_1 & a_3 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 0 \\ a_1/\epsilon & 1 \end{bmatrix}}_L \underbrace{\begin{bmatrix} \epsilon & a_2 \\ 0 & a_3 - a_2(a_1/\epsilon) \end{bmatrix}}_U$$

a_1/ϵ blir ett stort tal, vilket ger utskiftning i beräkningen av $u_{2,2} = a_3 - a_2(a_1/\epsilon)$. Låt oss anta att hela a_3 skiftas ut och att allt annat räknas ut exakt. Hur stort blir bakåttelet?

$$\underbrace{\begin{bmatrix} 1 & 0 \\ a_1/\epsilon & 1 \end{bmatrix}}_{\text{ber. } L} \underbrace{\begin{bmatrix} \epsilon & a_2 \\ 0 & -a_2(a_1/\epsilon) \end{bmatrix}}_{\text{ber. } U} = \underbrace{\begin{bmatrix} \epsilon & a_2 \\ a_1 & 0 \end{bmatrix}}_{\text{faktoriserad matris}}$$

Vi har alltså faktoriserat en matris som avviker mycket från A i (2,2)-elementet. Algoritmen behöver inte vara stabil.

Det kan vi dock lätt fixa. Kasta om raderna i systemet (byt ordning på ekvationerna), dvs. studera matrisen $B = PA$:

$$B = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_P A = \begin{bmatrix} a_1 & a_3 \\ \epsilon & a_2 \end{bmatrix}$$

LU-faktorisering blir nu:

$$\begin{bmatrix} a_1 & a_3 \\ \epsilon & a_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \epsilon/a_1 & 1 \end{bmatrix} \begin{bmatrix} a_1 & a_3 \\ 0 & a_2 - a_3(\epsilon/a_1) \end{bmatrix}$$

Notera att ϵ/a_1 är ett litet tal. Vi får alltså inte farlig utskiftning i $u_{2,2}$.

40

Slutligen ett exempel från flervariabelkursen. Låt $z = f(x, y)$ vara en reellvärd funktion av två variabler. Vi vill undersöka om f har ett strängt lokalt minimum i punkten (a, b) . Om f är tillräckligt snäll (har tillräckligt många kontinuerliga derivator) gäller att:

$$f(a+h, b+k) = f(a, b) + f'_x(a, b)h + f'_y(a, b)k + \frac{f''_{xx}(a, b)h^2 + 2f''_{xy}(a, b)hk + f''_{yy}(a, b)k^2}{2} + \dots$$

Ett nödvändigt villkor för minimum är att gradienten är nollvektorn, ty annars kan vi göra f mindre genom att gå i negativa gradientens riktning. Alltså gäller:

$$f(a+h, b+k) = f(a, b) + \frac{f''_{xx}(a, b)h^2 + 2f''_{xy}(a, b)hk + f''_{yy}(a, b)k^2}{2} + \dots$$

Nu är (där vi inte skriver ut (a, b))

$$f''_{xx}h^2 + 2f''_{xy}hk + f''_{yy}k^2 = f''_{xx}h^2 + f''_{xy}hk + f''_{yx}kh + f''_{yy}k^2 = \begin{bmatrix} h & k \end{bmatrix} \begin{bmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} = v^T H v, \text{ med } v = \begin{bmatrix} h \\ k \end{bmatrix}, H = \begin{bmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{bmatrix}$$

H är den sk Hessianen. Om H är positivt definit så har f ett strängt lokalt minimum i (a, b) .

Detta gäller allmänt. Om $w = f(x, y, z)$ där $\nabla f(a, b, c)$ är nollvektorn, så har f ett strängt lokalt minimum i (a, b, c) om

$$\begin{bmatrix} f''_{xx} & f''_{xy} & f''_{xz} \\ f''_{yx} & f''_{yy} & f''_{yz} \\ f''_{zx} & f''_{zy} & f''_{zz} \end{bmatrix}$$

är positivt definit (alla derivator är beräknade i (a, b, c)).

Notera också att om A är en $n \times n$ -matris och x en kolonnvektor så gäller att:

$$\sum_{j=1}^n \sum_{k=1}^n a_{j,k} x_j x_k = x^T A x$$

45

Konditionstalet för $Ax = b$ -problemet "Proof by example"

Låt oss se hur lösningen x ändrar sig när vi stör högerledet b . Vi studerar ett numeriskt exempel där A är diagonal.

$$\begin{bmatrix} 2 & 0 \\ 0 & 10^{-10} \end{bmatrix} x = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \Rightarrow x = \begin{bmatrix} 1 \\ 10^{10} \end{bmatrix}$$

Vi stör nu b med f och får då lösningen y , dvs. $Ay = b + f$. Hur mycket ändras x , dvs. hur stor är $y - x$?

$$y = A^{-1}(b + f) = A^{-1}b + A^{-1}f = x + A^{-1}f$$

så att

$$y - x = A^{-1}f = \begin{bmatrix} 1/2 & 0 \\ 0 & 10^{10} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} 0.5 f_1 \\ 10^{10} f_2 \end{bmatrix}$$

Det är inte alltid så här illa. Om vi i stället tar koefficientmatrisen

$$B = \begin{bmatrix} 2 & 0 \\ 0 & 0.1 \end{bmatrix}$$

så blir

$$y - x = B^{-1}f = \begin{bmatrix} 0.5 f_1 \\ 10 f_2 \end{bmatrix}$$

Slutsats: om A har ett eller flera diagonalelement nära noll, så kommer x att vara känslig för ändringar i högerledet.

A är "nästan singular" i följande mening:

$$\underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 10^{-10} \end{bmatrix}}_A + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & -10^{-10} \end{bmatrix}}_E = \underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}}_{\text{singulär}}$$

46

Den lilla störningen E (små element jämfört med det största elementet i A) gör A singular. A ligger alltså nära en singular matris.

B är inte nästan singular eftersom E måste innehålla ett stort element, -0.1 .

Allmänt gäller att x är känslig för störningar i b och A om A är nästan singular. Om A är långt från att vara singular, så är x relativt okänslig för störningar.

En nästan singular matris har en invers där åtminstone något element är stort. Eftersom $y - x = A^{-1}f$ så kommer x att ändras mycket om A^{-1} innehåller stora element.

Om matrisen inte är diagonal får vi ett mer komplicerat uppträdande. Antag att $\delta > 0$ är nära noll.

$$C = \begin{bmatrix} 1 & 1 \\ 1 & 1 + \delta \end{bmatrix} \Rightarrow C^{-1} = \frac{1}{\delta} \begin{bmatrix} 1 + \delta & -1 \\ -1 & 1 \end{bmatrix}$$

Vi ser att C^{-1} är proportionell mot $1/\delta$, så inversen är stor. C ligger också nära en singular matris, ty om vi subtraherar δ från $c_{2,2}$ så blir matrisen singular. Detta gäller allmänt.

Om C har element av storleksordningen ett:

storlek på $C^{-1} \approx 1 / \text{avståndet till närmaste singulära matris}$

För att göra riktiga satser krävs mer matematik, vektor- och matrisnormer.

47

Vektornormer

En vektornorm är en funktion som ger ett mått på storleken på elementen i en vektor. Om vektorn innehåller n element så sammanfattar vi storleken med ett ickenegativt tal, så normer kan vara trubbiga mätverktyg.

Det finns oändligt många normer. Vi kommer att använda tre så kallade L_p -normer som vi betecknar med $\|\cdot\|_p$:

$$\|x\|_p = \left[\sum_{k=1}^n |x_k|^p \right]^{1/p}, \quad p > 0$$

- $p = 1$, $\|x\|_1 = \sum_{k=1}^n |x_k|$, ettnormen
- $p = 2$, $\|x\|_2 = \left[\sum_{k=1}^n x_k^2 \right]^{1/2}$, tvånormen
- $p = \infty$, $\|x\|_\infty = \max_{1 \leq k \leq n} |x_k|$, maxnormen

Dessa tre normer (liksom alla vektornormer) uppfyller:

- $x \neq 0 \Rightarrow \|x\| > 0$ (positivitet), $\|0\| = 0$
- $\|\alpha x\| = |\alpha| \|x\|$ för alla $\alpha \in \mathfrak{R}$ (homogenitet)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangelolikheten)

Normer är olika stora

$$x = \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix}, \quad \|x\|_1 = 6, \quad \|x\|_2 = \sqrt{14}, \quad \|x\|_\infty = 3$$

48

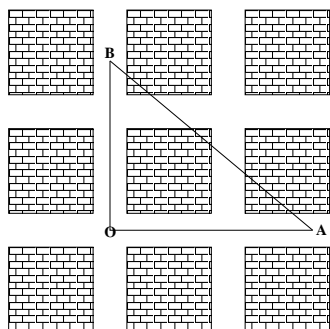
Varför räcker det inte med den "vanliga längden" av en vektor, dvs. det vi kallar tvånormen?

Det beror på att olika problemställningar kräver olika sätt att mäta storlek. Dessutom kan det vara så att det går att skapa starkare satser för en viss norm.

Exempel:

Antag att vi befinner oss i en stad där kvarteren ligger i ett rutnät. Vi vill ta oss den kortaste vägen från A till B. Låt O beteckna origo. Om vi kan flyga så är avståndet $\|B - A\|_2$. Om vi måste följa gatorna så är avståndet $\|B - A\|_1$ (dvs. längden av \overline{OA} plus längden av \overline{OB}).

$\|B - A\|_\infty$ kan tolkas som den största förflyttning som vi gör sidledes.



Innerprodukter

Vi kan definiera en norm givet en innerprodukt (skalärprodukt).

$$x \cdot y = (x, y) = \sum_{k=1}^n x_k y_k = x^T y$$

Så

$$\|x\|_2 = \sqrt{x^T x}$$

Notera att $x^T y$ är en skalär men $x y^T$ är en matris.

Exempel:

$$\begin{bmatrix} -1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = 4, \quad \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -2 & -1 \\ 6 & 4 & 2 \\ 9 & 6 & 3 \end{bmatrix}$$

Om $x \neq 0$ så säger vi att vektorn $x/\|x\|$ är normerad (har längd ett).

Definition av vinkel

$$x^T y = \|x\|_2 \|y\|_2 \cos \phi$$

Cauchy-Schwarz olikhet

$$|x^T y| \leq \|x\|_2 \|y\|_2$$

Matrisnormer

Matrisnormer är funktioner från $\mathfrak{R}^{m \times n}$ till \mathfrak{R} och uppfyller de tre vektornormsvillkoren ovan.

Användbara matrisnormer är dessutom submultiplikativa (konsistenta):

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

Tre olika normer kan vara inblandade ovan. Vi använder samma beteckning $\|\cdot\|$ för alla tre.

Vi kan bilda matrisnormer utgående från vektornormer.

En operatornorm mäter hur mycket multiplikation med en matris A kan förstora en vektor:

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

Vi noterar att $\|I\| = 1$ om $\|\cdot\|$ är en operatornorm.

Operatornormerna som svarar mot våra tidigare vektornormer:

- $p = 1$, $\|A\|_1 = \max_k \sum_{r=1}^m |a_{r,k}|$, ettnormen, största kolonnsumman
- $p = 2$, $\|A\|_2 = \max[\lambda(A^T A)]^{1/2}$ tvånormen
- $p = \infty$, $\|A\|_\infty = \max_r \sum_{k=1}^n |a_{r,k}|$, maxnormen, största radsumman

Exempel:

$$A = \begin{bmatrix} 1 & -2 & -3 \\ 6 & 4 & 2 \\ 9 & -6 & 3 \end{bmatrix}, \quad \|A\|_1 = 16, \quad \|A\|_2 \approx 11.9042, \quad \|A\|_\infty = 18$$

Konditionstal för $Ax = b$ -problemet

Vi skall nu använda normer för att studera konditionstalet för $Ax = b$ -problemet. Vi vill alltså studera vad som händer med x när vi ändrar A och b . Vi kommer endast att ändra b .

Sats: Antag att A är ickesingulär och att $Ax = b \neq 0$.

Om $Ay = b + f$ så gäller att

$$\frac{\|x - y\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|f\|}{\|b\|}$$

Bevis:

$$Ay = b + f \quad \text{och} \quad Ax = b \Rightarrow A(y - x) = f \Rightarrow$$

$$y - x = A^{-1}f \Rightarrow \|y - x\| = \|A^{-1}f\| \leq \|A^{-1}\| \|f\|$$

Men $Ax = b$ så att $\|A\| \|x\| \geq \|b\|$ eller $1/\|x\| \leq \|A\|/\|b\|$. ■

Man kan bevisa likartade satser för fallen när A eller A och b störs. Normalt betecknas konditionstalet med kappa, dvs. $\kappa(A) = \|A\| \|A^{-1}\|$.

Antag att $\|\cdot\|$ är en operatornorm, då gäller:

- $\kappa(A) \geq 1$ för alla A , ty $1 = \|AA^{-1}\| \leq \|A\| \|A^{-1}\|$
- I är perfekt konditionerad ty $\kappa(I) = 1$
- konditionstalet är skalningsoberoende $\kappa(\alpha A) = \kappa(A)$
- $\kappa(A) = \infty$ om A är singulär

Om A är singulär kan det finnas ingen eller oändligt många lösningar. Vi förväntar oss problem om A är näst singulär.

Om $\kappa(A)$ är stort så finns en matris, E , med liten norm $\|E\|$, så att $A + E$ är exakt singulär. A "ligger nära" mängden av singulära matriser, matrisen är "nästan" singulär.

Om $\kappa(A)$ är litet måste man ändra A mycket (stor E) för att $A + E$ skall bli singulär.

Man kan visa att de E som gör $A + E$ singulär och som har minsta norm uppfyller $\|E\| = \|A\|/\kappa(A)$.

Determinanten för A inte är något bra mått på nästan singulär.

Exempel:

$$\det(\alpha I) = \alpha^n, \quad \kappa(\alpha I) = 1$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad \det(A) = 0.1, \quad \kappa(A) = 10$$

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}, \quad \det(A) = 0.001, \quad \kappa(A) = 10$$

53

Exempel: Hur väl stämmer satsen?

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-8} \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \kappa_\infty(A) = 10^8$$

Tag

$$f = \begin{bmatrix} 10^{-9} \\ 10^{-9} \end{bmatrix} \Rightarrow y - x = \begin{bmatrix} 10^{-9} \\ 10^{-1} \end{bmatrix}, \quad \frac{\|x - y\|_\infty}{\|x\|_\infty} = \frac{0.1}{1} = 0.1$$

$$\kappa_\infty(A) \frac{\|f\|_\infty}{\|b\|_\infty} = 10^8 \frac{10^{-9}}{1} = 0.1$$

Så likhet i gränsen. Tag nu i stället

$$f = \begin{bmatrix} 10^{-9} \\ 0 \end{bmatrix} \Rightarrow y - x = \begin{bmatrix} 10^{-9} \\ 0 \end{bmatrix}, \quad \frac{\|x - y\|_\infty}{\|x\|_\infty} = \frac{10^{-9}}{1} = 10^{-9}$$

$$\kappa_\infty(A) \frac{\|f\|_\infty}{\|b\|_\infty} = 10^8 \frac{10^{-9}}{1} = 0.1$$

vilket ger en enorm överskattning.

54

Tolkning av satsen

Antag att elementen i x (resp. y) är ungefär lika stora. Då gäller att $x \approx x_k e$, där e är vektorn av ettor.

Analogt gäller att $y \approx y_k e$. Alltså:

$$\frac{\|x - y\|}{\|x\|} \approx \frac{\|(x_k - y_k)e\|}{\|x_k e\|} = \frac{|x_k - y_k|}{|x_k|}$$

Så normen uppskattar då det elementvisa felet. Felet kan då, i detta specialfall, begränsas enligt:

$$\frac{|x_k - y_k|}{|x_k|} \lesssim \kappa(A) \frac{\|f\|}{\|b\|}$$

Detta säger att relativa felet i varje lösningskomponent begränsas av det relativa felet i indata multiplicerat med $\kappa(A)$.

Om t.ex. $\|f\|/\|b\| = 0.5 \cdot 10^{-k}$ (k decimaler) och $\kappa(A) \approx 10^p$ så är

$$\frac{|x_k - y_k|}{|x_k|} \lesssim 10^p \cdot 0.5 \cdot 10^{-k} = 0.5 \cdot 10^{p-k}$$

så vi har tappat p siffror i svaret, vilket ger följande tumregel:

Om $\kappa(A) = 10^p$ så riskerar vi att tappa p siffror.

Antag nu att x innehåller element av olika storleksordning, t.ex. $x = [1, 10^{-3}]^T$ och att vi använder $\|\cdot\|_\infty$. Om $p - k = -3$ gäller att:

$$\max\{|1 - y_1|, |10^{-3} - y_2|\} \leq 0.5 \cdot 10^{-3}$$

så att

$$1 - 0.5 \cdot 10^{-3} \leq y_1 \leq 1 + 0.5 \cdot 10^{-3}$$

och

$$10^{-3} - 0.5 \cdot 10^{-3} \leq y_2 \leq 10^{-3} + 0.5 \cdot 10^{-3}$$

Normer kan vara trubbiga instrument.

55

Hur stora fel har vi i indata? Låt oss se på två fall.

Exakta indata: i detta fall får vi eventuellt avrundningsfel när $a_{j,k}$ och b_k lagras i datorns minne. Relativa felet (per komponent) är ungefär ϵ_{mach} . Vi får också avrundningsfel när vi löser $Ax = b$ -problemet.

Vi kan nog tillåta tämligen stora $\kappa(A)$, men det beror givetvis på hur många siffror vi behöver. Om vi har stora krav eller om $\kappa(A)$ är mycket stort, så kan vi minska ϵ_{mach} genom att t.ex. använda Maple eller Mathematica (räkna med fler siffror). Att räkna med många siffror går dock mycket långsammare (mjukvara och inte hårdvara).

Indata med osäkerhet (mätdata): detta fall ger normalt större begränsningar på hur stora $\kappa(A)$ som kan tillåtas, eftersom vi normalt inte mäter så väldigt noga.

Att räkna med mindre ϵ_{mach} ger normalt inte en bättre lösning eftersom $\kappa(A)$ är måttligt stort och mätfelen helt dominerar över avrundningsfelen.

Hur kan vi uppskatta $\kappa(A)$?

Att beräkna A^{-1} tar mycket tid och minne om A är stor. `cond` i Matlab använder `svd` för $\|\cdot\|_2$ och `explicit inv` för andra normer.

För stora matriser kan man använda `condest` som uppskattar $\|A^{-1}\|$ genom att lösa linjära ekvationssystem (+ listigheter).

Även LAPACK kan ge en sådan uppskattning när man löser $Ax = b$. Uppskattningen kostar nästan inget eftersom man utnyttjar den LU-faktorisering som redan beräknats.

56

Vad säger residualen $r = b - A\hat{x}$? \hat{x} beräknad lösning.

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-8} \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{x} = \begin{bmatrix} 1 \\ 10^4 \end{bmatrix}, \quad r = \begin{bmatrix} 0 \\ -10^{-4} \end{bmatrix}$$

Kan visa:

$$(A + E)\hat{x} = b, \quad \|E\|_2 = \|r\|_2 / \|\hat{x}\|_2$$

$\|E\|_2 \approx 10^{-8}$ i exemplet.

Så en liten residual betyder att vi löst nästan rätt problem.

I framåtriktningen kommer $\kappa(A)$ in (antag att A^{-1} existerar):

$$r = b - A\hat{x} = Ax - A\hat{x} = A(x - \hat{x}) \Leftrightarrow x - \hat{x} = A^{-1}r$$

Alltså gäller:

$$\|x - \hat{x}\| \leq \|A^{-1}\| \|r\|$$

Om $b \neq 0$ gäller:

$$\|b\| \leq \|A\| \|x\| \Leftrightarrow \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

Om vi kombinerar de två olikheterna erhåller vi:

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

Så felet i lösningen kan vara godtyckligt stort även om residualen är liten.

57

Minstakvadratproblem

Ett ofta återkommande problem är att vi har en matematisk modell och uppmätta värden, och vill bestämma parametrar i modellen.

En mycket vanlig modell är $b = ce^{\lambda t}$ (halveringstid, befolkningstillväxt, urladdning av kondensator). b kan vara befolkningen vid en given tidpunkt t . c är befolkningmängden vid tiden $t = 0$.

Antag att vi vill bestämma parametern λ genom att mäta b vid m olika tidpunkter. Vi alltså har m par (t_k, b_k) , $k = 1, \dots, m$ av mätvärden. Låt oss anta att vi känner c .

Hur ska vi beräkna λ ? Vi har ju m olika ekvationer

$$b_1 = ce^{\lambda t_1}, \quad b_2 = ce^{\lambda t_2}, \dots, \quad b_m = ce^{\lambda t_m}$$

och vi lär inte kunna hitta ett λ som satisfierar alla ekvationerna. Det är inte intressant att få m olika λ -värden.

En rimlig kompromiss är att hitta ett λ som approximativt satisfierar alla ekvationerna, dvs:

$$b_1 \approx ce^{\lambda t_1}, \quad b_2 \approx ce^{\lambda t_2}, \dots, \quad b_m \approx ce^{\lambda t_m}$$

Detta kan formuleras på följande vis:

Försök att göra alla avvikelserna (residualerna)

$$ce^{\lambda t_1} - b_1, \quad ce^{\lambda t_2} - b_2, \dots, \quad ce^{\lambda t_m} - b_m$$

så små (när noll) som möjligt.

Vi hade lika gärna kunnat studera avvikelserna $b_k - ce^{\lambda t_k}$.

58

Vi kan definiera "små" på oändligt många sätt, till exempel:

$$\min_{\lambda} \sum_{k=1}^m |ce^{\lambda t_k} - b_k|$$

$$\min_{\lambda} \left[\sum_{k=1}^m [ce^{\lambda t_k} - b_k]^2 \right]^{1/2} \quad \min_{\lambda} \max_{1 \leq k \leq m} |ce^{\lambda t_k} - b_k|,$$

Dessa förslag är inte slumpvis valda, utan låt oss införa en vektor, r , av alla residualerna (en residualvektor):

$$r = \begin{bmatrix} ce^{\lambda t_1} - b_1 \\ ce^{\lambda t_2} - b_2 \\ \vdots \\ ce^{\lambda t_m} - b_m \end{bmatrix}$$

Vår tre mått kan då skrivas:

$$\min_{\lambda} \|r\|_1, \quad \min_{\lambda} \|r\|_2, \quad \min_{\lambda} \|r\|_{\infty}$$

Vi kommer normalt att få olika värden på λ beroende på vilken norm vi utnyttjar. Varje λ är dock bäst för den givna normen.

Det finns oändligt många frågor, varje fråga med sitt svar; varje svar är dock ett korrekt svar på den givna frågan.

Det finns normalt inte ett bästa λ -värde.

59

Man kan givetvis ha modeller med flera parametrar. Ett vanligt problem är att anpassa en serie mätpunkter till en rät linje.

Vår modell kan då skrivas; $b = x_1 + x_2 t$. Här är x_1 och x_2 parametrar och (t_k, b_k) är uppmätta värden.

Hur ser residualvektorn ut i detta fall?

$$r = \begin{bmatrix} x_1 + x_2 t_1 - b_1 \\ x_1 + x_2 t_2 - b_2 \\ \vdots \\ x_1 + x_2 t_m - b_m \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_x - \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_b$$

Dvs. $r = Ax - b$. Vi vill alltså lösa minimeringsproblemet:

$$\min_x \|Ax - b\|$$

i någon lämplig norm.

Observera att detta normalt inte är ett linjärt ekvationssystem. Vi löser inte $Ax = b$, ty detta går normalt inte, eftersom vi får avvikelser i alla ekvationerna. Lite slarvigt kan man skriva $Ax \approx b$. Om vi kan lösa $Ax = b$ så är ju residualvektorn $r = Ax - b$ nollvektorn, varför mätpunkterna följer modellen exakt. Notera också att matrisen A har fler rader än kolonner.

När residualvektorn kan skrivas $r = Ax - b$ säger vi att problemet är linjärt. Modellen har utseendet:

$$b = \text{uttryck}_1 \text{ parameter}_1 + \dots + \text{uttryck}_n \text{ parameter}_n$$

där uttryck_k beror av mätvärdena och inte beror av någon parameter.

Vår första modell är icke-linjär eftersom parametern λ inte ingår linjärt i modellen.

60

I vissa fall kan vi via substitutioner eller andra transformationer skapa en linjär modell utifrån en icke linjär sådan. Vår första modell är enkel att transformera, förutsatt att b och c har samma tecken. Låt oss anta att både b och c är positiva:

$$b = ce^{\lambda t} \Leftrightarrow \log b = \log c + \lambda t$$

λ ingår nu linjärt i modellen.

Om vi nu antar att c inte är känd (vi mätte aldrig b för $t = 0$) så är c en parameter som nu ingår icke linjärt i modellen. Om vi sätter $x_1 = \log c$ har vi dock en linjär modell som är identisk med modellen för vår räta linje, $\log b = x_1 + \lambda t$.

För att göra analogin ännu tydligare sätter vi $x_2 = \lambda$ och får:

$$\min_x \left\| \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \log b_1 \\ \log b_2 \\ \vdots \\ \log b_m \end{bmatrix} \right\|$$

Efter att x är beräknad sätter vi så $c = e^{x_1}$ och $\lambda = x_2$.

När vi gör transformationer på detta sätt ändrar vi (ibland) på normen. Logaritmering, till exempel, har en utjämnande verkan, och minskar de stora residualernas inflytande. Detta kan jämföras med att minimera i en annan norm. Vi ställer en annan fråga, men den kan ju vara lika relevant.

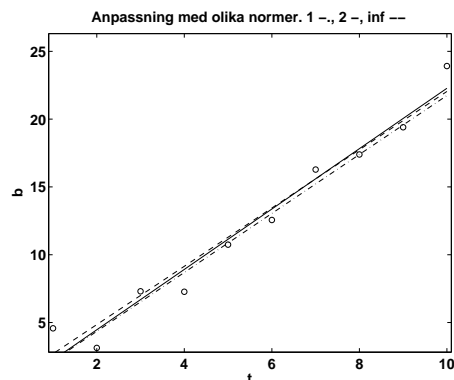
Ibland fäster vi olika stor vikt vid de olika residualerna. Mätapparaturen kanske mäter olika noga i olika mätområden. Det är då rimligt att ett osäkert värde får mindre inflytande än ett säkert. Vi kan åstadkomma detta med en viktad norm, t.ex.

$$\min_x \|V(Ax - b)\|, \quad V = \text{diag}(v_1, v_2, \dots, v_m)$$

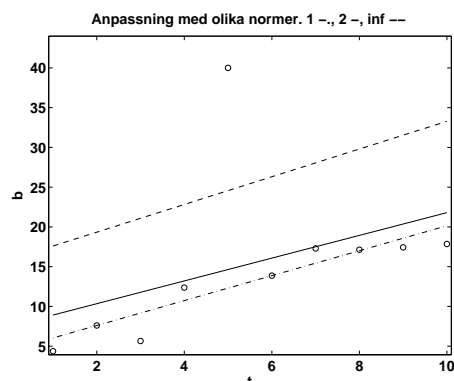
Residual, r_k , multipliceras alltså med vikten v_k .

61

Låt oss nu återvända till den räta linjen, $b = x_1 + x_2 t$. Följande bild visar anpassning i våra tre normer.



Accentueras när man har utliggare (outliers).



62

Vi studerar nu det linjära minstakvadratproblemet:

$$\min_x \|Ax - b\|_2$$

Det är enkelt att beskriva den optimala lösningen till detta problem. Vi ser på specialfallet när A har två kolonner, a_1 respektive a_2 , men det kan enkelt generaliseras till ett godtyckligt fall.

För en godtycklig $x \in \mathbb{R}^2$ gäller att $Ax = a_1 x_1 + a_2 x_2$ är en linjärkombination av A s kolonner. När x varierar över alla vektorer med två element så kommer mängden $a_1 x_1 + a_2 x_2$ att bilda ett plan, A s bildrum, $\mathcal{R}(A)$ (ty $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$).

Om b tillhör detta plan så existerar (minst) ett x så att $Ax = b$ med likhet. Residualvektorn $r = Ax - b$ blir då noll. T.ex.

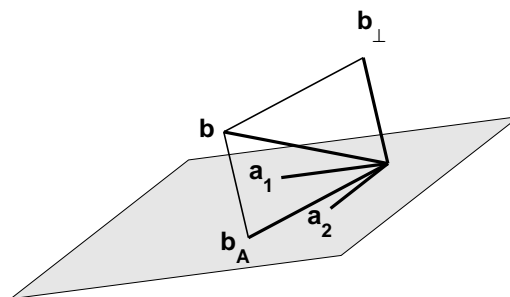
$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

I exemplet är $x = [1, 1]^T$. Normalt bildar dock b en vinkel mot planet, tag till exempel $b = [2, 1, 2]^T$. Vektorn b kan då inte skrivas som en linjärkombination av A s kolonner, men vi vill minimera avvikelserna, längden av residualvektorn $Ax - b$.

Dela upp b i två komponenter, b_A som ligger i planet och b_\perp som är ortogonal mot planet. Oavsett hur vi väljer x så kan vi inte nollställa någon del av b_\perp , eftersom b_\perp är ortogonal mot alla linjärkombinationer, Ax . Däremot kan vi nollställa b_A , eftersom b_A ligger i planet och den därmed är en linjärkombination av A s kolonner, dvs. det existerar (minst) ett x så att $b_A = Ax$. Detta x är det x vi söker.

Residualvektorn blir $r = Ax - b = Ax - b_A - b_\perp = -b_\perp$.

63



Här följer samma resonemang med normer:

Pythagoras sats: om y och z är ortogonala vektorer gäller:

$$\|y + z\|_2^2 = \|y\|_2^2 + \|z\|_2^2$$

ty

$$\|y + z\|_2^2 = (y + z)^T (y + z) = y^T y + \underbrace{y^T z}_0 + \underbrace{z^T y}_0 + z^T z = \|y\|_2^2 + \|z\|_2^2$$

Det x som löser $Ax = b_A$ är optimalt. Ty om så inte vore fallet existerar $z \neq 0$ så att $x + z$ ger ett mindre värde på normen. Vi testar:

$$\|A(x + z) - b\|_2^2 = \|A(x + z) - (b_A + b_\perp)\|_2^2 = \|\underbrace{Ax - b_A}_0 + Az - b_\perp\|_2^2 = \|Az\|_2^2 + \|b_\perp\|_2^2 \geq \|b_\perp\|_2^2$$

Med minimum då $z = 0$ (om A har linjärt oberoende kolonner).

64