180521 Matriisit

Timo Erkama
LaTeX-käännös Tommi Sallinen
Joensuun yliopisto
3. joulukuuta 2009

Sisältö

Osa I

Matriisiteoriaa

1 Peruskäsitteitä

1.1 Vektorit

Skalaarit ovat yleensä reaalilukuja $(\alpha, \beta, \gamma, \ldots)$.
Vektori on euklidisen avaruuden alkio $(\mathbf{x}, \mathbf{y}, \mathbf{a}, \mathbf{b}, \ldots)$.

n-ulotteisen avaruuden $\mathbb{R}^n$ alkio on n-vektori:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \tag{1.1}$$

Vektorin $\mathbf{x}$ *komponentit* $x_1, \ldots, x_n$ ovat skalaareja; niiden lukumäärä on vektorin $\mathbf{y}$ *dimensio*. Vektorit $\mathbf{x}$ ja $\mathbf{y}$ ovat yhtäsuuret, jos niillä on sama dimensio ja jos $x_i = y_i$ kaikille i. Vektorin (1.1) *transpoosi* $\mathbf{x}^T$ on *vaakavektori* ($1 \times n$-matriisi):

$$\mathbf{x}^T = (x_1, x_2, \ldots, x_n).$$

Vaakavektoreita voidaan merkitä myös esimerkiksi $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$.

Vektori voidaan kertoa skalaarilla γ:

$$\gamma\mathbf{x} = \begin{pmatrix} \gamma x_1 \\ \gamma x_2 \\ \vdots \\ \gamma x_n \end{pmatrix}$$

Olkoot $\mathbf{x}_1, \ldots, \mathbf{x}_k$ vektoreita, joilla on sama dimensio. Vektoreiden $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *lineaarikombinaatio* on muotoa

$$\gamma_1\mathbf{x}_1 + \ldots + \gamma_k\mathbf{x}_k, \tag{1.2}$$

missä $\gamma_1, \ldots, \gamma_k$ ovat lineaarikombinaation (1.2) *kertoimet*.

Lineaarikombinaatio (1.2) on *triviaali*, jos $\gamma_i = 0$ kaikille i, muussa tapauksessa lineaarikombinaatio on *epätriviaali*. Jos vektoreilla $\mathbf{x}$ ja $\mathbf{y}$ on sama dimensio n, niiden sisätulo $\mathbf{x}^T\mathbf{y}$ on skalaari

$$\mathbf{x}^T\mathbf{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n = \sum_{i=1}^{n} x_iy_i$$

Sisätulo on

- vaihdannainen: $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}$
- distributiivinen: $\mathbf{a}^T(\mathbf{b} + \mathbf{c}) = \mathbf{a}^T\mathbf{b} + \mathbf{a}^T\mathbf{c}$

n-vektorin $\mathbf{x}$ euklidinen *normi* on

$$\sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2 + \ldots + x_n^2}$$

Jos vektorin $\mathbf{x}$ euklidinen normi on 1, niin $\mathbf{x}$ on *normalisoitu*. Jokaisesta vektorista $\mathbf{x} \neq \mathbf{0}$ saadaan samansuuntainen normalisoitu vektori $\mathbf{u}$ jakamalla $\mathbf{x}$ omalla normillaan:

$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{x}^T\mathbf{x}}}\mathbf{x}$$

n-vektorit ovat *ortogonaaliset*, jos $\mathbf{x}^T\mathbf{y} = 0$.

1.2 Matriisit

Olkoon A $m \times n$-matriisi (m vaakarivia, n pystyriviä). Merkitään A_{ij} tai a_{ij} alkiota, joka sijaitsee vaakarivillä i ja pystyrivillä j.

Matriisin A alkio a_{ij} on

- diagonaalinen, jos $i = j$
- epädiagonaalinen, jos $i \neq j$
- subdiagonaalinen, jos $i > j$
- superdiagonaalinen, jos $i < j$

Matriisit A ja B ovat *yhtäsuuret*, jos niillä on sama tyyppi (esim $n \times m$) ja jos $a_{ij} = b_{ij}$ kaikilla i, j.

Nollamatriisin jokainen alkio on 0.

$m \times n$-matriisin A *transpoosi* A^T on $n \times m$-matriisi, jolle pätee:

$$\left(A^T\right)_{ij} = A_{ji} \;\; (1 \leq i \leq m, \;\; a \leq j \leq n)$$

Matriisi A on *symmetrinen*, jos $A^T = A$, siis jos $a_{ij} = a_{ji}$ kaikille i, j. Symmetrinen matriisi on aina neliömatriisi.

Merkitään matriisin A pystyvektoria j vektorilla $\mathbf{a}_j$.

Esimerkiksi

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 4 & 6 \end{pmatrix}, \text{ niin } \mathbf{a}_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \mathbf{a}_3 = \begin{pmatrix} -1 \\ 6 \end{pmatrix}$$

Yleisesti, jos A on $m \times n$-matriisi, niin $\mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}$ ja merkitään

$$A = (\mathbf{a}_1 \ \mathbf{a}_2 \ \ldots \ \mathbf{a}_n) \tag{1.3}$$

Joskus matriisin A vaakavektoria i merkitään $\mathbf{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{in})$, jolloin

$$A = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_m' \end{pmatrix} \tag{1.4}$$

Kerrottaessa $m \times n$-matriisi A skalaarilla γ, saadaan $m \times n$-matriisi γA, jolle pätee $(\gamma A)_{ij} = \gamma A_{ij}$.

Matriisin A ja n-vektorin $\mathbf{x}$ tulo on m-vektori $A\mathbf{x}$ siten, että

$$A\mathbf{x} = \begin{pmatrix} \mathbf{a}_1'\mathbf{x} \\ \mathbf{a}_2'\mathbf{x} \\ \vdots \\ \mathbf{a}_n'\mathbf{x} \end{pmatrix}$$

Siis vektorin $A\mathbf{x}$ i:s komponentti on sisätulo $\mathbf{a}_i'\mathbf{x} = \sum_{k=i}^{n} a_{ik}x_k$.
Toisin sanoen

$$A\mathbf{x} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \ldots \ \mathbf{a}_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_n x_n.$$

on matriisin A sarakkeiden lineaarikombinaatio, jonka kertoimet ovat vektorin $\mathbf{x}$ komponentit.

Kuvaus $\mathbf{x} \mapsto A\mathbf{x}$ on lineaarinen, sillä

$$A(\alpha\mathbf{x} + \beta\mathbf{y}) = A(\alpha\mathbf{x}) + A(\beta\mathbf{y}) = \alpha A\mathbf{x} + \beta A\mathbf{y} \tag{1.5}$$

aina, kun α ja β ovat skalaareja ja $\mathbf{x}$ ja $\mathbf{y}$ n-vektoreita.

Kahden matriisin A ja B tulo AB on määritelty, jos A on tyyppiä $m \times l$ ja B tyyppiä $l \times n$.

Jos, kuten kaavassa (1.3), kirjoitetaan $B = (\mathbf{b}_1\ \mathbf{b}_2\ \ldots\ \mathbf{b}_m)$, niin

$$AB = A\,(\mathbf{b}_1\ \mathbf{b}_2\ \ldots\ \mathbf{b}_n) = (A\mathbf{b}_1\ A\mathbf{b}_2\ \ldots A\mathbf{b}_n)$$

Jokainen matriisin AB sarake on siis matriisin A sarakkeiden lineaarikombinaatio ja matriisin AB j:s sarake riippuu matriisin A ohella vain matriisin B j:nnestä sarakkeesta. Näiden pystyvektorien asemasta voidaan tulon AB karakterisoinnissa käyttää myös matriisin A vaakavektoreita:

$$AB = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_n' \end{pmatrix} B = \begin{pmatrix} \mathbf{a}_1'B \\ \mathbf{a}_2'B \\ \vdots \\ \mathbf{a}_n'B \end{pmatrix}; \quad \mathbf{a}_1'B = (\mathbf{a}_1'\mathbf{b}_1\ \ldots\ \mathbf{a}_1'\mathbf{b}_n)$$

Nähdään, että matriisin AB i:s vaakarivi on matriisin B vaakarivien lineaarikombinaatio, joka riippuu matriisin B ohella vain matriisin A vaakarivistä i.

Tulon AB ik:s alkio on $\mathbf{a}_i'\mathbf{b}_k$.

Matriisin kertolasku on assosiatiivinen ja distributiivinen:

$$A\,(BC) = (AB)\,C \quad \text{ja} \quad A\,(B + C) = AB + AC$$

aina, kun yhtälöiden vasemmalla puolella olevat lausekkeet on määritelty.

Matriisitulo ei kuitenkaan ole vaihdannainen: yleensä $AB \neq BA$.

Tulon transpoosille on helppo todistaa kaava:

$$(AB)^T = B^T A^T,$$

jonka avulla voidaan myös nähdä, että kahden symmetrisen matriisin tulo ei yleensä ole symmetrinen.

Neliömatriisin A k:s potenssi määritellään induktiivisesti seuraavasti:

$$A^k = A \cdot A^{k-1}$$

Sovitaan myös: $A^0 = I$, kun $A \neq 0$

Usein joudutaan tarkastelemaan matriiseja, joiden alkiot ovat tarkasteltavan matriisin *alimatriiseja* tai lohkoja.

Esimerkki:

$$A = \begin{pmatrix} 1 & 0 & 2 & 3 & 4 \\ 0 & 1 & 1 & 0 & 5 \\ 2 & 1 & 2 & 0 & 0 \\ 3 & 0 & 0 & 3 & 0 \\ 4 & 5 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} A_1 & A_2^T \\ A_2 & A_3 \end{pmatrix}$$

missä $A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $A_2 = \begin{pmatrix} 2 & 1 \\ 3 & 0 \\ 4 & 5 \end{pmatrix}$ ja $A_3 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Esitystä lohkomuodossa käytetään mm. silloin, kun lohkoilla on joitakin erityisominaisuuksia. Mikäli lohkomatriiseilla on sopiva tyyppi, voidaan niillä suorittaa yhteen- ja kertolaskutoimituksia ikäänkuin lohkot itsessään olisivat skalaareja.

Esimerkki:

$$\begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} + \begin{pmatrix} B_1 & B_1 \\ B_3 & B_4 \end{pmatrix} = \begin{pmatrix} A_1 + B_1 & A_2 + B_2 \\ A_3 + B_3 & A_4 + B_4 \end{pmatrix}$$

ja

$$\begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix} = \begin{pmatrix} A_1B_1 + A_2B_3 & A_1B_2 + A_2B_4 \\ A_3B_1 + A_4B_3 & A_3B_2 + A_4B_4 \end{pmatrix}$$

edellyttäen, että kussakin lohkossa olevat lausekkeet ovat määriteltyjä.

1.3 Tärkeitä nimityksiä

Kertalukua n oleva *yksikkömatriisi* I_n (merkitään myös I).

Esimerkiksi $I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Jos A on $n \times n$-matriisi, niin $AI_n = A$ ja $I_nA = A$.

Yksikkömatriisin I_n j:s sarake on avaruuden $\mathbb{R}^n$ yksikkövektori, jota merkitään $\mathbf{e}_j$.

Huomautus: $A\mathbf{e}_j = \mathbf{a}_j; \quad \mathbf{e}_j^T A = \mathbf{a}_i'; \quad a_{ij} = \mathbf{e}_i^T A \mathbf{e}_j$

Matriisi on *diagonaalinen*, jos kaikki sen epädiagonaaliset alkiot ovat nollia. Tekstissä nollat jätetään yleensä merkitsemättä.

Esimerkiksi $\begin{pmatrix} -1 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{pmatrix} = \begin{pmatrix} -1 & & \\ & 6 & \\ & & 8 \end{pmatrix} = \mathrm{diag}(-1, 6, 8)$

Yleisemmin diag(...) tarkoittaa diagonaalimatriisia, jonka lävistäjän alkiot on ilmoitettu sulkujen sisäpuolella; tällöin kysymys on yleensä neliömatriisista. Jos esimerkiksi A on $n \times n$-matriisi, niin $\mathrm{diag}(a_{ii})$ tarkoittaa matriisin A diagonaalialkiosta muodostettua diagonaalimatriisia.

Jos $D = \mathrm{diag}(d_i)$, niin

$$DA = \begin{pmatrix} d_1 \mathbf{a}_1' \\ \vdots \\ d_i \mathbf{a}_i' \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1' A \\ \vdots \\ \mathbf{d}_i' A \\ \vdots \end{pmatrix}$$

kun taas

$$AD = (d_1 \mathbf{a}_1 \ \ldots \ d_j \mathbf{a}_j) = (A\mathbf{d}_1 \ \ldots \ A\mathbf{d}_j)$$

Matriisi U on *yläkolmiomatriisi*, jos sen kaikki $u_{ij} = 0$ kaikilla $i > j$ ts. jos subdiagonaaliset alkiot ovat nollia. Vastaavasti, matriisi L on *alakolmiomatriisi*, jos kaikki $l_{ij} = 0$ kaikilla $i < j$ ts. jos superdiagonaaliset alkiot ovat nollia.

Joukko $\{\mathbf{p}_1, \ldots \mathbf{p}_k\}$ nollasta eriäviä vektoreita on *ortogonaalinen*, jos

$$\mathbf{p}_i^T \mathbf{p}_j = 0, \quad \text{kun } i \neq j \tag{1.6}$$

Jos lisäksi $\mathbf{p}_i^T \mathbf{p}_i = 1$ kaikille $i = 1, \ldots, k$, niin vektorijoukko on *ortonormaali*.

Esimerkiksi vektoreiden

$$\mathbf{p}_1 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \text{ ja } \mathbf{p}_2 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \tag{1.7}$$

muodostama joukko on ortonormaali.

$m \times n$-matriisin P sarakkeiden joukko on ortonormaali, jos ja vain jos

$$P^T P = I_n$$

$$\begin{aligned} \mathbf{p}_j &= P\mathbf{e}_j \\ \mathbf{p}_i^T &= \mathbf{e}_i^T P^T \\ \mathbf{p}_i^T \mathbf{p}_j &= \mathbf{e}_i^T \underbrace{P^T P}_{I_n} \mathbf{e}_j = \mathbf{e}_i^T \mathbf{e}_j = \begin{cases} 0, & \text{kun } i \neq j \\ 1, & \text{kun } i = j \end{cases} \end{aligned}$$

Tällöin matriisin P vaakarivien joukko ei välttämättä ole ortonormaali, paitsi tapauksessa $m = n$.

Esimerkiksi vektoreista (1.7) muodostettu 4×2-matriisin $(\mathbf{p}_1\ \mathbf{p}_2)$ vaakarivit eivät muodosta ortonormaalia joukkoa.

Määritelmä 1. Neliömatriisi Q on *ortogonaalinen*, jos sen sarakkeiden joukko on ortonormaali, ts. jos

$$Q^T Q = I \tag{1.8}$$

Tällöin myös $QQ^T = I$, joten myös vaakarivien joukko on ortonormaali.

Esimerkki

$$Q = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

Vektoreiden $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$ ja $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ *ulkoinen tulo* on $m \times n$-matriisi

$$\mathbf{x}\mathbf{y}^T = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_2 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \dots & x_m y_n \end{pmatrix}$$

Alkeismatriisi on muotoa $E = I - \alpha \mathbf{u}\mathbf{v}^T$, missä I on $n \times n$-yksikkömatriisi, α skalaari ja $\mathbf{u}$ sekä $\mathbf{v}$ n-vektoreita. Jos $\mathbf{x}$ on mielivaltainen n-vektori, niin

$$E\mathbf{x} = \left(I - \alpha \mathbf{u}\mathbf{v}^T\right)\mathbf{x} = \mathbf{x} - \alpha \mathbf{u}\left(\mathbf{v}^T \mathbf{x}\right) = \mathbf{x} - \xi \mathbf{u},$$

missä $\xi = \alpha\left(\mathbf{v}^T\mathbf{x}\right)$.

Näin ollen $E\mathbf{x}$ saadaan $\mathbf{x}$:stä vähentämällä siitä eräs $\mathbf{u}$:n suuntainen vektori.

2 Vektoriavaruudet

2.1 Lineaarinen riippuvuus ja riippumattomuus

Joukko $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ jonkin euklidisen avaruuden vektoreita on *lineaarisesti riippuva*, jos jokin sen eri vektoreista muodostettu epätriviaali lineaarikombinaatio on yhtäsuuri kuin nollavektori. Muussa tapauksessa vektorit ovat *lineaarisesti riippumattomat*.

Jokaisessa avaruuden $\mathbb{R}^n$ lineaarisesti riippumattomassa osajoukossa on korkeintaan n alkiota, ja mikään niistä ei ole nollavektori.

Matriisin A pystyvektorit ovat lineaarisesti riippuvia, jos ja vain jos

$$A\mathbf{z} = \mathbf{0} \text{ jollekin vektorille } \mathbf{z} \neq \mathbf{0}. \tag{2.1}$$

Mikäli $A\mathbf{z} = \mathbf{0}$ toteutuu vain jos

$$\mathbf{z} = \mathbf{0}, \tag{2.2}$$

niin matriisin A sarakkeet ovat lineaarisesti riippumattomat.

Tarkastellaan *lineaarista yhtälöryhmää*

$$A\mathbf{x} = \mathbf{b}, \tag{2.3}$$

missä A on matriisi ja $\mathbf{x}$ ja $\mathbf{b}$ ovat vektoreita.

Matriisin A ja vektorin $\mathbf{b}$ ollessa annettuja, sanotaan, että (2.3) on *ratkeava*, jos on olemassa vektori $\mathbf{x}$ siten, että (2.3) pätee; muussa tapauksessa sanotaan, että (2.3) on *ratkeamaton*.

Selvästi (2.3) on ratkeava, jos ja vain jos vektori $\mathbf{b}$ matriisin A sarakkeiden lineaarikombinaatio.

Jos matriisin A sarakkeet ovat lineaarisesti riippuvia ja (2.3) on ratkeava,

niin systeemillä (2.3) on äärettömän monta ratkaisua. Kaavoista (2.1) ja (2.3) seuraa nimittäin, että (vrt. (1.5))

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

jokaiselle skalaarille α. Siten myös $\mathbf{x} + \alpha\mathbf{z}$ toteuttaa yhtälön (2.3). Vektorin $\mathbf{b}$ esitys matriisin A sarakkeiden lineaarikombinaationa ei tällöin ole yksikäsitteinen. Tapauksessa $\mathbf{b} = \mathbf{0}$ nähdään, erityisesti, että homogeenisellä yhtälöryhmällä $A\mathbf{x} = \mathbf{0}$ on ei-triviaaleja ratkaisuja, mikäli matriisin A sarakkeet ovat lineaarisesti riippuvia. Jos sen sijaan matriisin A sarakkeet ovat lineaarisesti riippumattomat ja (2.3) on ratkeava, niin ratkaisu on yksikäsitteinen. Tämä nähdään tarkastelemalla kahta ratkaisua $\mathbf{x}$ ja $\bar{\mathbf{x}}$, joille siis

$$A\mathbf{x} = \mathbf{b} \text{ ja } A\bar{\mathbf{x}} = \mathbf{b}.$$

Vähentämällä yhtälöt puolittain, saadaan (vrt. (1.5))

$$A\mathbf{x} - A\bar{\mathbf{x}} = A(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{b} - \mathbf{b} = \mathbf{0},$$

josta yhtälön (2.2) perusteella seuraa $\mathbf{x} - \bar{\mathbf{x}} = 0$ eli $\mathbf{x} = \bar{\mathbf{x}}$.

Pystyvektorien lineaarinen riippumattomuus takaa siis ratkaisujen yksikäsitteisyyden, mutta ei sen olemassaoloa (paitsi jos A on neliömatriisi, vrt. luku 2.3).

Esimerkki:

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 2 \\ 3 \\ 2 \end{pmatrix} \tag{2.4}$$

ei ole ratkeava.

2.2 Maaliavaruus ja nolla-avaruus

Olkoon S joukko m-vektoreita. Sanotaan, että S on avaruuden $\mathbb{R}^m$ aliavaruus, jos ehdosta $\mathbf{x}, \mathbf{y} \in S$ seuraa $\alpha\mathbf{x}+\beta\mathbf{y} \in S$ aina kun α ja β ovat skalaareja. Aliavaruus sisältää aina nollavektorin (vrt. jos $\alpha = \beta = 0$).

Esimerkiksi avaruuden $\mathbb{R}^2$ aliavaruuksia ovat kaikki origon kautta kulkevat suorat ja avaruuden $\mathbb{R}^3$ aliavaruuksia ovat kaikki origon kautta kulkevat suorat ja tasot.

Sanotaan, että joukon S osajoukko $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ virittää joukon S, jos jokainen joukon S alkio on vektoreiden $\mathbf{a}_1, \ldots, \mathbf{a}_k$ lineaarikombinaatio.

Nollavektorin muodostama yksiö $\{\mathbf{0}\}$ on avaruuden $\mathbb{R}^m$ *triviaali* aliavaruus, jonka dimensio on 0. Epätriviaalin aliavaruuden *dimensio* = kantavektoreiden lukumäärä. Joukon S *kanta* on mikä hyvänsä joukon S virittävä osajoukko, jonka alkiot ovat lineaarisesti riippumattomat.

$m \times n$-matriisin A *maaliavaruus* tai *maali* (range) on avaruuden $\mathbb{R}^m$ osajoukko

$$R(A) = \{A\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n\}$$

$R(A)$ on avaruuden $\mathbb{R}^m$ aliavaruus, jos $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ ja α_1, α_2 ovat skalaareja, niin kuvauksen $\mathbf{x} \mapsto A\mathbf{x}$ lineaarisuuden nojalla

$$\alpha_1 (A\mathbf{x}_1) + \alpha_2 (A\mathbf{x}_2) = A (\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) \in R(A)$$

$R(A)$ sisältää kaikki matriisin A sarakkeet ($A\mathbf{e}_j = \mathbf{a}_j$) ja $\dim R(A) \leq m$.

Esimerkiksi kaavasta (2.4) $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}$ ja maaliavaruus $R(A)$ käsittää kaikki ne vektorit, jotka ovat muotoa

$$A \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} \gamma_1 + \gamma_2 \\ \gamma_1 \\ \gamma_1 - \gamma_2 \end{pmatrix}, \tag{2.5}$$

missä γ_1 ja γ_2 ovat mielivaltaisia skalaareja. Koska (2.4) ei ole ratkeava, niin sen oikealla puolella oleva vektori $\begin{pmatrix} 2 \\ 3 \\ 2 \end{pmatrix}$ ei kuulu avaruuteen $R(A)$.

Maaliavaruuden $R(A)$ dimensio on matriisin A *aste*.

Selvästi $\dim R(A) \leq n$, koska matriisin A sarakkeet $\mathbf{a}_1, \ldots, \mathbf{a}_n$ virittävät maaliavaruuden $R(A)$, jonka mielivaltaisessa kannassa on siten korkeintaan n alkiota.

Matriisin A *sarakeaste* on matriisin A sarakkeiden virittämän avaruuden $\mathbb{R}^m$ aliavaruuden dimensio, ja matriisin A *riviaste* on matriisin A vaakavektorien (transpoosien) virittämän avaruuden $\mathbb{R}^n$ aliavaruuden dimensio.

Jos matriisin A sarakkeet ovat lineaarisesti riippumattomat, ne muodostavat maaliavaruuden $R(A)$ kannan, jolloin $r(A) = \dim R(A) = n$; tällöin sanotaan, että A on *täyttä sarakeastetta*. Vastaavasti matriisi A on *täyttä riviastetta*, jos $\dim R(A^T) = m$.

Matriisi on *täysasteinen*, jos se on joko täyttä rivi- tai sarakeastetta. Jos matriisi A ei ole täysasteinen, siis jos $r(A) < \min(n, m)$, sanotaan, että matriisi A on *vajaa-asteinen*.

Esimerkki

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix},\ B = \begin{pmatrix} 1 & 2 \\ 0 & 0 \\ 1 & 2 \end{pmatrix},\ C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},\ D = \begin{pmatrix} 1 & 5 & 3 \\ -1 & -5 & -3 \end{pmatrix}$$

Näiden matriisien asteet ovat ≤ 2. Koska matriisin A sarakkeet ovat lineaarisesti riippumattomat, niin $r(A) = 2$ ja matriisi A on täyttä sarakeastetta. Sen sijaan matriisi B on vajaa-asteinen, sillä sen kumpikin sarake saadaan toisesta skalaarilla kertomalla; näin ollen $r(B) = 1$. Edelleen matriisi C on täyttä riviastetta, mutta sarakkeet ovat lineaarisesti riippuvia, ja $r(D) = 1$.

Lause 2. *Matriiseilla A ja A^T on sama aste.*

Todistus. Osoitetaan ensin, että $\dim R(A) \leq \dim R(A^T)$. Olkoon $r = \dim R(A^T)$, $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ on avaruuden $R(A^T)$ kanta ja $V = (\mathbf{v}_1 \ldots \mathbf{v}_r)$ on vastaava $n \times r$-matriisi. Lisäksi $C = (\mathbf{c}_1 \ldots \mathbf{c}_m)$ $r \times m$-matriisi.
Koska $R(A^T)$ sisältää kaikki matriisin A^T sarakkeet, voidaan matriisin A^T sarake i esittää muodossa

$$\left(A^T\right)_i = V\mathbf{c}_i$$

$$A^T = VC \Leftrightarrow A = C^T V^T$$

$A\mathbf{x} = C^T V^T \mathbf{x}$ matriisin C^T sarakkeiden lineaarikombinaatio.
Jokainen $A\mathbf{x} \in R(A)$ kuuluu siis matriisin C^T sarakkeiden virittämään aliavaruuteen $R(C^T)$, jonka dimensio on enintään matriisin C^T sarakkeiden lukumäärä r. Siis $R(A) \subset R(C^T)$, joten

$$\dim R(A) \leq \dim R(C^T) \leq r \leq \dim R(A^T)$$

Vastaavasti (korvaamalla matriisi A matriisilla A^T) nähdään, että

$$\dim R(A^T) \leq \dim R((A^T)^T) = \dim R(A)$$

Siis

$$\dim R(A) = \dim(A^T)$$

□

Matriisin A sarakeaste = matriisin A riviaste = matriisin A aste.

$m \times n$-matriisin A *nolla-avaruus* $\mathcal{N}(A)$ on lineaarikuvauksen $\mathbf{x} \mapsto A\mathbf{x}$ ydin $\{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{0}\}$. Selvästi $\mathcal{N}(A)$ on avaruuden $\mathbb{R}^n$ aliavaruus. $\mathcal{N}(A^T) = \left\{\mathbf{z} \in \mathbb{R}^m \mid A^T\mathbf{z} = \mathbf{0}\right\}$.

Väite 3. *Avaruus $\mathcal{N}(A^T)$ koostuu niistä avaruuden $\mathbb{R}^m$ vektoreista, jotka ovat kohtisuorassa jokaista avaruuden $R(A)$ vektoria vastaan.*

Todistus. Olkoon $\mathbf{z} \in \mathcal{N}(A^T)$. Tällöin $A^T\mathbf{z} = \mathbf{0}$. Tästä seuraa

$$(A\mathbf{x})^T \mathbf{z} = \mathbf{x}^T \left(A^T\mathbf{z}\right) = \mathbf{0} \ \forall \mathbf{x} \in \mathbb{R}^n.$$

Kääntäen: oletetaan, että $\mathbf{z} \perp A\mathbf{x}\ \forall \mathbf{x} \in \mathbb{R}^n$. Valitaan $\mathbf{x} = A^T\mathbf{z}$, jolloin

$$\mathbf{0} = \mathbf{z}^T A\mathbf{x} = \mathbf{z}^T AA^T\mathbf{z} = \left\|A^T\mathbf{z}\right\|^2.$$

Tästä seuraa $A^T\mathbf{z} = \mathbf{0} \Rightarrow \mathbf{z} \in \mathcal{N}(A^T)$. □

Aliavaruudet $R(A)$ ja $\mathcal{N}(A^T)$ ovat toistensa *ortogonaalisia komplementteja*: jokainen avaruuden $\mathbb{R}^m$ vektori voidaan esittää kahden ortogonaalisen vektorin summana, joista toinen kuuluu maaliavaruuteen $R(A)$ ja toinen nolla-avaruuteen $\mathcal{N}(A^T)$. Edelleen

$$R(A) \cap \mathcal{N}(A^T) = \{\mathbf{0}\} \text{ ja } \dim R(A) + \dim \mathcal{N}(A^T) = m.$$

Jos matriisin A sarakkeet ovat lineaarisesti riippumattomat, jolloin $\dim R(A) = n \leq m$, niin $\dim \mathcal{N}(A^T) = m - n$. Korvaamalla matriisi A matriisilla A^T nähdään vastaavasti, että jokainen n-vektori voidaan esittää avaruuksien $R(A^T)$ ja $\mathcal{N}(A)$ vektorien summana, ja

$$\dim R(A^T) + \dim \mathcal{N}(A) = n.$$

Esimerkiksi kaavan (2.4) matriisille A, $A^T = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$, pätee

$$A^T\mathbf{z} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} z_1 + z_2 + z_3 \\ z_1 - z_3 \end{pmatrix},$$

aina, kun $\mathbf{z} \in \mathbb{R}^3$. Näin ollen $\mathbf{z} \in \mathcal{N}(A^T)$ jos ja vain jos

$$\begin{pmatrix} z_1 + z_2 + z_3 \\ z_1 - z_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Tämän yhtälön ratkaisu on

$$\mathbf{z} = \begin{pmatrix} \gamma_3 \\ -2\gamma_3 \\ \gamma_3 \end{pmatrix}, \tag{2.6}$$

missä γ_3 on mielivaltainen skalaari. Näin ollen $\dim \mathcal{N}(A^T) = 1$.

Koska $R(A) \cap \mathcal{N}(A^T) = \{\mathbf{0}\}$, jokaiselle m-vektorille $\mathbf{b}$ on yksikäsitteinen esitys muodossa

$$\mathbf{b} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}}, \text{ missä } \mathbf{b}_R \in R(A) \text{ ja } \mathbf{b}_{\mathcal{N}} \in \mathcal{N}(A^T) \tag{2.7}$$

Perustelua:

$$\begin{aligned} \mathbf{b} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}} &= \mathbf{b}'_R + \mathbf{b}'_{\mathcal{N}} \\ \mathbf{b}_R - \mathbf{b}'_R &= \mathbf{b}'_{\mathcal{N}} - \mathbf{b}_{\mathcal{N}} \end{aligned}$$

Koska $\mathbf{b}_R$ ja $\mathbf{b}_{\mathcal{N}}$ ovat ortogonaaliset ja $\mathbf{b}_R^T \mathbf{b}_{\mathcal{N}} = 0$, niin

$$\mathbf{b}^T \mathbf{b} = \mathbf{b}_R^T \mathbf{b}^T + \mathbf{b}_{\mathcal{N}}^T \mathbf{b}^T.$$

Korvaamalla matriisi A matriisillla A^T nähdään vastaavasti, että jokaisen n-vektorin $\mathbf{x}$ esitys

$$\mathbf{x} = \mathbf{x}_R + \mathbf{x}_{\mathcal{N}}, \text{ missä } \mathbf{x}_R \in R(A^T) \text{ ja } \mathbf{x}_{\mathcal{N}} \in \mathcal{N}(A),$$

on yksikäsitteinen.

Olkoon

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -2 \end{pmatrix} \text{ ja } \mathbf{b} = \begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix}$$

Vektorin $\mathbf{b}$ esitys muodossa (2.7) saadaan kaavojen (2.5) ja (2.6) avulla:

$$\mathbf{b} = \begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}} = \begin{pmatrix} \gamma_1 + \gamma_2 \\ \gamma_1 \\ \gamma_1 - \gamma_2 \end{pmatrix} + \begin{pmatrix} \gamma_3 \\ -2\gamma_3 \\ \gamma_3 \end{pmatrix}$$

Tämän yhtälöryhmän

$$\begin{cases} \gamma_1 + \gamma_2 + \gamma_3 = 2 \\ \gamma_1 \qquad - 2\gamma_3 = 0 \\ \gamma_1 - \gamma_2 + \gamma_3 = 4 \end{cases}$$

ratkaisu on $(\gamma_1, \gamma_2, \gamma_3) = (2, -1, 1)$, joten

$$\mathbf{b}_R = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ ja } \mathbf{b}_{\mathcal{N}} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

Edelleen $\mathbf{b}_R^T \mathbf{b}_{\mathcal{N}} = 1 \cdot 1 + 2 \cdot (-2) + 3 \cdot 1 = 0$ ja $\mathbf{b}^T \mathbf{b} = 20 = \underbrace{\mathbf{b}_R^T \mathbf{b}_R}_{14} + \underbrace{\mathbf{b}_{\mathcal{N}}^T \mathbf{b}_{\mathcal{N}}}_{6}$

2.3 Säännölliset ja singulaariset matriisit

Neliömatriisi on *säännöllinen*, jos sen sarakkeet ovat lineaarisesti riippumattomat. Jos sarakkeet eivät ole lineaarisesti riippumattomat, niin sanotaan, että neliömatriisi on *singulaarinen*. Jos A on säännöllinen $n \times n$-matriisi, niin $R(A) = R(A^T) = \mathbb{R}^n$ (vrt. luku 2.2, Lause 2). Tällöin

$$\mathcal{N}(A^T) = \mathcal{N}(A) = \{\mathbf{0}\}.$$

Luvun 2.1 tarkasteluista seuraa välittömästi, kun A on neliömatriisi:

- A on säännöllinen, ja yhtälöllä $A\mathbf{x} = \mathbf{0}$ on triviaaliratkaisu $\mathbf{x} = \mathbf{0}$.
- Jos A on säännöllinen, niin yhtälöllä $A\mathbf{x} = \mathbf{b}$ on aina täsmälleen yksi ratkaisu.
- A on singulaarinen, jos ja vain jos $A\mathbf{x} = \mathbf{0}$ jollekin vektorille $\mathbf{x} \neq \mathbf{0}$.
- Jos A on singulaarinen ja yhtälö $A\mathbf{x} = \mathbf{b}$ on ratkeava, sillä on äärettömän monta ratkaisua.

Jokaisella säännöllisellä matriisilla A on yksikäsitteisesti määrätty *käänteismatriisi* A^{-1} siten, että

$$A^{-1}A = AA^{-1} = I.$$

Kahden säännöllisen matriisin A ja B tulo on säännöllinen, ja

$$(AB)^{-1} = B^{-1}A^{-1}.$$

Jos matriisi A on säännöllinen, niin matriisi A^T on säännöllinen ja

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T.$$

Jos $D = \operatorname{diag}(d_1, \ldots, d_n)$ on säännöllinen, niin myös D^{-1} on diagonaalimatriisi ja $D^{-1} = \operatorname{diag}\left(d_1^{-1}, \ldots, d_n^{-1}\right)$.

Diagonaalinen neliömatriisi on singulaarinen täsmälleen silloin, kun joku sen diagonaalialkioista on 0. Jos Q on ortogonaalinen neliömatriisi, niin (1.8):n perusteella $Q^{-1} = Q^T$.

Säännöllisen *alkeismatriisin* $E = I - \alpha \mathbf{u}\mathbf{v}^T$ käänteismatriisi on alkeismatriisi; jos $\alpha \mathbf{u}^T \mathbf{v} \neq 1$, niin voidaan näyttää, että

$$E^{-1} = \left(I - \alpha \mathbf{u}\mathbf{v}^T\right)^{-1} = I - \beta \mathbf{u}\mathbf{v}^T, \quad \text{missä } \beta = \frac{\alpha}{\alpha \mathbf{u}^T \mathbf{v} - 1}.$$

Käänteismatriisin käsite on teoreettisesti tärkeä, mutta sen laskemista tulee välttää.

3 Ominaisarvot ja singulaariarvot

3.1 Ominaisarvot ja ominaisvektorit

Kompleksinen n-vektori on avaruuden $\mathbb{C}^n$ alkio. *Kompleksinen matriisi* $A = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ on järjestetty joukko kompleksisia n-vektoreita.

Kompleksisille vektoreille ja matriiseille määritellään yhteen- ja vähennyslaskuoperaatiot kuten reaalisessa tapauksessa.

Esimerkiksi, jos $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ on kompleksinen n-vektori ja $\alpha \in \mathbb{C}$, niin

$$\alpha \mathbf{x} = \begin{pmatrix} \alpha \mathbf{x}_1 \\ \vdots \\ \alpha \mathbf{x}_n \end{pmatrix}$$

Määritelmä 4. Kompleksinen $n \times n$-neliömatriisi on *singulaarinen*, jos $A\mathbf{x} = \mathbf{0}$ jollekin $\mathbf{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$.

Lause 5. *Matriisi A on singulaarinen jos ja vain jos* $\det A = 0$.

Määritelmä 6. Olkoon A $n \times n$ matriisi. *Karakteristinen polynomi* on astetta n oleva muuttujan λ polynomi

$$\det\left(A - \lambda I\right),$$

jonka nollakohtia kutsutaan matriisin A *ominaisarvoiksi.*

Olkoon esimerkiksi

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

jolloin

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3$$

Matriisin A ominaisarvot ovat yhtälön $\lambda^2 - 4\lambda + 3 = 0$ juuret $\lambda_1 = 1$ ja $\lambda_2 = 3$.

Koska polynomilla, jonka aste on n, on korkeintaan n nollakohtaa, niin $n \times n$-matriisilla on korkeintaan n ominaisarvoa.

Esimerkki: Olkoon

$$A = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$$

Tällöin

$$\det(A - \lambda I) = \begin{vmatrix} 3-\lambda & 1 \\ 0 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 = 0$$

kun $3 - \lambda = 0$ eli kun $\lambda = 3$. Siis matriisilla A on vain yksi ominaisarvo $\lambda = 3$.

Jos λ on $m \times n$-matriisin A ominaisarvo, niin matriisi $A - \lambda I$ on singulaarinen. Näin ollen on olemassa reaalinen tai kompleksinen vektori $\mathbf{u} \neq \mathbf{0}$ siten että

$$(A - \lambda I)\,\mathbf{u} = \mathbf{0}$$

Tällaista vektoria $\mathbf{u} \neq \mathbf{0}$ kutsutaan matriisin A ominaisvektoriksi:

$$A\mathbf{u} = \lambda\mathbf{u} \tag{3.1}$$

on *ominaisarvoa* λ *vastaava* matriisin A ominaisvektori.

Edellisessä esimerkissä ominaisarvoa $\lambda = 3$ vastaava ominaisvektori on esimerkiksi $\mathbf{u} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, sillä

$$A\mathbf{u} = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 3\mathbf{u} = \lambda\mathbf{u}$$

Kompleksisen $n \times n$-matriisin A nolla-avaruus $\mathcal{N}(A)$ on matriisin A määräämän lineaarikuvauksen $\mathbf{x} \mapsto A\mathbf{x}$ $\mathbb{C}^n \to \mathbb{C}^n$ ydin ja on siten avaruuden $\mathbb{C}^n$ *kompleksinen aliavaruus*. $\alpha\mathbf{x} \in \mathcal{N}(A)$ aina, kun $\alpha \in \mathbb{C}$ ja $\mathbf{x} \in \mathcal{N}(A)$.

Vektorin $\mathbf{x}$ *konjugaatti* $\bar{\mathbf{x}}$ saadaan konjugoimalla kaikki vektorin $\mathbf{x}$ komponentit:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix}$$

Matriisin A ominaisarvoa λ vastaava ominaisvektori on nolla-avaruuden $\mathcal{N}(A-\lambda I)$ nollasta eroava alkio, jolloin $A\mathbf{u} = \lambda\mathbf{u}$.

Esimerkiksi: $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $\lambda = i$, $\mathbf{u} = \begin{pmatrix} -i \\ 1 \end{pmatrix}$, jolloin

$$A\mathbf{u} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} -i \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ i \end{pmatrix} = i \begin{pmatrix} -i \\ 1 \end{pmatrix}$$

Konjugoimalla puolittain yhtälö $A\mathbf{u} = i\mathbf{u}$ saadaan $A\bar{\mathbf{u}} = -i\bar{\mathbf{u}}$, joten $\bar{\mathbf{u}} = \begin{pmatrix} i \\ 1 \end{pmatrix}$ on ominaisarvoa $\lambda = -i$ vastaava ominaisvektori.

Kahden kompleksisen n-vektorin $\mathbf{u}$ ja $\mathbf{v}$ *sisätulo* määritellään lausekkeella

$$\bar{\mathbf{u}}^T\mathbf{v} = \bar{u}_1 v_1 + \ldots + \bar{u}_n v_n.$$

Tällöin

$$\bar{\mathbf{u}}^T\mathbf{u} = |u_1|^2 + \ldots + |u_n|^2 > 0, \ \mathbf{u} \neq 0 \text{ ja } \|\mathbf{u}\| = \sqrt{\bar{\mathbf{u}}^T\mathbf{u}}.$$

Jos $\mathbf{u}$ on ominaisarvoa λ vastaava matriisin A ominaisvektori ja $\gamma \neq 0$ on skalaari, niin

$$A(\gamma\mathbf{u}) = \gamma A\mathbf{u} = \gamma\lambda\mathbf{u} = \lambda(\gamma\mathbf{u}),$$

joten myös $\gamma\mathbf{u}$ on ominaisvektori. Näin ollen jokaista ominaisarvoa vastaa aina normalisoitu ominaisvektori, jonka pituus on 1.

Huomautus:

$$\left.\begin{array}{l} A\mathbf{u} = \lambda\mathbf{u} \\ A\mathbf{v} = \lambda\mathbf{v} \end{array}\right\} \Rightarrow A(\mathbf{u}+\mathbf{v}) = \lambda(\mathbf{u}+\mathbf{v})$$

Kertomalla (3.1) puolittain säännöllisellä matriisilla S saadaan

$$SA\mathbf{u} = SA\underbrace{S^{-1}S}_{I}\mathbf{u} = \underbrace{\left(SAS^{-1}\right)}_{B}(S\mathbf{u}) = \lambda(S\mathbf{u}) \tag{3.2}$$

Näin ollen $S\mathbf{u}$ on matriisin $B = SAS^{-1}$ ominaisarvoa λ vastaava ominaisvektori, ja nähdään, että jokainen matriisin A ominaisarvo on myös matriisin B ominaisarvo.

Itse asiassa matriiseilla A ja B on samat ominaisarvot.

Edelleen, kun $B = SAS^{-1}$, saadaan $S^{-1}BS = S^{-1}SAS^{-1}S = A$; $T = S^{-1}$, $T^{-1} = S$, eli

$$TBT^{-1} = A$$

Sanotaan, että matriisit A ja B ovat *similaariset*, jos niillä on samat ominaisarvot, mutta mahdollisesti eri ominaisvektorit. Kuvaus $A \mapsto SAS^{-1}$ on *similaarimuunnos*.

Neliömatriisi on singulaarinen, jos ja vain jos sen ominaisarvo on 0. Säännöllisen matriisin A käänteismatriisin A^{-1} ominaisarvot ovat matriisin A ominaisarvojen käänteislukuja (käänteisluvut).

Matriisin A on *positiivisesti definiitti*, jos

$$\mathbf{x}^T A\mathbf{x} > 0$$

kaikille vektoreille $\mathbf{x} \neq 0$. Tällöin kaikki matriisin A ominaisarvot ovat positiivisia reaalilukuja.

$$\left(A\mathbf{x} = \lambda\mathbf{x} \Rightarrow \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(\lambda\mathbf{x}) = \lambda \underbrace{\left(\mathbf{x}^T\mathbf{x}\right)}_{>0,\ \mathbf{x}\neq 0} > 0 \Rightarrow \lambda > 0 \right)$$

Esimerkiksi $A = \begin{pmatrix} 3 & 5 \\ 1 & 3 \end{pmatrix}$ ei ole positiivisesti definiitti, sillä

$$(1, -1)\, A \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0.$$

Kuitenkin matriisin ominaisarvot ovat $\frac{1}{2}\left(6 \pm \sqrt{20}\right) > 0$.

Olkoon esimerkiksi $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ mielivaltainen 2×2-matriisi. Matriisin T ominaisarvot saadaan toisen asteen yhtälön

$$\begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} = \lambda^2 - (a + d)\,\lambda + ad - bc = 0$$

juurina

$$\lambda_1, \lambda_2 = \frac{1}{2}\left(a + d \pm \sqrt{(a-d)^2 + 4bc}\right) \tag{3.3}$$

Lause 7. *Olkoon A reaalinen symmetrinen $n \times n$-matriisi. Silloin*

(i) Matriisin A ominaisarvot ovat reaalisia

(ii) Matriisin A eri ominaisarvoja vastaavat ominaisvektorit ovat ortogonaaliset

(iii) Avaruudella $\mathbb{R}^n$ on ortonormaali kanta, jonka alkiot ovat matriisin A ominaisvektoreita.

Todistus.
(i) Olkoon $\mathbf{u}$ ominaisarvoa λ vastaava (mahdollisesti kompleksinen) ominaisvektori.

Konjugoidaan ja transponoidaan yhtälö (3.1):

$$\begin{aligned}
& A\mathbf{u} = \lambda\mathbf{u} \\
\Rightarrow\ & A\bar{\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}} \\
\Rightarrow\ & \bar{\mathbf{u}}^T A^T = \bar{\lambda}\bar{\mathbf{u}}^T \\
\Rightarrow\ & \bar{\mathbf{u}}^T A = \bar{\lambda}\bar{\mathbf{u}}^T \\
\Rightarrow\ & \bar{\mathbf{u}}^T \underbrace{A\mathbf{u}}_{\lambda\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}^T\mathbf{u} \\
\Rightarrow\ & \bar{\mathbf{u}}^T\lambda\mathbf{u} = \bar{\lambda}\bar{\mathbf{u}}^T\mathbf{u} \\
\Rightarrow\ & \left(\lambda - \bar{\lambda}\right) \underbrace{\bar{\mathbf{u}}^T\mathbf{u}}_{\mathbf{u}\neq 0\ \Rightarrow >0} = 0 \\
\Rightarrow\ & \lambda - \bar{\lambda} = 0 \\
\Rightarrow\ & \lambda = \bar{\lambda} \\
\Rightarrow\ & \lambda \text{ reaalinen}
\end{aligned}$$

(ii) Olkoon $A\mathbf{u} = \lambda\mathbf{u}$ ja $A\mathbf{v} = \mu\mathbf{v}$ $(\lambda \neq \mu)$. Silloin $(A\mathbf{u})^T = \lambda\mathbf{u}^T$, ja siis $\mathbf{u}^T A^T = \lambda\mathbf{u}^T$. Kerrotaan oikealta puolelta vektorilla $\mathbf{v}$, jolloin saadaan

$$\mathbf{u}^T A^T \mathbf{v} = \lambda\mathbf{u}^T\mathbf{v}$$

Koska matriisi A on symmetrinen, on $A^T = A$, siis vasen puoli on

$$\mathbf{u}^T A\mathbf{v} = \mathbf{u}^T(\mu\mathbf{v}) = \mu\mathbf{u}^T\mathbf{v}$$

Näin ollen $\lambda \mathbf{u}^T\mathbf{v} = \mu \mathbf{u}^T\mathbf{v} \;\Rightarrow\; (\lambda - \mu)\,\mathbf{u}^T\mathbf{v} = 0$, ja koska oletettiin, että $\lambda \neq \mu$, niin $\mathbf{u}^T\mathbf{v} = 0$ eli vektorit $\mathbf{u}$ ja $\mathbf{v}$ ovat ortogonaaliset.

(iii) Jos matriisilla A on n reaalista (erisuurta) ominaisarvoa $\lambda_1, \ldots, \lambda_n$, löydetään ortonormaali kanta etsimällä kutakin ominaisarvoa λ_j vastaava normalisoitu ominaisvektori $\mathbf{u}_j$. Kohdan (ii) perusteella vektorit $\mathbf{u}_1, \ldots, \mathbf{u}_n$ ovat nimittäin keskenään ortogonaaliset, ja niiden virittämän avaruuden $\mathbb{R}^n$ aliavaruuden dimensio on n.

Jos matriisin A ominaisarvojen lukumäärä on $< n$, todistus on hankalampi. Rajoitutaan tapaukseen $n = 2$, jolloin A on muotoa

$$A = \begin{pmatrix} a & b \\ b & d \end{pmatrix} \tag{3.4}$$

ja ominaisarvot ovat

$$\lambda_1, \lambda_2 = \frac{1}{2}\left(a + d \pm \sqrt{(a-d)^2 + 4b^2}\right). \tag{3.5}$$

Koska ominaisarvojen lukumäärä on $< n = 2$, niin $\lambda_1 = \lambda_2$, jolloin on oltava $(a-d)^2 = 4b^2 = 0$. Toisin sanoen $A = aI$, ja kaikki avaruuden $\mathbb{R}^2$ (nollasta eroavat) vektorit ovat ominaisarvoa λ vastaavia ominaisvektoreita.

□

Lause 8. *Olkoon A (reaalinen) symmetrinen $n \times n$-matriisi. Silloin on olemassa ortogonaalinen $n \times n$-matriisi U ja diagonaalimatriisi $\Lambda = diag\,(\lambda_1, \ldots, \lambda_n)$ siten, että luvut $\lambda_1, \ldots, \lambda_n$ ovat matriisin A ominaisarvoja ja*

$$A = U\Lambda U^T \tag{3.6}$$

$$U^T A = \underbrace{U^T U}_{I}\Lambda U^T = \Lambda U^T$$

$$U^T A U = \Lambda \underbrace{U^T U}_{I} = \Lambda$$

Symmetrinen matriisi voidaan siis diagonalisoida ortogonaalisella similaarisuusmuunnoksella. Sanotaan, että (3.6) on matriisin A *spektraalihajotelma.*

Todistus. Lauseen 7 nojalla avaruuden $\mathbb{R}^n$ on ortonormaali kanta $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$, jonka alkiot ovat matriisin A ominaisvektoreita. Tällöin

$$A\mathbf{u}_i = \lambda_i \mathbf{u}_i \;\; (1 \leq i \leq n)\,, \tag{3.7}$$

missä luvut λ_i ovat matriisin A ominaisarvoja. Olkoon U matriisi, jonka i:s sarake on $\mathbf{u}_i$ ja olkoon $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Silloin (3.7) voidaan kirjoittaa muotoon

$$AU = U\Lambda$$

ja kertomalla oikealta matriisilla U^T, saadaan matriisin U ortogonaalisuden perusteella (3.6):

$$A\underbrace{UU^T}_{I} = U\Lambda U^T \Leftrightarrow A = U\Lambda U^T$$

□

3.2 Singulaariarvohajotelma

Huomautus: Ennen lukua 3.2 on luennoitu luvut 4.1 ja 4.2.

Lemma 9. *Olkoon $\boldsymbol{u}$ n-vektori siten, että $\|\boldsymbol{u}\|_2 = 1$. Silloin on olemassa ortogonaalinen $m \times m$-matriisi U siten, että $\boldsymbol{u} = \boldsymbol{u}_i$.*

Todistus. $\dim R(\mathbf{u}) = 1 \Rightarrow \dim \mathcal{N}(\mathbf{u}^T) = m - 1$. Valitaan avaruuden $\mathcal{N}(\mathbf{u}^T)$ ortonormaali kanta $\{\mathbf{u}_2, \ldots, \mathbf{u}_n\}$. Silloin $U = (\mathbf{u}\ \mathbf{u}_2\ \ldots\ \mathbf{u}_n)$. □

Lause 10. *Olkoon A mielivaltainen $m \times n$-matriisi. Silloin on olemassa ortogonaalinen $m \times m$-matriisi U, ortogonaalinen $n \times n$-matriisi V ja diagonaalinen $m \times n$-matriisi $S = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ siten, että $p = \min(m, n)$, $\sigma_i \geq 0$ kaikille $i = 1, \ldots, p$ ja*

$$A = USV^T \tag{3.8}$$

Sanotaan, että luvut σ_i ovat matriisin A *singulaariarvoja*, ja että (3.8) on matriisin A *singulaariarvohajotelma*. Hajotelma (3.8) voidaan aina muodostaa siten, että $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$, jolloin $\sigma_1(A)$ on matriisin A suurin singulaariarvo.

$$AV = US$$
$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i$$

Todistus. Oletetaan, että $m \geq n$ ja todistetaan väite induktiolla n:n suhteen. Voidaan myös olettaa, että $A \neq 0$, sillä tapauksessa $A = 0$ (3.8) pätee triviaalisti $S = 0$.

Tapaus $n = 1$: Merkitään $\sigma = \sqrt{A^T A}$ ja $\mathbf{u} = \sigma^{-1} A$. Silloin $\mathbf{u}$ on yksikkövektori. $\mathbf{u}^T\mathbf{u} = \sigma^{-2} A^T A = 1$. Lemman 9 mukaan on olemassa m-vektorit $\mathbf{u}_2, \ldots, \mathbf{u}_m$ siten, että matriisi $U = (\mathbf{u}\ \mathbf{u}_2\ \ldots\ \mathbf{u}_m)$ on ortogonaalinen. Silloin

$$U^T A = \begin{pmatrix} \mathbf{u}^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_m^T \end{pmatrix} \sigma\mathbf{u} = \sigma \begin{pmatrix} \mathbf{u}^T\mathbf{u} \\ \mathbf{u}_2^T\mathbf{u} \\ \vdots \\ \mathbf{u}_m^T\mathbf{u} \end{pmatrix} = \begin{pmatrix} \sigma \\ 0 \\ \vdots \\ 0 \end{pmatrix} = S$$

on diagonaalimatriisi ja

$$A = UU^T A = US = USI$$

Hajotelma (3.8) pätee siis matriisin V ollessa yksikkömatriisi.

Induktio-oletus ja induktioaskel. Oletetaan, että väite pätee arvolla n, ja tarkastellaan tyyppiä $(m+1) \times (n+1)$ olevaa matriisia A, missä $m \geq n$. Funktio $\mathbf{v} \mapsto \|A\mathbf{v}\|_2$ on jatkuva avaruudessa $\mathbb{R}^{n+1}$, joten se saavuttaa maksiminsa avaruuden $\mathbb{R}^{n+1}$ suljetussa ja rajoitetusa joukossa $\{\mathbf{v} \in \mathbb{R}^{n+1}\ ;\ \|\mathbf{v}\|_2 = 1\}$. Valitaan yksikkövektori $\mathbf{v}_1 \in \mathbb{R}^{n+1}$ siten, että

$$\|A\mathbf{v}_1\|_2 = \max\{\|A\mathbf{v}\|_2\ ;\ \|\mathbf{v}\|_2 = 1\} \tag{3.9}$$

ja merkitään $\sigma_1 = \|A\mathbf{v}_1\|_2$. Lemman 9 nojalla on olemassa $n+1$-vektorit $\mathbf{v}_2, \ldots, \mathbf{v}_{n+1}$ siten, että matriisi $V_1 = (\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_{n+1})$ on ortogonaalinen. Soveltamalla Lemmaa 9 avaruuden $\mathbb{R}^{n+1}$ yksikkövektoriin $\mathbf{u}_1 = \sigma_1^{-1} A\mathbf{v}_1$, nähdään vastaavasti, että on olemassa $(m+1)$-vektorit $\mathbf{u}_2, \ldots, \mathbf{u}_{m+1}$ siten, että matriisi $U = (\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_{m+1})$ on ortogonaalinen.
Tarkastellaan matriisia

$$\begin{aligned} U_1^T A V_1 &= \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_{m+1}^T \end{pmatrix} (A\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_{n+1}) \\ &= \begin{pmatrix} \mathbf{u}_1^T A\mathbf{v}_1 & \mathbf{u}_1^T A\mathbf{v}_2 \ldots \mathbf{u}_1^T A\mathbf{v}_{n+1} \\ \mathbf{u}_2^T A\mathbf{v}_1 & \\ \vdots & A_2 \\ \mathbf{u}_{m+1}^T A\mathbf{v}_1 & \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & \mathbf{z}^T \\ 0 & A_2 \end{pmatrix}, \end{aligned} \tag{3.10}$$

sillä

$\mathbf{u}_1^T A\mathbf{v}_1 = \mathbf{u}_1^T \sigma_1 \mathbf{u}_1 = \sigma_1 \|\mathbf{u}_1\|_2^2 = \sigma_1$
$\mathbf{u}_2^T A\mathbf{v}_1 = \mathbf{u}_2^T \sigma_1 \mathbf{u}_1 = 0,$ sillä $\mathbf{u}_1$ ja $\mathbf{u}_2$ ovat ortogonaaliset
$\vdots$
$\mathbf{u}_{m+1}^T A\mathbf{v}_1 = \mathbf{u}_{m+1}^T \sigma_1 \mathbf{u}_1 = \sigma_1 \mathbf{u}_{m+1}^T \mathbf{u}_1 = 0,$ sillä $\mathbf{u}_1$ ja $\mathbf{u}_{m+1}$ ovat ortogonaaliset.

Yhtälössä (3.10) $\mathbf{z} = \begin{pmatrix} \mathbf{u}_1^T A\mathbf{v}_1 \\ \vdots \\ \mathbf{u}_1^T A\mathbf{v}_{n+1} \end{pmatrix}$ ja A_2 on $m \times n$-matriisi. Induktio-oletuksen nojalla on olemassa ortogonaalinen $m \times m$-matriisi

$$S_2 = \operatorname{diag}(\sigma_, \ldots, \sigma_{n+1})$$

siten, että $\sigma_i \geq 0$ kaikille $i = 2, \ldots, n+1$ ja

$$A_2 = U_2 S_2 V_2^T$$

Vektorin $\mathbf{z}$ tutkimiseksi tarkastellaan avaruuden $\mathbb{R}^{n+1}$ yksikkövektoria

$$\zeta = \frac{1}{\sqrt{\sigma_1^2 + \mathbf{z}^T\mathbf{z}}} \begin{pmatrix} \sigma_1 \\ \mathbf{z} \end{pmatrix}$$

Koska V_1 on ortogonaalinen, myös $V_1\zeta$ on yksikkövektori (vrt. (4.7)), ja

$$\sigma_1^2 \geq \|AV_1\zeta\|_2^2 \tag{3.11}$$

Toisaalta, matriisin U_1 (eli myös matriisin U_1^T) ortogonaalisuudesta seuraa, että

$$\|AV_1\zeta\|_2^2 = \left\|U_1^T AV_1\zeta\right\|_2^2 \tag{3.12}$$

Tässä oikea puoli on vähintään yhtäsuuri kuin vektorin $U_1^T AV\zeta$ ensimmäisen komponentin neliö, joten voidaan laskea kaavan (3.10) perusteella

$$\left\|U_1^T AV_1\zeta\right\|_2^2 = \left(\frac{\left(\sigma_1^2 + \mathbf{z}^T\mathbf{z}\right)}{\sqrt{\sigma_1^2 + \mathbf{z}^T\mathbf{z}}}\right)^2 = \sigma_1^2 + \mathbf{z}^T\mathbf{z} \tag{3.13}$$

Yhdistämällä (3.11), (3.12) ja (3.13) saadaan lopulta

$$\sigma_1^2 \geq \sigma_1^2 + \mathbf{z}^T\mathbf{z},$$

mikä voi toteutua vain, jos $\mathbf{z}^T\mathbf{z} = \|\mathbf{z}\|_2^2 = 0$, eli kun $\mathbf{z} = 0$. Näin ollen (3.10) saadaan muotoon

$$U_1^T AV_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & A_2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & U_2 S_2 V_2^T \end{pmatrix} \tag{3.14}$$

Määritellään ortogonaaliset lohkomatriisit

$$U_2' = \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} \text{ ja } V_2' = \begin{pmatrix} 1 & 0 \\ 0 & V_2 \end{pmatrix}.$$

Silloin (3.14) voidaan kirjoittaa

$$U_1^T A V_1 = U_2' \begin{pmatrix} \sigma_1 & 0 \\ 0 & S_2 \end{pmatrix} (V_2')^T = U_2' S (V_2')^T.$$

Kertomalla vasemmalta matriisilla U_1 ja oikealta matriisilla V_1^T, saadaan lopulta

$$\underbrace{U_1 U_1^T}_{I} A \underbrace{V_1 V_1^T}_{I} = A = \underbrace{U_1 U_2'}_{U} S \underbrace{V_1 V_2'}_{V}$$

(3.8) pätee valitsemalla $U = U_1 U_2'$ ja $V = V_1 V_2'$.
Tapauksessa $m < n$ tarkastellaan matriisin A asemasta matriisia A^T, jolle äskeisen todistuksen perusteella saadaan singulaariarvohajotelma

$$A^T = USV^T$$

Tällöin $A = \left(A^T\right)^T = \left(USV^T\right)^T = VS^T U^T$ on matriisin A singulaariarvohajotelma. □

Huomautus: Voidaan näyttää, että matriisin A aste on yhtäsuuri kuin positiivisten singulaariarvojen lukumäärä.

Singulaariarvojen yksikäsitteisyys:

$$A^T A = \left(USV^T\right)^T \left(USV^T\right) = VS^T \underbrace{U^T U}_{I} SV^T = VS^T SV^T,$$

(joka on matriisin $A^T A$ spektraalihajotelma), joten matriisin A singulaariarvot ovat yhtä kuin matriisin $A^T A$ ominaisarvojen ei-negatiiviset neliöjuuret, mikäli $m \geq n$. Jos $m < n$, niin matriisin $S^T S$ diagonaalilla on lukujen σ_i^2 ohella myös nollia. Tällöinkin *singulaariarvot saadaan matriisin $A^T A$ ominaisarvojen ei-negatiivisina neliöjuurina.*

Jos matriisi A on symmetrinen, niin sen singulaariarvot ovat matriisin A ominaisarvojen itseisarvoja:

$$A = U\Lambda U^T \; ; \; A^T A = U\Lambda^2 U^T$$

4 Normit

4.1 Vektorinormit

Vektorinormilla $\|\cdot\|$ on seuraavat kolme ominaisuutta

(i) $\|\mathbf{x}\| > 0$ kaikille $\mathbf{x} \neq 0$

(ii) jokaiselle skalaarille γ pätee $\|\gamma\mathbf{x}\| = |\gamma| \, \|\mathbf{x}\|$

(iii) mielivaltaisille vektoreille $\mathbf{x}$ ja $\mathbf{y}$ pätee kolmioepäyhtälö $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

Esimerkiksi n-vektorin *p-normi* määritellään yhtälöllä

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad (p \geq 1, p \in \mathbb{Z}) \tag{4.1}$$

Tavallisimmat vektorinormit ovat

- 1-normi: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- 2-normi eli euklidinen normi: $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$
- maksiminormi: $\|\mathbf{x}\|_\infty = \max\{|x_i| \ ; \ 1 \leq i \leq n\}$

Esimerkiksi jos $\mathbf{x} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$, niin

- $\|\mathbf{x}\|_1 = |1| + |-2| = 3$
- $\|\mathbf{x}\|_2 = \sqrt{1^2 + (-2)^2} = \sqrt{5}$
- $\|\mathbf{x}\|_\infty = \max\{|1|, |-2|\} = 2$

Väite 11. *Kuvaus $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|$ on jatkuva.*

Todistus.

$$\begin{aligned} \|\mathbf{x}\| - \|\mathbf{y}\| &= \|\mathbf{x} - \mathbf{y} + \mathbf{y}\| - \|\mathbf{y}\| \\ &\leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y}\| - \|\mathbf{y}\| \\ &= \|\mathbf{x} - \mathbf{y}\| \\ \|\mathbf{y}\| - \|\mathbf{x}\| &\leq \|\mathbf{y} - \mathbf{x}\| = \|\mathbf{x} - \mathbf{y}\| \end{aligned}$$

Siis $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\| \to 0$, kun $\mathbf{x} \to \mathbf{y}$. □

Vektoreiden $\mathbf{x}$ ja $\mathbf{y}$ välisen kulman θ antaa kaava

$$\cos\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2} \tag{4.2}$$

Kulma voidaan laskea, sillä avaruudessa $\mathbb{R}^n$ pätee *Schwartzin epäyhtälö*

$$\left|\mathbf{x}^T\mathbf{y}\right| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \tag{4.3}$$

Vektorin $\mathbf{x} \neq \mathbf{0}$ suuntainen normalisoitu *yksikkövektori* $\mathbf{u}$ toteuttaa ehdon $\|\mathbf{u}\| = 1$ ja riippuu käytetystä normista:

$$\mathbf{u} = \frac{1}{\|\mathbf{x}\|}\mathbf{x} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \tag{4.4}$$

4.2 Matriisinormit

Matriisinormi toteuttaa samat kolme aksioomaa kuin vektorinormi:

(i) $\|A\| > 0$ kaikille matriiseille $A \neq 0$

(ii) jokaiselle skalaarille γ pätee $\|\gamma A\| = |\gamma| \|A\|$

(iii) $\|A + B\| \leq \|A\| + \|B\|$

Lisäksi vaaditaan, että

(iv) $\|AB\| \leq \|A\| \|B\|$,

aina, kun tulo AB on määritelty.

Jokaiseen vektorinormiin liittyy vastaava matriisinormi, jota kutsutaan vektorinormin *indusoimaksi* matriisinormiksi:

$$\|A\| = \max \|A\mathbf{u}\|, \ \|\mathbf{u}\| = 1$$

Vaihtoehtoinen määritelmä saadaan kaavan (4.4) nojalla:

$$\|A\| = \max_{x \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \tag{4.5}$$

$$\left(\mathbf{u} = \frac{1}{\|\mathbf{x}\|}\mathbf{x},\ A\mathbf{u} = \frac{1}{\|\mathbf{x}\|}A\mathbf{x},\quad \|A\mathbf{u}\| = \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \right)$$

Määritelmästä seuraa heti, että

$$\|A\mathbf{x}\| \leq \|A\| \, \|\mathbf{x}\| \tag{4.6}$$

kaikille vektoreille $\mathbf{x}$.
Jos A on tyyppiä $m \times n$, ja $\mathbf{x}$ on n-vektori, $A\mathbf{x} \in \mathbb{R}^m$. (Ts. kaikki normit määritelty eri avaruuksissa.)

Vektoreiden 1-normia, 2-normia ja maksiminormia vastaavat matriisinormit ovat:

- $\|A\|_1 = \max_j \|\mathbf{a}_j\|_1$ (suurin sarakesumma)
- $\|A\|_2 = \sigma_1(A)$ (suurin singulaariarvo)
- $\|A\|_\infty = \max_i \|\mathbf{a}'_i\|_1$ (suurin rivisumma)

Vektorinormi $\|\cdot\|$ ja matriisinormi $\|\cdot\|'$ ovat *yhteensopivat*, jos

$$\|A\mathbf{x}\| \leq \|A\|' \, \|\mathbf{x}\|$$

kaikille matriiseille A ja vektoreille $\mathbf{x}$. Kaavan (4.6) perusteella vektorinormi on aina yhteensopiva indusoimansa matriisinormin kanssa.

Esimerkiksi $m \times n$-matriisin A *Frobeniuksen normi* on

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}$$

Voidaan osoittaa, että $\|\cdot\|_F$ ei ole minkään vektorinormin indusoima, ja että

$$\|A\|_F^2 = \operatorname{tr}\left(A^T A\right),$$

jossa tr tarkoittaa matriisin *jälkeä* eli diagonaalialkioiden summaa.

Kuitenkin $\|\cdot\|_F$ on yhteensopiva euklidisen normin kanssa:

$$\|A\mathbf{x}\|_2 \leq \|A\|_F \|\mathbf{x}\|_2$$

Kaikki vektorinormit ovat ekvivalentteja seuraavassa mielessä: jos $\|\cdot\|$ ja $\|\cdot\|'$ ovat vektorinormeja, on olemassa tarkasteltavan avaruuden dimensiosta riippumattomat positiiviset vakiot c_l ja c_u siten, että

$$c_l \|\mathbf{x}\| \leq \|\mathbf{x}\|' \leq c_u \|\mathbf{x}\|$$

kaikille vektoreille $\mathbf{x}$. Vastaava tulos pätee myös matriisinormeille.

Jos A on ortogonaalinen $n \times n$-matriisi, niin kaikille n-vektoreille pätee

$$\|A\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \tag{4.7}$$

Osa II

Numeerinen laskenta ja häiriöalttius

1 Virhe ja sen kertaluku

1.1 Absoluuttinen ja suhteellinen virhe

Jos $\bar{x}$ on reaaliluvun x likiarvo, niin vastaava *absoluuttinen virhe* on

$$e_A = |x - \bar{x}|$$

Likiarvon $\bar{x}$ suhteellinen virhe on

$$e_R = \frac{|x - \bar{x}|}{|x|} \tag{1.1}$$

1.2 Virheen kertaluku

Olkoon τ jokin kymmenen potenssi siten, että $\tau = 10^k$, missä $k \in \mathbb{Z}$. Sanotaan, että reaaliluku $x \neq 0$ on *kertalukua* τ, jos

$$|x| = \mu\tau, \text{ missä } 1 \leq \mu < 10. \tag{1.2}$$

Huomautus: Sanaa *kertaluku* tullaan käyttämään myös muissa yhteyksissä.

Esimerkiksi n-ulotteisen lineaarisen yhtälöryhmän ratkaisemiseen tarvittavien laskutoimitusten lukumäärä on ”kertalukua n^3” tai ”kertalukua $\frac{2}{3}n^3$”; tällöin on kyseessä asymptoottinen arvio laskutoimitusten lukumäärästä tai sen suuruusluokasta, kun $n \to \infty$.

2 Pyöristys- ja katkaisuvirheistä

2.1 Liukuluvut

Olkoon $b \geq 2$ tietokoneen käyttämän lukujärjestelmän kantaluku (esim. $b = 2$ tai $b = 16$). Jokainen reaaliluku ξ voidaan esittää muodossa

$$\xi = \pm m \cdot b^z, \tag{2.1}$$

missä z on kokonaisluku, toisin sanoen eksponentti, ja m on ei-negatiivinen reaaliluku, *mantissa*.

Sanotaan, että esitys (2.1) on *normalisoitu*, jos $b^{-1} \leq m < 1$; tällöin ξ määrää luvun m yksikäsitteisesti, mikäli $\xi \neq 0$. Usein käytetään myös normalisointia, jossa $1 \leq m < b$; esimerkiksi käsin laskettaessa kirjoitetaan mieluummin

$1.5 \cdot 10^2$, vaikka normalisoitu esitys olisi $0.15 \cdot 10^3$.

Tietokoneessa reaaliluku (2.1) korvataan niin sanotulla *liukuluvulla*, joka on muotoa

$$x = \pm 0.m_1 m_2 \ldots m_t \cdot b^z \tag{2.2}$$

Tässä mantissaa m vastaava osa $0, m_1 m_2 \ldots m_t$ on b-järjestelmän reaaliluku

$$\sum_{j=1}^{t} m_j b^{-j}$$

ja jokainen m_j on ehdon $0 \leq m_j < b$ toteuttava kokonaisluku.

Normalisoidussa liukulukuesityksessä $m_1 \neq 0$, mikäli $x \neq 0$. Numeroiden m_j lukumäärä t riippuu tietokoneesta ja käytettävästä tarkkuudesta.

IEEE-standardin mukaan $b = 2$ ja $t = 23$ (normaali tarkkuus) tai $t = 52$ (kaksoistarkkuus). Eksponentti z vaihtelee konekohtaisesti jollakin välillä

$$-k_1 \leq z \leq k_2, \tag{2.3}$$

missä k_1 ja k_2 ovat positiivisia kokonaislukuja.

IEEE-standardin normaalitarkkuudessa $k_1 = 126$ ja $k_2 = 127$; tällöin käytetyssä liukulukuaritmetiikassa pienin ja suurin positiiviluku ovat vastaavasti noin $2^{-126} \approx 1 \cdot 10^{-38}$ ja $2^{127} \approx 2 \cdot 10^{38}$.

Ylivuoto tai *alivuoto* syntyy, jos jonkin laskutoimituksen tuloksena saadun liukuluvun eksponentti ei toteuta ehtoa (2.3).

Käytetystä liukulukujärjestelmästä riippuu, mitkä reaaliluvut voidaan esittää muodossa (2.2) siten, että (2.3) pätee; sanotaan, että tällainen luku x on parametreihin b, t, k_1 ja k_2 liittyvä liukulukujärjestelmän *jäsen*. On olemassa reaalilukuja (esim. π ja $\sqrt{2}$), jotka eivät ole minkään liukulukujärjestelmän jäseniä. Toisaalta, esimerkiksi $\frac{1}{10}$ on jäsen desimaalijärjestelmässä ($b = 10$), mutta ei 2-järjestelmässä.

Tietokoneissa tapahtuvia laskutoimituksia varten jokainen reaaliluku x on korvattava luvun x *liukuvastineella*

$$\bar{x} = fl(x),$$

joka on tietokoneessa käytetyn liukulukujärjestelmän jäsen.

Kuvaus $x \mapsto fl(x)$ riippuu käytetystä pyöristyssäännöstä, mutta yli- ja alivuototapauksia lukuunottamatta $fl(x)$ on yleensä lukua x lähinnä sijaitseva liukulukujärjestelmän jäsen (*normaalipyöristys*). Tällöin luvun x likiarvon $\bar{x}$ absoluuttisella virheellä on yläraja

$$|fl(x) - x| \leq \frac{1}{2} b^{1-t} |x| \tag{2.4}$$

ja suhteellinen virhe on enintään $\frac{1}{2} b^{1-t}$. Käytettäessä niin sanottua *katkaisevaa pyöristystä*, mantissasta poistetaan kaikki numeron m_t jälkeiset numerot. Tällöin luvuilla x ja $fl(x)$ on sama eksponentti ja myös mantissojen t ensimmäistä numeroa ovat samat; $x = \pm 0, m_1 m_2 \ldots m_t \underbrace{m_{t+1} \ldots}_{\text{pois}} \cdot b^z$.

Jos esimerkiksi $b = 10$ ja $t = 3$, niin luvuilla 0.3141 ja 0.3149 on sama liukulukuvastine 0.314.

Pyöristysvirheellä (katkaisevassa pyöristyksessä) on yläraja:

$$|fl(x) - x| \leq b^{1-t} |x| \tag{2.5}$$

ja suhteellinen virhe on enintään b^{1-t}.

Pyöristysyksikkö $\bar{u}$ on kaavoissa (2.4) ja (2.5) esiintyvä pyöristyksessä syntyvän suhteellisen virheen yläraja:

$$\bar{u} = \begin{cases} \frac{1}{2} b^{1-t} & \text{normaalipyöristyksessä} \\ b^{1-t} & \text{katkaisevassa pyöristyksessä} \end{cases}$$

IEEE-standardissa noudatetaan normaalipyöristystä, jolloin

$$\bar{u} = \begin{cases} 2^{-23} \approx 10^{-7} & \text{normaalitarkkuudessa} \\ 2^{-53} \approx 2 \cdot 10^{-16} & \text{kaksoistarkkuudessa} \end{cases}$$

2.2 Liukulukuaritmetiikkaa

Olkoon a ja b liukulukujärjestelmän jäseniä. Käytetään merkintää "op" jollekin laskutoimituksista "+", "−", "·" tai "/". Monesti $fl(a \text{ op } b) \neq a \text{ op } b$.

Jos esimerkiksi $b = 10$, $t = 4$, ja $a = 1.6$, $b = 0.002342$. Tällöin $a + b = 1.602342$ ja $fl(a + b) = 1.602 \neq a + b$; yhteenlaskun lopputulos sisältää pyöristysvirheen

Tietokoneessa yleisesti

$$fl(a \text{ op } b) = (a \text{ op } b)(1 + \delta), \tag{2.6}$$

missä $|\delta| \leq \bar{u}$.

Pyöristysvirheiden takia tavalliset reaalilukujen laskusäännöt eivät päde liukulukuartmetiikassa. Jos esimerkiksi x, y ja z ovat jäseniä liukulukujärjestelmässä, niin monesti

$$fl(fl(x+y)+z) \neq fl(x+fl(y+z)),$$

eli liukulukujen yhteenlasku ei ole liitännäinen laskutoimitus. Numeerinen esimerkki tapauksessa $b = 10$, $t = 4$; $x = -1$, $y = 1$ ja $z = 0.001$:

$$fl(x+y) = 0.0 \ ; \ fl(fl(x+y)+z) = 0.0001$$
$$fl(y+z) = 1.0 \ ; \ fl(x+fl(y+z)) = 0.0$$

2.2.1 Suppeneminen

Esimerkiksi harmoninen sarja $1 + \frac{1}{2} + \frac{1}{3} + \ldots$ saadaan suppenevaksi korvaamalla sen termit näiden liukulukuvastineilla.

2.3 Merkitsevien numeroiden kumoutuminen

Eksponenttifunktion e^x arvoja voidaan laskea käyttämällä Taylorin sarjaa

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \tag{2.7}$$

Esimerkiksi $x = -5.5$, $b = 10$, $t = 5$. Tämän liukulukujärjestelmän aritmetiikassa sarjan summaksi tulee 0.0026363. Lopputulos on epätarkka, sillä oikein pyöristettynä $e^{-5,5} \approx 0.0040868$. Syy epätarkkuuteen on ns. *merkitsevien numeroiden kumoutuminen*, jollaista yleisemmin tapahtuu vähennyslaskussa silloin, kun vähennettävä ja vähentäjä ovat likimain yhtäsuuret.

Sarja (2.7) on vuorotteleva arvolla $x = -5.5$; alkupään termit tarkasteltavassa liukulukujärjestelmässä laskettuna ovat 1.0000, -5.5000, 15.125, -27.730, 38.129, -41.942 ja 38.446. Pyöristyksessä aiheutuva virhe kussakin termissä voi olla esimerkiksi kertalukua 10^{-4} Näiden virheiden kasautuminen aiheuttaa epätarkan lopputuloksen. Oikean viisinumeroisen likiarvon saavuttamiseksi sarjan termit tulisi laskea yhdentoista merkitsevän numeron tarkkuudella.

2.4 Toisen asteen yhtälöistä

Toisen asteen yhtälön

$$ax^2 + bx + c = 0 \tag{2.8}$$

juuret saadaan kaavalla

$$x_1, x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{2.9}$$

Liukulukuaritmetiikassa käytettäessä ratkaisukaava (2.9) voi johtaa hyvin epätarkkaan lopputulokseen. Olkoon esimerkiksi $a = 1$, $b = -12.4$ ja $c = 0.494$. Silloin juuret ovat kuuden merkitsevän numeron tarkkuudella $x_1 = 12.3600$ ja $x_2 = 0.0399675$. Jos juuret lasketaan kaavasta (2.9) käyttämällä kolminumeroista aritmetiikkaa ja katkaisevaa pyöristystä, saadaan

$$x = \frac{12.4 \pm \sqrt{153 - 1.97}}{2} = \frac{12.4 \pm 12.2}{2}$$

eli $\bar{x}_1 = 12,3$ ja $\bar{x}_2 = 0,1$. Pienemmässä juuressa ei siis ole yhtään merkitsevää numeroa. Syynä on jälleen merkitsevien numeroiden kumoutuminen.

Tarkkuutta voidaan parantaa valitsemalla toinen ratkaisualgoritmi. Lasketaan ensin kaavan (2.9) se juuri, jolla valittu luvun $b^2 - 4ac$ neliöjuuren haara on samanmerkkinen kuin $-b$; tällöin merkitsevien numeroiden kumoutumista ei tapahdu. Jäljellä oleva juuri saadaan tämän jälkeen ratkaisemalla yhtälö

$$x_1 x_2 = \frac{c}{a} \tag{2.10}$$

Äskeisessä esimerkissä saataisiin siis $\bar{x}_1 = 12.3$ ja kaavan (2.10) perusteella

$$\bar{x}_2 = fl\left(\frac{0.494}{12.3}\right) = 0.0401$$

Nyt pienemmän juuren likiarvo on huomattavasti tarkempi.

Pyöristysvirheiden vaikutus siis *korostuu* vähennyslaskuissa silloin, kun vähennettävä ja vähentäjä ovat likimain yhtäsuuret.

3 Probleeman häiriöalttius

Numeerisissa probleemissa lähtöarvot muodostavat yleensä äärellisen lukujonon (vektorin) $\mathbf{d}$, ja probleeman ratkaisu on jokin vektorin $\mathbf{d}$ (yleensä vektoriarvoinen) funktio $s(\mathbf{d})$. Probleeman häiriöalttius riippuu kuvauksesta $\mathbf{d} \mapsto s(\mathbf{d})$ ja sitä voidaan mitata tutkimalla osamäärän

$$\frac{\|s(\mathbf{d}_1) - s(\mathbf{d}_2)\|}{\|\mathbf{d}_1 - \mathbf{d}_2\|} \tag{3.1}$$

mahdollisia arvoja. Jos osamäärä (3.1) on pieni kaikille mahdollisille arvopareille $\mathbf{d}_1$ ja $\mathbf{d}_2$, $\mathbf{d}_1 \neq \mathbf{d}_2$, on probleema *hyvänlaatuinen*. Jos taas osamäärä (3.1) voi saada hyvin suuria arvoja, probleema on *pahanlaatuinen* eli häiriöaltis. Häiriöalttius ei riipu funktion $s(\mathbf{d})$ arvojen laskemiseen käytetystä algoritmista.

Esimerkki: Neljännen asteen yhtälön ratkaiseminen. Tarkastellaan yhtälöitä

$$(x-1)^4 = 0 \tag{3.2}$$

ja

$$(x-1)^4 = 10^{-8} \tag{3.3}$$

Yhtälön (3.2) ainoa juuri (nelinkertainen) on $x = 1$ ja eräs yhtälön (3.3) juurista on $1 - 10^{-2}$. Neljännen asteen yhtälöä ratkaistaessa vektorin $\mathbf{d}$ muodostavat yhtälön kertoimet ja vektorin $s(\mathbf{d})$ yhtälön neljä juurta — esimerkiksi yhtälölle (3.2)

$$s(\mathbf{d}) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Yhtälöillä (3.2) ja (3.3) lähtöarvojen etäisyys (osamäärän (3.1) nimittäjä) on kertalukua 10^{-8}. Juurivektorien etäisyys (osamäärän (3.1) osoittaja) on sen sijaan vähintään kertalukua 10^{-2}. Osamäärä (3.1) on tällöin vähintään kertalukua 10^6, ja probleema on siis pahanlaatuinen.

Esimerkki: 20. asteen yhtälön ratkaiseminen. Olkoon

$$W(x) = x^{20} - 210x^{19} + \cdots + 20! = (x-1)(x-2)\ldots(x-19)(x-20).$$

$W(x)$ on niin sanottu *Wilkinsonin polynomi*, jonka nollakohdat ovat $1, 2, \ldots, 19$ ja 20. Tarkastellaan yhtälön $W(x) = 0$ ohella toista 20. asteen yhtälöä

$$W(x) = 2^{-23}x^{19}, \tag{3.4}$$

jonka kertoimet ovat 19. asteen kerrointa lukuunottamatta samat kuin yhtälöllä $W(x) = 0$. Yhtälö (3.4) syntyy, kun alkuperäisen yhtälön $W(x) = 0$ 19. asteen kertoimessa tapahtuu kertalukua 10^{-7} oleva häiriö (häiriön suhteellinen kertaluku on vain 10^{-9}). Häiriön tuloksena osa juurista on kompleksisia, ja etäisyys alkuperäisestä juurivektorista on vähintään kertalukua 1 (riippuu käytetystä normista). Probleema on näin ollen pahanlaatuinen.

4 Virheanalyysiä

4.1 Etenevä virheanalyysi

Edellä tarkastellun numeerisen probleeman $\mathbf{d} \mapsto s(\mathbf{d})$ täsmällinen ratkaisu s poikkeaa yleensä ratkaisussa käytetyn algoritmin avulla saadusta likiarvosta, jota merkitään $s_c(\mathbf{d})$. Monesti on kuitenkin olemassa vektorista $\mathbf{d}$ riippumattomat ei-negatiiviset vakiot δ_f ja ε_f siten, että

$$\|s_c(\mathbf{d}) - s\| \leq \delta_f + \varepsilon_f \|s\| \tag{4.1}$$

Jos lisäksi nämä ns. *virhevakiot* ovat pieniä, sanotaan, että algoritmi on (numeerisesti) *stabiili*.

Esimerkki: Oletetaan, että tehtävänä on laskea reaaliluvun x neliö x^2. Valitaan algoritmi, joka aluksi etsii luvun x liukulukuvastineen $\tilde{x} = fl(x)$ ja antaa lopputulokseksi luvun $\tilde{x}^2$ liukulukuvastineen $s_c = fl(\tilde{x}^2)$.

$$\begin{aligned} &|fl(x) - x| \leq \bar{u}\,|x| \\ (2.6) \Rightarrow &\left|fl(x^2) - \tilde{x}^2\right| \leq \bar{u}\,|\tilde{x}|^2 \end{aligned}$$

$$\begin{aligned} \left|s_c - x^2\right| &= \left|fl(\tilde{x}^2) - x^2\right| \\ &= \left|\left(fl(\tilde{x}^2) - x^2\right) + \left(\tilde{x}^2 - x^2\right)\right| \\ &\leq \left|fl(\tilde{x}^2) - \tilde{x}^2\right| + \left|\tilde{x}^2 - x^2\right| \\ &\leq \bar{u}\,|\tilde{x}|^2 + \left|\tilde{x}^2 - x^2\right| \\ &= \bar{u}\left(\tilde{x}^2 - x^2 + x^2\right) + \left|\tilde{x}^2 - x^2\right| \\ &\leq (\bar{u}+1)\left|\tilde{x}^2 - x^2\right| + \bar{u}x^2 \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{|s_c - x^2|}{x^2} &\leq (\bar{u}+1)\left|\frac{\tilde{x}^2 - x^2}{x^2}\right| + \bar{u} \\ &= (\bar{u}+1)\,\frac{\overbrace{|\tilde{x}-x|}^{\leq \bar{u}|x|}\ \overbrace{|\tilde{x}+x|}^{(\tilde{x}-x)+2x}}{x^2} + \bar{u} \\ &\leq (\bar{u}+1)\,\frac{\bar{u}\,|x|\,(\overbrace{|\tilde{x}-x|}^{\leq \bar{u}|x|} + 2\,|x|)}{x^2} + \bar{u} \\ &\leq (\bar{u}+1)\,\bar{u}\,(\bar{u}+2) + \bar{u} \\ &= \bar{u}\left(\bar{u}^2 + 3\bar{u} + 3\right) \\ &\approx 3\bar{u} \end{aligned} \tag{4.2}$$

$$\Rightarrow \left|s_c - x^2\right| \leq \bar{u}\left(\bar{u}^2 + 3\bar{u} + 3\right) x^2$$

Näin ollen (4.1) pätee valitsemalla $\delta_f = 0$ ja $\varepsilon_f = \bar{u}\left(\bar{u}^2 + 3\bar{u} + 3\right) \approx 3\bar{u}$.

Käytetään esimerkkinä kolminumeroista aritmetiikkaa ja katkaisevaa pyöristystä, jolloin $\bar{u} = 10^{-2}$. Soveltamalla äskeistä algoritmia reaalilukuun $x = 2.129$ saadaan $\tilde{x} = 2.12$ ja $s_c = 4.49$. Koska $x^2 = 4.532641$, algoritmin lopputuloksen suhteellinen virhe on

$$\frac{|s_c - x^2|}{x^2} = 9,4 \cdot 10^{-3},$$

joka on selvästi pienempi kuin $3\bar{u}$.

4.2 Peräytyvä virheanalyysi

Numeerisen probleeman $\mathbf{d} \mapsto s(\mathbf{d})$ likimääräisen ratkaisun $s_c(\mathbf{d})$ tarkkuutta voidaan arvioida myös toisesta näkökulmasta. Usein nimittäin $s_c(\mathbf{d})$ on samalla funktion $\mathbf{d} \mapsto s(\mathbf{d})$ *tarkka* arvo jossakin vektorin $\mathbf{d}$ lähellä sijaitsevassa pisteessä $\mathbf{d}_N$: $s_c(\mathbf{d}) = s(\mathbf{d}_N)$. Likiarvon $s_c(\mathbf{d})$ "laatua" voidaan tällöin mitata tutkimalla etäisyyttä $\|\mathbf{d}_N - \mathbf{d}\|$. Jos on olemassa ei-negatiiviset vakiot δ_b ja ε_b siten, että

$$\|\mathbf{d}_N - \mathbf{d}\| \leq \delta_b + \varepsilon_b \|\mathbf{d}\| \tag{4.3}$$

lähtövektoreista $\mathbf{d}$ riippumatta, niin δ_b ja ε_b ovat numeerisen probleeman *peräytyvät virhevakiot*. Jos nämä vakiot ovat lisäksi pieniä, niin sanotaan, että algoritmi on *peräytyvästi (numeerisesti) stabiili*.

Olkoot x ja s_c kuten aiemmassa esimerkissä luvussa 4.1. Silloin on olemassa reaaliluku $\tilde{x} \approx x$ siten, että $s_c = \hat{x}^2$. Edelleen (4.2):n perusteella

$$\begin{aligned}
\frac{|\hat{x} - x|}{|x|} &= \frac{|\hat{x} - x|\,|\hat{x} + x|}{|x|\,|\hat{x} + x|} \\
&= \frac{|\hat{x}^2 - x^2|}{|x|\,|\hat{x} + x|} \\
&= \frac{|s_c - x^2|}{x^2} \frac{\overbrace{x}^{\leq 1}}{|\hat{x} + x|} \\
&\leq \frac{|s_c - x^2|}{x^2} \\
&\leq \bar{u}\left(\bar{u}^2 + 3\bar{u} + 3\right) \\
&\approx 3\bar{u},
\end{aligned}$$

joten (4.3) pätee valitsemalla $\delta_b = 0$ ja $\varepsilon_b \approx 3\bar{u}$.

Luvun 4.1 esimerkin luvuilla $x = 2.129$ ja $s_c = 4.49$ saadaan $\hat{x} = 2.1190$ ja $\dfrac{|\hat{x} - x|}{x} = 4.7 \cdot 10^{-3}$.

Esimerkki: Säännöllisen 2×2-matriisin kääntäminen on joskus pahanlaatuinen tehtävä. Olkoon esimerkiksi $\varepsilon \neq 0$ ja

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1 & 1 \end{pmatrix}, \quad \text{jolloin} \quad A^{-1} = \begin{pmatrix} -\frac{1}{\varepsilon} & 1+\frac{1}{\varepsilon} \\ \frac{1}{\varepsilon} & \frac{1}{\varepsilon} \end{pmatrix}.$$

Jos $|\varepsilon|$ on pieni, niin liukulukuaritmetiikkaa käytettäessä matriisin A^{-1} laskettu likiarvo voisi olla esimerkiksi

$$X = \begin{pmatrix} \frac{1}{\varepsilon} & \frac{1}{\varepsilon} \\ \frac{1}{\varepsilon} & -\frac{1}{\varepsilon} \end{pmatrix}.$$

Singulaarisena matriisina X ei kuitenkaan voi olla minkään matriisin käänteismatriisi. Vaikka siis matriisi X voi olla suhteellisen tarkka matriisin A^{-1} likiarvo, ei peräytyvä virheanalyysi ole tässä tapauksessa mahdollinen.

5 Lineaarisen systeemin häiriöalttius

5.1 Säännöllisen matriisin häiriöalttius

Tarkastellaan lineaarista systeemiä

$$A\mathbf{x} = \mathbf{b}, \tag{5.1}$$

missä A on säännöllinen neliömatriisi. Pieni muutos (häiriö) $\delta\mathbf{b}$ vektorissa $\mathbf{b}$ aiheuttaa vastaavan häiriön $\delta\mathbf{x}_b$ yhtälöryhmän ratkaisussa:

$$A\left(\mathbf{x} + \delta\mathbf{x}_b\right) = \mathbf{b} + \delta\mathbf{b} \tag{5.2}$$

Kun yhtälö (5.1) vähennetään puolittain yhtälöstä (5.2), saadaan

$$A\delta\mathbf{x}_b = \delta\mathbf{b}$$

ja edelleen

$$\delta\mathbf{x}_b = A^{-1}\delta\mathbf{b}$$

Vektorinormia ja indusoitua matriisinormia käyttäen tästä seuraa

$$\|\delta\mathbf{x}_b\| = \left\|A^{-1}\delta\mathbf{b}\right\| \leq \left\|A^{-1}\right\| \|\delta\mathbf{b}\| \tag{5.3}$$

Toisaalta, $A\mathbf{x} = \mathbf{b}$, joten $\|\mathbf{b}\| \leq \|A\| \, \|\mathbf{x}\|$ ja kertomalla puolittain (epä)yhtälön (5.3) kanssa saadaan

$$\|\mathbf{b}\| \, \|\delta\mathbf{x}_b\| \leq \left\|A^{-1}\right\| \|A\| \, \|\mathbf{x}\| \, \|\delta\mathbf{b}\|$$

Ratkaisun *suhteelliselle* häiriölle saadaan tästä yläraja

$$\frac{\|\delta\mathbf{x}_b\|}{\|\mathbf{x}\|} \leq \left\|A^{-1}\right\| \|A\| \, \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \tag{5.4}$$

Jos vektorin $\mathbf{b}$ asemasta häiritään yhtälön (5.1) kerroinmatriisia A, saadaan yhtälö

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}_A) = \mathbf{b}, \tag{5.5}$$

missä $\delta\mathbf{x}_A$ on kerroinmatriisin A häiriön δA aiheuttama muutos systeemin (5.1) ratkaisuun. Vähentämällä yhtälö (5.1) yhtälöstä (5.5), saadaan

$$A\delta\mathbf{x}_A + \delta A(\mathbf{x} + \delta\mathbf{x}_A) = \mathbf{0}$$

ja kertomalla matriisilla A^{-1} vasemmalta saadaan

$$\delta\mathbf{x}_A = A^{-1}\left[-\delta A(\mathbf{x} + \delta\mathbf{x}_A)\right]$$

Tästä seuraa jälleen yhteensopivia vektori- ja matriisinormeja käyttäen

$$\|\delta\mathbf{x}_A\| \leq \left\|A^{-1}\right\| \|\delta A\| \, \|\mathbf{x} + \delta\mathbf{x}_A\|$$

ja kaavaa (5.5) vastaava epäyhtälö

$$\frac{\|\delta\mathbf{x}_A\|}{\|\mathbf{x} + \delta\mathbf{x}_A\|} \leq \left\|A^{-1}\right\| \|A\| \, \frac{\|\delta A\|}{\|A\|} \tag{5.6}$$

Huomautus: kerroinmatriisin häiriö saattaa johtaa ratkeamattomaan systeemiin, jos $A + \delta A$ on singulaarinen; tällöin yhtälö (5.5) ei välttämättä toteudu millekään vektorille $\delta\mathbf{x}_A$. Jos kuitenkin $\|\delta A\|$ on kyllin pieni, niin $A + \delta A$ on säännöllinen.

Määritelmä 12. Säännöllisen matriisin A häiriöalttius $\varkappa(A)$ on

$$\varkappa(A) = \left\|A^{-1}\right\| \|A\| \tag{5.7}$$

Huomautus: Häiriöalttius $\varkappa(A)$ on siis positiivinen reaaliluku, mutta sen arvo riippuu käytetystä normista. Systeemin (5.1) vasemmalla tai oikealla puolella tapahtuvat häiriöt aiheuttavat systeemin ratkaisuun muutoksen, jonka

suhteelliselle suuruudelle saadaan kaavojen (5.5) ja (5.5) perusteella häiriöalttiudesta riippuva yläraja.

Vektorinormin indusoimalle matriisinormille yksikkömatriisin I normi on aina $\|I\| = 1$. Koska $I = A^{-1}A$ ja

$$1 = \|I\| = \left\|A^{-1}A\right\| \leq \left\|A^{-1}\right\| \|A\| = \varkappa(A),$$

nähdään, että $\varkappa(A) \geq 1$.

Jos matriisin häiriöalttius on kertalukua 1, matriisi on *hyvänlaatuinen*. *Pahanlaatuisen* matriisin häiriöalttius on $>> 1$.

Häiriöalttiudelle saadaan epäyhtälöistä (5.5) ja (5.6) alarajat

$$\varkappa(A) \geq \frac{\frac{\|\delta \mathbf{x}_b\|}{\|\mathbf{x}\|}}{\frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|}} \quad \text{ja} \quad \varkappa(A) \geq \frac{\frac{\|\delta \mathbf{x}_A\|}{\|\mathbf{x}+\delta \mathbf{x}_A\|}}{\frac{\|\delta A\|}{\|A\|}} \tag{5.8}$$

Olkoon esimerkiksi $A = \begin{pmatrix} 0.550 & 0.423 \\ 0.484 & 0.372 \end{pmatrix}$ ja $\mathbf{b} = \begin{pmatrix} 0.127 \\ 0.112 \end{pmatrix}$. Systeemin $A\mathbf{x} = \mathbf{b}$ ratkaisu on $\mathbf{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Olkoon $\delta\mathbf{b} = \begin{pmatrix} 0.00007 \\ 0.00028 \end{pmatrix}$. Silloin $\mathbf{b} + \delta\mathbf{b} = \begin{pmatrix} 0.12707 \\ 0.11228 \end{pmatrix}$. Tätä vektoria vastaava ratkaisu on $\mathbf{x} + \delta\mathbf{x}_b = \begin{pmatrix} 1.7 \\ -1.91 \end{pmatrix}$, jolloin $\delta\mathbf{x}_b = \begin{pmatrix} 0.7 \\ -0.91 \end{pmatrix}$.
Suhteelliset häiriöt maksiminormilla mitattuna ovat

$$\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \frac{0.00028}{0.127} \approx 2.2 \cdot 10^{-3} \text{ ja } \frac{\|\delta\mathbf{x}_b\|}{\|\mathbf{x}\|} = 0.91$$

Huomataan, että ratkaisun suhteellinen häiriö on yli 400 kertaa suurempi, kuin suhteellinen häiriö systeemin oikealla puolella. Näin ollen $\varkappa(A) \geq 400$ (vrt. (5.8)).

Matriisin A pahanlaatuisuus voidaan todeta myös tarkastelemalla systeemiä $(A + \delta A)(\mathbf{x} + \delta\mathbf{x}_A) = \mathbf{b}$.
Olkoon esimerkiksi $A + \delta A = \begin{pmatrix} 0.550 & 0.423 \\ 0.483 & 0.372 \end{pmatrix}$, jolloin

$$\frac{\|\delta A\|}{\|A\|} \approx 0.01.$$

Systeemin ratkaisu on $\mathbf{x}+\delta\mathbf{x}_A = \begin{pmatrix} -0.4536 \\ 0.8900 \end{pmatrix}$, siis $\|\delta\mathbf{x}_A\| = 1.89$ ja ratkaisun suhteellisen häiriö on lähes 2000 kertaa matriisin A suhteellinen häiriö. $A^{-1} = \begin{pmatrix} -2818.2 & 3204.5 \\ 3666.7 & -4166.7 \end{pmatrix}$, jolloin $\|A^{-1}\| = 7833.4$ ja $\|A\| = 0.973$. ∞-normilla mitattuna $\varkappa(A) = \|A^{-1}\|\,\|A\| = 7833.3 \cdot 0.973 \approx 7622$.

Pian nähdään, että A on pahanlaatuinen, jos on olemassa kaksi samanpituista vektoria $\mathbf{x}_1$ ja $\mathbf{x}_2$ siten, että

$$\|A\mathbf{x}_1\| << \|A\mathbf{x}_2\|.$$

Olkoon esimerkiksi $A = \begin{pmatrix} 10^4 & 0 \\ 0 & 10^{-4} \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ja $\hat{\mathbf{x}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Silloin $A\mathbf{x} = \begin{pmatrix} 10^4 \\ 0 \end{pmatrix}$ ja $A\hat{\mathbf{x}} = \begin{pmatrix} 0 \\ 10^{-4} \end{pmatrix}$. Koska $\|A\mathbf{x}\| = 10^4 >> 10^{-4} = \|A\hat{\mathbf{x}}\|$, on A pahanlaatuinen. Maksiminormilla mitattuna $\varkappa(A) = 10^8$.

Pahanlaatuinen matriisi on ”lähes singulaarinen”, mutta sen determinantti ei ole pieni.

Edellisen esimerkin pahanlaatuisen matriisin determinantti on yksi. Toisaalta $n \times n$-diagonaalimatriisin $A = \mathrm{diag}(10^{-10})$ determinantti on $\det(A) = 10^{(-10)n} \approx 0$. Silti A on mitä hyvälaatuisin: $\varkappa(A) = 1$. Determinanttia ei siis voi pitää häiriöalttiuden mittana yleensä.

Hyvälaatuisen matriisin ominaisarvojen määrääminen voi olla pahanlaatuinen tehtävä: esimerkiksi yksikkömatriisin I_n karakteristinen polynomi on $(1-\lambda)^n$; sen nollakohtien määrääminen on suurilla n:n arvoilla pahanlaatuinen tehtävä.

5.2 Häiriöalttius ja singulaariarvohajotelma

Säännöllisen $n \times n$-matriisin A singulaariarvohajotelma on muotoa

$$A = USV^T,$$

missä U ja V ovat ortogonaalisia $n \times n$-matriiseja, $S = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_n)$ ja $\sigma_i > 0$ kaikille $i = 1, \ldots, n$. Voidaan olettaa, että $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$, jolloin σ_1 on matriisin A suurin singulaariarvo. Ehto $\sigma_i > 0$ seuraa siitä, että säännöllisen matriisin kaikki singulaariarvot ovat positiivisia.

Lause 13. $\sigma_1 = \|A\|$

Todistus. Määritelmän mukaan

$$\|A\|_2 = \max_{x \neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \tag{5.9}$$

$$\|A\mathbf{x}\|_2^2 = (A\mathbf{x})^T (A\mathbf{x}) = \mathbf{x}A^T A\mathbf{x} \tag{5.10}$$

Lausekkeen (5.10) tutkimista varten todetaan aluksi, että matriisin V sarakkeiden joukko $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ on avaruuden $\mathbb{R}^n$ ortonormaali kanta. Näin ollen jokaista n-vektoria x vastaa yksi n-vektori $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ siten, että

$$\mathbf{x} = \alpha_1 \mathbf{v}_1 + \ldots + \alpha_n \mathbf{v}_n = V\alpha$$

Koska V on ortogonaalinen, tästä seuraa, että

$$\mathbf{x}^T\mathbf{x} = \alpha^T V^T V\alpha = \alpha^T\alpha = \sum_{i=1}^{n} \alpha_i^2 \tag{5.11}$$

Edelleen

$$A\mathbf{x} = USV^T (V\alpha) = US \underbrace{\left(V^T V\right)}_{I} \alpha = US\alpha = \sum_{i=1}^{n} (\sigma_i \alpha_i)\, \mathbf{u}_i, \tag{5.12}$$

sillä $S = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Erikoistapauksessa $\alpha = \mathbf{e}_i$ (jolloin $\mathbf{x} = \mathbf{v}_i$). Yhtälö (5.12) yksinkertaistuu muotoon

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i \;\; (1 \leq i \leq n) \tag{5.13}$$

Huomaa (spektraali- ja singulaariarvohajotelmat):

$$\begin{aligned} & A = U\Lambda U^T \\ \Leftrightarrow\; & AU = U\Lambda \\ \Leftrightarrow\; & A\mathbf{u}_i = \lambda_i \mathbf{u}_i \end{aligned}$$

$$\begin{aligned} & A = USV^T \\ \Leftrightarrow\; & AV = US \\ \Leftrightarrow\; & A\mathbf{v}_i = \sigma_i \mathbf{v}_i \end{aligned}$$

Lauseke (5.10) voidaan yhtälön (5.13) perusteella kirjoittaa

$$\mathbf{x}^T A^T A\mathbf{x} = \alpha S^T \underbrace{U^T U}_{I} S\alpha = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2$$

Näin ollen, kun huomioidaan (5.11),

$$\frac{\|A\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{\sum_{i=1}^{n} \alpha_i^2 \sigma_i^2}{\sum_{i=1}^{n} \alpha_i^2} \tag{5.14}$$

on lukujen σ_i^2 painotettu keskiarvo. Sen suurin mahdollinen arvo σ_1^2 saavutetaan tapauksessa $\alpha_1 \neq 0, \alpha_2 = \ldots = \alpha_n = 0$. Siis

$$\|A\|_2 = \max_{x \neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_1$$

□

Matriisin A suurin singulaariarvo antaa siis ylärajan vektorin pituuden suhteelliselle kasvulle kuvauksessa $\mathbf{x} \mapsto A\mathbf{x}$; tämä yläraja saavutetaan aina, kun $\mathbf{x}$ on vektorin $\mathbf{v}_1$ suuntainen. Jos taas $\mathbf{x}$ on vektorin $\mathbf{v}_n$ suuntainen, jolloin $\alpha_1 = \ldots = \alpha_{n-1} = 0$, painotettu keskiarvo (5.14) saa pienimmän mahdollisen arvonsa σ_n^2. Säännöllisen matriisin A pienin singulaariarvo σ_n ilmoittaa siis vektorin euklidisen pituuden suhteellisen *vähimmäiskasvun* kuvauksessa $\mathbf{x} \mapsto A\mathbf{x}$.

Että voitaisiin tutkia matriisin A häiriöalttiutta, todetaan aluksi, että yhtälöstä $A = USV^T$ seuraa heti käänteismatriisin A^{-1} singulaariarvohajotelma:

$$A^{-1} = VS^{-1}U^T \tag{5.15}$$

Erityisesti matriisin A^{-1} singulaariarvot ovat matriisin A singulaariarvojen käänteisluvut $\sigma_1^{-1}, \ldots, \sigma_n^{-1}$. Näin ollen $\|A^{-1}\|_2 = \sigma_n^{-1}$, ja matriisin A^{-1} pienin singulaariarvo on σ_1^{-1}. Matriisin A häiriöalttius 2-normilla mitattuna on siis

$$\varkappa(A) = \left\|A^{-1}\right\|_2 \|A\|_2 = \sigma_n^{-1}\sigma_1 = \frac{\sigma_1}{\sigma_n} \tag{5.16}$$

Häiriöalttiuden ohella singulaariarvohajotelma antaa tietoa kuvauksesta $\delta\mathbf{b} \to \delta\mathbf{x}_b$ yhtälössä (5.2)

$$\begin{aligned} A(\mathbf{x} + \delta\mathbf{x}_b) &= \mathbf{b} + \delta\mathbf{b} \\ A(\delta\mathbf{x}_b) &= \delta\mathbf{b} \\ (5.14): \ \sigma_n^2 \leq \frac{\|A(\delta\mathbf{x}_b)\|_2^2}{\|\delta\mathbf{x}_b\|_2^2} &= \frac{\|\delta\mathbf{b}\|_2^2}{\|\delta\mathbf{x}_b\|_2^2} \leq \sigma_1^2 \\ \frac{1}{\sigma_1} \leq \frac{\|\delta\mathbf{x}_b\|}{\|\delta\mathbf{b}\|} &\leq \frac{1}{\sigma_n} \end{aligned}$$

Jos matriisi A on hyvänlaatuinen, jolloin kaikki singulaariarvot ovat samaa suuruusluokkaa, on vektorien $\delta\mathbf{x}_b$ ja $\delta\mathbf{b}$ euklidisten pituuksien suhde samaa kertaluokkaa riippumatta häiriön $\delta\mathbf{b}$ suunnasta.

Suhde $\dfrac{\|\delta\mathbf{x}_b\|_2}{\|\delta\mathbf{b}\|_2}$ saa suurimman arvonsa $\frac{1}{\sigma_n}$ vektorin $\delta\mathbf{b}$ ollessa vektorin $\mathbf{u}_n$ suuntainen, sillä kaavan (5.13) perusteella pätee tällöin

$$\delta\mathbf{b} = \|\delta\mathbf{b}\|_2 \mathbf{u}_n \overset{A\mathbf{v}_n=\sigma_n\mathbf{u}_n}{=} \|\delta\mathbf{b}\|_2 \frac{1}{\sigma_n} A\mathbf{v}_n = A\left(\frac{\|\delta\mathbf{b}\|_2}{\sigma_n}\mathbf{v}_n\right),$$

ja siis

$$\delta\mathbf{x}_b = \frac{\|\delta\mathbf{b}\|_2}{\sigma_n}\mathbf{v}_n \Rightarrow \frac{\|\delta\mathbf{x}_b\|_2}{\|\delta\mathbf{b}\|_2} = \frac{\|\mathbf{v}_n\|_2}{\sigma_n} = \frac{1}{\sigma_n}$$

Suhteen $\dfrac{\|\delta\mathbf{x}_b\|_2}{\|\delta\mathbf{b}\|_2}$ pienin mahdollinen arvo $\frac{1}{\sigma_1}$ saavutetaan vastaavasti vektorin $\delta\mathbf{b}$ ollessa vektorin $\mathbf{u}_1$ suuntainen.

Luvun 5.1 esimerkin matriisilla $A = \begin{pmatrix} 0.550 & 0.423 \\ 0.484 & 0.372 \end{pmatrix}$ on singulaarihajotelma $A = USV^T$, missä neljän numeron tarkkuudella $U = \begin{pmatrix} 0.7508 & 0.6605 \\ 0.6605 & -0.7508 \end{pmatrix}$, $V = \begin{pmatrix} 0.7928 & -0.6095 \\ 0.6095 & 0.7928 \end{pmatrix}$ ja $S = \operatorname{diag}(0.9242,\ 1.428 \cdot 10^{-4})$. Matriisin A häiriöalttius 2-normilla mitattuna saadaan kaavasta (5.16):

$$\varkappa(A) = \frac{\sigma_1}{\sigma_2} = 6470$$

Suuruusluokka on sama kuin ∞-normilla mitattuna (5.1). Yhtälön (5.13) mukaan voidaan kirjoittaa eksplisiittisesti

$$A\mathbf{v}_1 = \begin{pmatrix} 0.550 & 0.423 \\ 0.484 & 0.372 \end{pmatrix} \begin{pmatrix} 0.7928 \\ 0.6095 \end{pmatrix} = \begin{pmatrix} 0.6939 \\ 0.6104 \end{pmatrix} = \sigma_1\mathbf{u}_1$$

$$A\mathbf{v}_2 = \begin{pmatrix} 0.550 & 0.423 \\ 0.484 & 0.472 \end{pmatrix} \begin{pmatrix} -0.6095 \\ 0.7928 \end{pmatrix} = 10^{-3} \begin{pmatrix} 0.0943 \\ -0.1072 \end{pmatrix} = \sigma_2\mathbf{u}_2$$

Oikeanpuoleisten vektorien 2-normit ovat vastaavasti

$$\|\sigma_1\mathbf{u}_1\|_2 = \sigma_1 = 0.9242 \quad \text{ja} \quad \|\sigma_2\mathbf{u}_2\|_2 = \sigma_2 = 1.428 \cdot 10^{-4}$$

Merkitsevien numeroiden kumoutumisen takia vektoria $A\mathbf{v}_2$ laskettaessa on käytettävä vektorille $\mathbf{v}_2$ tarkempaa likiarvoa, sillä nelinumeroista likiarvoa

käytettäessä saadaan täysin virheellinen tulos:

$$\begin{pmatrix} 0.550 & 0.423 \\ 0.484 & 0.372 \end{pmatrix} \begin{pmatrix} -0.6095 \\ 0.7928 \end{pmatrix} = 10^{-3} \begin{pmatrix} 0.1294 \\ -0.0764 \end{pmatrix}$$

Osa III

Lineaariset systeemit

1 Johdanto

Tarkastellaan lineaarista yhtälöryhmää

$$A\mathbf{x} = \mathbf{b}, \tag{1.1}$$

jonka kerroinmatriisi A on säännöllinen. Tällainen systeemi on aina ratkeava.

Olkoon esimerkiksi

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix} \text{ ja } \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 6 \end{pmatrix}$$

Tällöin (1.1) voidaan kirjoittaa

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ 2x_1 + 3x_2 + x_3 = 2 \\ 3x_1 - x_2 - x_3 = 6 \end{cases} \tag{1.2}$$

Systeemi (1.2) ratkaistaan palauttamalla se niin sanottuun *kolmiomuotoon*, joka on helposti ratkaistavissa. Kolmiomuotoon päästään niin sanotulla *eliminointimenetelmällä*, jossa systemaattisesti eliminoidaan tuntemattomia yhtälöryhmästä (1.2) esimerkiksi seuraavasti:

- eliminoidaan x_1 toisesta ja kolmannesta yhtälöstä:

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ x_2 - 3x_3 = -4 \\ -4x_2 - 7x_3 = -3 \end{cases} \tag{1.3}$$

- eliminoidaan x_2 kolmannesta yhtälöstä:

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ x_2 - 3x_3 = -4 \\ -19x_3 = -19 \end{cases} \tag{1.4}$$

Systeemit (1.2), (1.3) ja (1.4) ovat yhtäpitävät. Ratkaisu lasketaan helposti yhtälöryhmän (1.4) perusteella niin sanotulla *takaisinsijoituksella*:

$$\begin{cases} -19x_3 = -19 \Rightarrow x_3 = 1 \\ x_2 - 3x_3 = 4 \Rightarrow x_2 = 3x_3 - 4 = 3 \cdot 1 - 4 = -1 \\ x_1 + x_2 + 2x_3 = 3 \Rightarrow x_1 = -x_2 - 2x_3 + 3 = 1 - 2 \cdot 1 + 3 = 2 \end{cases}$$

Siis $\mathbf{x} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$.

Myöhemmin nähdään, että siirtyminen kolmiomuotoon (1.4) voidaan toteuttaa kertomalla (1.1) puolittain sopivilla alkeismatriiseilla.

2 Kolmiomuotoiset matriisit

2.1 Permutaatiomatriiiseista

$n \times n$-matriisi on *permutaatiomatriisi*, jos sillä on järjestystä vaille samat sarakkeet (tai rivit) kuin yksikkömatriisilla I_n. Esimerkiksi

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{2.1}$$

Permutaatiomatriisi voidaan karakterisoida antamalla sen *rivijärjestys* tai *sarakejärjestys*. $n \times n$-permutaatiomatriisin P rivijärjestys $\{i_1, \ldots, i_n\}$ ilmoittaa, että matriisin P k:s vaakarivi on sama kuin matriisin I_n vaakarivi i_k. Matriisin P sarakejärjestys $\{j_1, \ldots, j_n\}$ ilmoittaa vastaavasti, että matriisin P k:s sarake on matriisin I_n sarake j_k. Esimerkiksi matriisin (2.1) rivijärjestys on $\{3, 1, 2\}$ ja sarakejärjestys on $\{2, 3, 1\}$.

Olkoon $A = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \mathbf{a}_3' \end{pmatrix} = (\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_3)$. Tällöin $PA = \begin{pmatrix} \mathbf{a}_3' \\ \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix}$ ja $AP = (\mathbf{a}_2\ \mathbf{a}_3\ \mathbf{a}_1)$

Yleisemmin pätee: kerrottaessa matriisia A permutaatiomatriisilla P vasemmalta matriisin A rivien järjestys vaihtuu matriisin P rivijärjestyksen mukaiseksi. Vastaavasti, kertomalla oikealta matriisilla P matriisin A sarakkeiden järjetys vaihtuu matriisin P sarakejärjestyksen mukaiseksi.

Määritelmä 14. Neliömatriisi T on *kolmioituva*, jos on olemassa permutaatiomatriisit P_l ja P_r siten, että $L = P_l T P_r$ on alakolmiomatriisi.

Esimerkki: Jokainen 3×3-yläkolmiomatriisi on kolmioituva - valitsemalla $P_l = P_r = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ kääntyy sekä rivien että sarakkeiden järjestys, ja

$$P_l \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ & u_{22} & u_{23} \\ & & u_{33} \end{pmatrix} P_r = \begin{pmatrix} u_{33} & & \\ u_{23} & u_{22} & \\ u_{13} & u_{12} & u_{11} \end{pmatrix}$$

Muita kolmioituvia matriisityyppejä ovat

$$T_1 = \begin{pmatrix} & & \times \\ & \times & \times \\ \times & \times & \times \end{pmatrix} \quad \text{ja} \quad T_2 = \begin{pmatrix} \times & \times & \times \\ \times & \times & \\ \times & & \end{pmatrix} \tag{2.2}$$

sekä

$$T_3 = \begin{pmatrix} \times & \times & \\ \times & & \\ \times & \times & \times \end{pmatrix} \quad \text{ja} \quad T_4 = \begin{pmatrix} & \times & \\ \times & \times & \times \\ & \times & \times \end{pmatrix} \tag{2.3}$$

Jos systeemin (1.1) kerroinmatriisi on kolmioituva, niin systeemi on helppo ratkaista takaisinsijoituksella kuten yhtälöryhmässä (1.4). Tällaista systeemiä tutkittaessa voidaan olettaa, että kerroinmatriisi on ylä- tai alakolmiomatriisi.

2.2 Kolmiomuotoisen systeemin ratkaiseminen

Tarkastellaan systeemiä

$$L\mathbf{x} = \mathbf{b},$$

missä L on säännöllinen $n \times n$ alakolmiomatriisi. Siis

$$\begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{2.4}$$

Koska L on säännöllinen, sen diagonaaliset alkiot ovat $l_{ii} \neq 0$.

Systeemin (2.4) ylimmästä yhtälöstä $l_{11}x_1 = b_1$ seuraa

$$x_1 = \frac{b_1}{l_{11}} \tag{2.5}$$

Toisesta yhtälöstä $l_{21}x_1 + l_{22}x_2 = b_2$ seuraa vastaavasti

$$x_2 = \frac{1}{l_{22}} (b_2 - l_{21}x_1) \tag{2.6}$$

ja niin edelleen. Yleisesti, arvoilla $1 < k \leq n$:

$$x_k = \frac{1}{l_{kk}} (b_k - l_{k1}x_1 - l_{k2}x_2 - \ldots - l_{k,k-1}x_{k-1}) \tag{2.7}$$

Ratkaisuun tarvittavien laskutoimitusten lukumäärä:

x_1: yksi jakolasku

x_2: yksi jakolasku ja yksi floppi (eli yhteenlasku ja kertolasku)

x_k: yksi jakolasku ja $k-1$ floppia

Yhteensä systeemin $L\mathbf{x} = \mathbf{b}$ ratkaisemiseen riittää n jakolaskua ja $1 + 2 + \ldots + (n-1) = \frac{1}{2}n\,(n-1)$ floppia. Suurilla n:n arvoilla toistettavien floppien lukumäärä on suuruusluokkaa $\frac{1}{2}n^2$. Vastaava määrä operaatioita tarvitaan esimerkiksi suoritettaessa pelkkä kertolasku $L\mathbf{y}$, missä $\mathbf{y}$ on n-vektori. Lausekkeen $\mathbf{c} = L\mathbf{y}$ komponentit ovat nimittäin

$$c_1 = l_{11}y_1$$

$$c_2 = l_{21}y_1 + l_{22}y_2$$

$$\vdots$$

$$c_n = l_{n1}y_1 + l_{n2}y_2 + \ldots + l_{nn}y_n$$

ja esimerkiksi komponentin c_n laskemiseen tarvitaan n kertolaskua ja $n-1$ yhteenlaskua.

3 Gaussin eliminointimenetelmä

3.1 Palauttaminen kolmiomuotoon

Seuraavassa näytetään, että systeemi $A\mathbf{x} = \mathbf{b}$ voidaan palattaa kolmiomuotoon kertomalla vasemmalta sopivalla säännöllisellä matriisilla M siten, että MA on yläkolmiomatriisi. Matriisi M voidaan esittää äärellisen monen alkeismatriisin tulona.

Muotoa $E = I - \mathbf{u}\mathbf{v}^T$ olevan alkeismatriisin ja vektorin $\mathbf{a}$ tulo on

$$E\mathbf{a} = \left(I - \mathbf{u}\mathbf{v}^T\right)\mathbf{a} = \mathbf{a} - \mathbf{u}\underbrace{\mathbf{v}^T\mathbf{a}}_{\text{skalaari}} = \mathbf{a} - \left(\mathbf{v}^T\mathbf{a}\right)\mathbf{u} \tag{3.1}$$

Vektori $E\mathbf{a}$ saadaan siis vähentämällä vektorista $\mathbf{a}$ eräs vektorin $\mathbf{u}$ suuntainen vektori.

Ratkaistaessa systeemiä $A\mathbf{x} = \mathbf{b}$ pyritään ensimmäisessä eliminointivaiheessa saamaan kerroinmatriisiksi sellainen matriisi, jonka ensimmäinen sarake on $(a_{11}\ 0\ \ldots\ 0)^T$. Tämä saadaan aikaan kertomalla vasemmalta alkeismatriisilla $M_1 = I - \mathbf{u}_1\mathbf{v}_1^T$, jossa

$$\mathbf{u}_1 = \begin{pmatrix} 0 \\ \frac{a_{21}}{a_{11}} \\ \vdots \\ \frac{a_{n1}}{a_{11}} \end{pmatrix} \text{ ja } \mathbf{v}_1 = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.2}$$

edellyttäen, että $a_{11} \neq 0$. Lukua a_{11} kutsutaan *tukialkioksi*. Tapauksessa $a_{11} \neq 0$ merkitään

$$m_{i1} = \frac{a_{i1}}{a_{11}} \quad (2 \leq i \leq n) \tag{3.3}$$

Tällöin $M_1 = I - \mathbf{u}_1 \mathbf{v}_1^T$ on alakolmiomatriisi:

$$M_1 = I - \mathbf{u}_1 \mathbf{v}_1^T = \begin{pmatrix} 1 & 0 & & & 0 \\ -m_{21} & 1 & & & \\ -m_{31} & 0 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -m_{n1} & 0 & 0 & \dots & 1 \end{pmatrix} \tag{3.4}$$

Matriisilla M_1 on siis ensimmäistä saraketta lukuunottamatta samat sarakkeet kuin yksikkömatriisilla I_n.

Ensimmäisen eliminointivaiheen jälkeen kerroinmatriisi on siis $A^{(2)} = M_1 A$, jonka ensimmäinen sarake on $(a_{11}\ 0\ \dots\ 0)^T$ ja ensimmäinen vaakarivi $\mathbf{a}_1^{(2)'} = \mathbf{a}_1'$. Matriisin $A^{(2)}$ j:s sarake on

$$a_j^{(2)} = M_1 \mathbf{a}_j = \left(I - \mathbf{u}_1 \mathbf{v}_1^T\right) \mathbf{a}_j = \mathbf{a}_j - \underbrace{\left(\mathbf{v}_1^T \mathbf{a}_j\right)}_{\mathbf{e}_1^T \mathbf{a}_j = a_{1j}} \mathbf{u}_1 = \mathbf{a}_j - a_{1j}\mathbf{u}_j$$

$$a_{1j}^{(2)} = a_{1j} - a_{1j} \underbrace{u_{11}}_{0} = a_{1j}$$

Matriisin $A^{(2)}$ ij:s alkio on

$$a_{ij}^{(2)} = a_{ij} - a_{1j} \left(u_1\right)_i = a_{ij} - m_{i1} a_{1j} \ \ (2 \leq i, j \leq n) \tag{3.5}$$

Näin ollen matriisin $A^{(2)}$ i:s vaakarivi on matriisin A ensimmäisen ja i:nnen vaakarivin lineaarikombinaatio.

$$M_1 A = A^{(2)} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' - m_{21}\mathbf{a}_1' \\ \vdots \\ \mathbf{a}_n' - m_{n1}\mathbf{a}_1' \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix} \tag{3.6}$$

Olkoon esimerkiksi (vrt. (1.2)) $A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix}$. Tukialkio $a_{11} = 1$ ja kaavan (3.3) perusteella $m_{21} = \frac{a_{21}}{a_{11}} = 2$ ja $m_{31} = \frac{a_{31}}{a_{11}} = 3$. Siis

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}$$

$$M_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -3 \\ 0 & -4 & -7 \end{pmatrix}$$

Toinen eliminointivaihe toteutetaan kertomalla äsken saatu matriisi alkeismatriisilla $M_2 = I - \mathbf{u}_2 \mathbf{e}_2^T$, missä $\mathbf{u}_2 = (0, 0, m_{32} \dots m_{n2})^T$ ja $m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}}$ arvoilla $3 \leq i \leq n$.

Yleisesti k:nnessa eliminointivaiheessa käytetään alkeismatriisia

$$M_k = I - \mathbf{u}_k \mathbf{e}_k^T, \text{ missä } (u_k)_i = \begin{cases} 0, & \text{kun } 1 \leq i \leq k \\ m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, & \text{kun } k+1 \leq i \leq n \end{cases} \tag{3.7}$$

Menettely onnistuu vain, jos tukialkio $a_{kk} \neq 0$. Äskeisessä esimerkissä

$$m_{32} = \frac{-4}{1} = -4 \text{ ja } M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix}.$$

$$M_2 A^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -3 \\ 0 & -4 & -7 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -3 \\ 0 & 0 & -19 \end{pmatrix} = U$$

Yleisesti: jos kaikki tukialkiot ovat nollasta eroavia, niin $n-1$ eliminointivaiheen jälkeen saatava matriisi

$$U = M_{n-1} \cdots M_2 M_1 A \tag{3.8}$$

on yläkolmiomatriisi.

Alakolmiomatriisien M_k tulo $M = M_{n-1} \cdots M_2 M_1$ on alakolmiomatriisi, jonka diagonaalialkiot ovat ykkösiä.

Gaussin eliminointimenetelmässä tarvittavien aritmeettisten laskutoimitusten lukumääristä:

- 1. eliminointivaiheen kertoimien $m_{21}, \dots, m_{n1}$ laskemiseen tarvitaan $n-1$ jakolaskua; matriisin $A^{(2)}$ jokaista vaakariviä muodostettaessa suoritetaan $n-1$ kerto- ja yhteenlaskutoimitusta - yhteensä matriisin $A^{(2)}$ laskeminen vaatii siis $(n-1)^2$ floppia.
- 2. eliminointivaihe: $n-2$ jakolaskua ja $(n-2)^2$ floppia.

Eliminoinnin kaikki $n-1$ vaihetta edellyttävät yhteensä $(n-1)^2 + (n-2)^2 + \dots + 1 = \frac{1}{6}(n-1)(2n-1) \approx \frac{1}{3}n^3$ floppia ja lisäksi noin $\frac{1}{2}n^2$ jakolaskua. Aritmeettisten laskutoimitusten lukumäärän kertaluku on siis $\frac{1}{3}n^3$ floppia.

3.2 LU-hajotelma

$n \times n$-matriisin A LU-hajotelma on muotoa

$$A = LU, \tag{3.9}$$

missä L on alakolmiomatriisi, jonka diagonaalialkiot ovat ykkösiä, ja U yläkolmiomatriisi.

Jos hajotelma (3.9) on olemassa ja U on säännöllinen, matriisien U ja L alkiot voidaan määrittää seuraavasti:

Ensiksikin matriisien A ja U ensimmäiset vaakarivit ovat identtiset, sillä yhtälön (3.9) perusteella

$$\mathbf{a}_1' = \mathbf{l}_1' U = \mathbf{e}_1^T U = \mathbf{u}_1'.$$

Edelleen

$$\mathbf{a}_2' = \mathbf{l}_2' U = l_{21}\mathbf{u}_1' + l_{22}\mathbf{u}_2' = l_{21}\mathbf{u}_1' + \mathbf{u}_2', \tag{3.10}$$

koska L on alakolmiomatriisi ja $l_{22} = 1$. Erityisesti

$$a_{21} = l_{21}u_{11} + u_{21} = l_{21}u_{11}$$

ja koska U on säännöllinen, niin $u_{11} \neq 0$. Siis

$$l_{21} = \frac{a_{21}}{u_{11}} = \frac{a_{21}}{a_{11}},$$

ja matriisin U toinen vaakarivi voidaan määrätä yhtälön (3.10) perusteella:

$$\mathbf{u}_2' = \mathbf{a}_2' - l_{21}\mathbf{u}_1' = \mathbf{a}_2' - \left(\frac{a_{21}}{a_{11}}\right)\mathbf{u}_1'.$$

Matriisien L ja U kolmannet vaakarivit voidaan määrätä samaan tapaan ja menettelyä jatkamalla saadaan selville loputkin vaakarivit.

Matriisien L ja U alkiot voidaan määrätä myös sarakkeittain, mikäli edelleen oletetaan, että matriisi U on säännöllinen. Yhtälöstä (3.9) seuraa aluksi

$$\mathbf{a}_1 = L\mathbf{u}_1 = u_{11}\mathbf{l}_1, \tag{3.11}$$

sillä $u_{21} = \ldots = u_{n1} = 0$. Näin ollen matriisin L ensimmäinen sarake on

$$\mathbf{l}_1 = \left(\frac{1}{u_{11}}\right)\mathbf{a}_1.$$

Koska $l_{11} = 1$, yhtälöstä (3.11) seuraa

$$a_{11} = u_{11} \quad \text{ja} \quad l_{k1} = \frac{a_{k1}}{a_{11}} \ \ (2 \leq k \leq n) \tag{3.12}$$

Matriisien U ja L toiset sarakkeet määrätään yhtälön $\mathbf{a}_2 = L\mathbf{u}_2$ perusteella: koska sarakkeen $\mathbf{u}_2$ ainoat mahdollisesti nollasta eroavat komponentit ovat u_{12} ja u_{22}, saadaan

$$\mathbf{a}_2 = u_{12}\mathbf{l}_1 + u_{22}\mathbf{l}_2. \tag{3.13}$$

Ensimmäisiä komponentteja tarkastelemalla tästä seuraa

$$a_{12} = u_{12}l_{11} + u_{22}\underbrace{l_{12}}_{=0} = 1 \cdot u_{12} = u_{12}$$

ja toisille komponenteille pätee vastaavasti

$$a_{22} = u_{12}l_{21} + u_{22}l_{22} = u_{12}l_{21} + u_{22}.$$

Näin ollen $u_{22} = a_{22} - u_{12}l_{21}$, ja koska l_{21} tiedetään (3.12):n perusteella, matriisin U toisen sarakkeen alkiot u_{21} ja u_{22} on määrätty. Matriisin L toisen sarakkeen alkiot saadaan tämän jälkeen tarkastelemalla yhtälön (3.13) jäljellä olevia komponentteja $3, 4, \ldots, n$.

Näin jatkamalla voidaan matriisien L ja U alkiot määrätä sarakkeittain, mikäli kaikki matriisin U diagonaaliset alkiot ovat nollasta eroavia. Tällöin matriisin A LU-hajotelma on siis yksikäsitteisesti määrätty.

Olkoon esimerkiksi $A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix}$, jonka LU-hajotelma löydetään kuten edellä; joko riveittäin tai sarakkeittain:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -4 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -3 \\ 0 & 0 & -19 \end{pmatrix}$$

LU-hajotelmalla (3.9) on läheinen yhteys hajotelmaan (3.8). Jälkimmäisenä mainitussa esiintyvien alkeismatriisien tulo $M = M_{n-1} \cdots M_2 M_1$ on alakolmiomatriisi, jonka diagonaalialkiot ovat ykkösiä. Koska (3.8):n mukaisesti $MA = U$ on yläkolmiomatriisi, niin $A = M^{-1}U$ on matriisin A LU-hajotelma, sillä myös matriisi M^{-1} on alakolmiomatriisi, jonka diagonaalialkiot ovat ykkösiä.

LU-hajotelman matriisi $L = M^{-1} = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$ voidaan esittää eksplisiittisesti kaavan (3.7) vakioiden m_{ik} avulla:

$$L = \begin{pmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ m_{n1} & m_{n2} & m_{n3} & \cdots & 1 \end{pmatrix} \tag{3.14}$$

Näin ollen matriisien L ja U alkiot voidaan määrätä myös eliminointialgoritmin yhteydessä, jolloin

$$\begin{aligned} l_{ik} = m_{ik} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \quad (k+1 \leq i \leq n) \\ u_{kj} &= a_{kj}^{(k)} \quad (k \leq j \leq n)\,. \end{aligned}$$

Matriisi U on eliminointialgoritmin tuloksena syntyvä matriisi (3.8).

3.3 Kolmiohajotelmat ja systeemi $A\mathbf{x} = \mathbf{b}$

Hajotelmia (3.8) ja (3.9) voidaan soveltaa systeemin $A\mathbf{x} = \mathbf{b}$ ratkaisemiseen seuraavasti: hajotelmassa (3.8) esiintyvä alakolmiomatriisi $M = M_{n-1} \cdots M_2 M_1$ on säännöllinen ja toteuttaa ehdon $MA = U$. Näin ollen $A\mathbf{x} = \mathbf{b}$ täsmälleen silloin, kun

$$MA\mathbf{x} = U\mathbf{x} = M\mathbf{b}. \tag{3.15}$$

Systeemin $A\mathbf{x} = \mathbf{b}$ asemasta riittää ratkaista kolmiomuotoinen systeemi

$$U\mathbf{x} = M\mathbf{b}.$$

Käytettäessä LU-hajotelmaa (3.9) voidaan systeemi $A\mathbf{x} = \mathbf{b}$ korvata systeemillä $LU\mathbf{x} = L\,(U\mathbf{x}) = \mathbf{b}$, joka on yhtäpitävä yhtälöparin

$$\begin{cases} L\mathbf{y} = \mathbf{b} \\ U\mathbf{x} = \mathbf{y} \end{cases} \tag{3.16}$$

kanssa. Molemmat yhtälöparin (3.16) systeemeistä ovat kolmiomuotoisia.

Koska $MA = U$ ja $A = LU$, niin $M\,(LU) = U$ ja siis $M = L^{-1}$. Näin ollen yhtälöparin (3.16) ylemmän systeemin $L\mathbf{y} = \mathbf{b}$ ratkaisu on vektori $M\mathbf{b}$ systeemin (3.15) oikealla puolella.

LU-hajotelman muodostamiseen tarvittavien floppien lukumäärä on kertalukua $\frac{1}{3}n^3$. Kolmiomuotoisen systeemin ratkaisemiseen kuluu lähes $\frac{1}{2}n^2$ floppia, joten kaiken kaikkiaan systeemin $A\mathbf{x} = \mathbf{b}$ ratkaiseminen Gaussin eliminointimenetelmällä vaatii kertalukua $\frac{1}{3}n^3$ olevan määrän floppeja.

4 Tuenta

4.1 Tuennan periaate

Eliminointi ja LU-hajotelman muodostaminen onnistuvat kuten luvussa 3, vain jos kussakin eliminointivaiheessa tukialkio on nollasta poikkeava. Jossakin vaiheessa tukialkio häviää, kuten esimerkiksi, jos

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 2 \\ 4 & 5 & 6 \end{pmatrix}. \tag{4.1}$$

Tässä päästään pälkähästä vaihtamalla yhtälöiden järjestystä yhtälöryhmässä $A\mathbf{x} = \mathbf{b}$, jolloin kerroinmatriisin vaakarivien järjestys vaihtuu samalla tavalla.

Matriisin (4.1) tapauksessa voitaisiin vaihtaa systeemin $A\mathbf{x} = \mathbf{b}$ toinen tai kolmas rivi ensimmäiseksi yhtälöksi kertomalla systeemi puolittain sopivalla permutaatiomatriisilla. Ensimmäinen ja kolmas yhtälö vaihtavat paikkaa kerrottaessa matriisilla

$$P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \text{ jolloin } P_1 A = \begin{pmatrix} 4 & 5 & 6 \\ 1 & 1 & 2 \\ 0 & 1 & 1 \end{pmatrix}$$

Tätä niinsanottua tuentaa käytetään yleisesti Gaussin eliminointimenetelmässä siten, että k:nnessa eliminointivaiheessa toteutetaan ensin mahdollinen rivienvaihto kertomalla sopivalla permutaatiomatriisilla P_k ja vasta tämän jälkeen tapahtuu varsinainen eliminointi eli kertominen alkeismatriisilla M_k. Kerroinmatriisin A saattaminen yläkolmiomuotoon voidaan tällöin esittää yhtälöllä

$$M_{n-1}P_{n-1} \cdots M_2 P_2 M_1 P_1 A = U. \tag{4.2}$$

Tuennalla täydennetty eliminointialgoritmi onnistuu periaatteessa aina, mikäli matriisin A on säännöllinen. Numeerisen stabiiliuden saavuttamiseksi matriisit P_k on kuitenkin pyrittävä valitsemaan siten, että lopulliset tukialkiot eivät ole liian pieniä.

Teoriassa riittäisi yksi ainoa rivienvaihto. Voidaan nimittäin näyttää, että jokaista säännöllistä matriisia A kohden on olemassa permutaatiomatriisi P siten, että matriisilla PA on LU-hajotelma

$$PA = LU. \tag{4.3}$$

Systeemi $A\mathbf{x} = \mathbf{b}$ palautuu tällöin systeemiksi

$$PA\mathbf{x} = LU\mathbf{x} = P\mathbf{b}$$

ja edelleen, kuten luvussa 3.3, pariksi kolmiomutoisia systeemejä:

$$\begin{cases} L\mathbf{y} = P\mathbf{b} \\ U\mathbf{x} = \mathbf{y} \end{cases}$$

Johdetaan kaava (4.3) tapauksessa $n = 3$, jolloin yhtälö (4.2) on muotoa

$$M_2 P_2 M_1 P_1 A = U \tag{4.4}$$

Koska P_2 on ortogonaalinen, eli $P_2^T P_2 = I$, niin yhtälösta (4.4) seuraa

$$M_2 P_2 M_1 P_2^T P_2 P_1 A = U.$$

Merkitsemällä $\widetilde{M}_1 = P_2 M_1 P_2^T$ ja $P = P_2 P_1$, saadaan

$$M_2 \widetilde{M}_1 P A = U. \tag{4.5}$$

Tässä sekä M_2 että $\widetilde{M}_1$ ovat alakolmiomatriiseja, joiden diagonaaliset alkiot ovat ykkösiä. $\widetilde{M}_1$ saadaan nimittäin matriisista M_1 permutoimalla sen ensimmäisen sarakkeen alkiot permutaatiomatriisin P_2 rivijärjestyksen mukaiseksi.

Jos esimerkiksi $P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ ja $M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -m_{21} & 1 & 0 \\ -m_{31} & 0 & 1 \end{pmatrix}$, niin

$$\widetilde{M}_1 = P_2 M_1 P_2^T = \begin{pmatrix} 1 & 0 & 0 \\ -m_{31} & 1 & 0 \\ -m_{21} & 0 & 1 \end{pmatrix}.$$

Määrittelemällä $\widetilde{M} = M_2 \widetilde{M}_1$, saadaan yhtälössä (4.5) etsitty hajotelma

$$PA = LU,$$

missä $L = \widetilde{M}^{-1}$.

4.2 Tuennan käytäntö

Jos jokin tukialkioista on nollasta eroava mutta pieni, eliminointialgoritmi saattaa olla numeerisesti epästabiili.

Ratkaistaan eliminointimenetelmällä esimerkiksi systeemi

$$A\mathbf{x} = \begin{pmatrix} 0.0001 & 0.5 \\ 0.4 & -0.3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.1 \end{pmatrix} = \mathbf{b}. \tag{4.6}$$

Suorittamalla laskutoimitukset nelinumeroisessa aritmetiikassa käyttämällä normaalipyöristystä, merkitään kaavan (3.3) mukaisesti

$$m_{21} = fl\left(\frac{a_{21}}{a_{11}}\right) = \frac{0.4}{0.0001} = 4000.$$

Kaavan (3.5) mukaisesti

$$a_{22}^{(2)} = fl\left(a_{22} - m_{21}a_{12}\right) = fl\left(-0.3 - 4000 \cdot 0.5\right) = -2000.$$

Näin ollen alkion a_{22} sisältämä informaatio ”hukkuu” pyöristysvirheeseen. Syypää on pieni tukialkio a_{11}, jonka ansiosta kertoja m_{21} on suuri.

Systeemin (4.6) oikealla puolella saadaan ensimmäisessä eliminointivaiheessa

$$b_2^{(2)} = fl\left(b_2 - m_{21}b_1\right) = fl\left(0.1 - 4000 \cdot 0.5\right) = -2000.$$

Havaitaan, että myös luvun b_2 sisältämä informaatio ”hukkuu”.

Eliminoinnin tuloksena syntyy kolmiomuotoinen systeemi

$$\begin{pmatrix} 0.0001 & 0.5 \\ 0 & -2000 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -2000 \end{pmatrix}. \tag{4.7}$$

Takaisinsijoituksessa ei tapahdu tällä kertaa pyöristysvirheitä, ja ratkaisu on $\mathbf{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Systeemin (4.6) oikea ratkaisu neljän numeron tarkkuudella on $\mathbf{x} = \begin{pmatrix} 0.9999 \\ 0.9998 \end{pmatrix}$.

Epätarkkuuden aiheuttaa pieni tukialkio kasvattaessaan lopullisen kolmiomuotoisen systeemin toisen vaakarivin alkioita niin suuriksi, että kyseisen vaakarivin alkuperäinen informaatio katoaa pyöristysten yhteydessä: systeemiin (4.7) olisi päädytty myös, jos alkuperäisen systeemin alkiot $a_{22} = -0.3$ ja $b_2 = 0.1$ olisi valittu nolliksi. Laskettu ratkaisu on itse asiassa tällä tavoin häirityn systeemin

$$\begin{pmatrix} 0.0001 & 0.5 \\ 0.4 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$$

tarkka ratkaisu. Koska häiriötä ei voida katsoa pieneksi, peräytyvän virheanalyysin kannalta algoritmi ei ole stabiili.

Numeerisen stabiiliuden saavuttamiseksi tulee Gaussin eliminointialgoritmia modifioida siten, ettei liian pieniä tukialkioita esiintyisi. Yleensä tämä saadaan aikaan osittaistuennalla vaihtamalla kerroinmatriisin vaakarivien järjestystä sopivalla tavalla kussakin eliminointivaiheessa.

Käytännössä lasketaan k:nnessa eliminointivaiheessa aluksi vastaavan kerroinmatriisin sarakkeen k subdiagonaalisten alkioiden itseisarvot ja merkitään

$$\gamma_k = \max\left\{\left|a_{ik}^{(k)}\right|,\ i > k\right\}.$$

Jos $\left|a_{kk}^{(k)}\right| \geq \gamma_k$ ei kyseisessä eliminointivaiheessa suoriteta rivinvaihtoa. Jos taas $\left|a_{kk}^{(k)}\right| < \gamma_k$, etsitään pienin kokonaisluku $l > k$ siten, että $\gamma_k = \left|a_{lk}^{(k)}\right|$. Tämän jälkeen vaihdetaan keskenään vaakarivit k ja l, ja suoritetaan varsinainen eliminointi. Vastaavan alkeismatriisin alkioille pätee tällöin

$$|m_{ik}| = \frac{\left|a_{ik}^{(k)}\right|}{\left|a_{kk}^{(k)}\right|} \leq 1 \quad (1 \leq k \leq n-1;\ k+1 \leq i \leq n).$$

Näin ollen algoritmissa syntyvän LU-hajotelman matriisin L jokainen alkio on itseisarvoltaan korkeintaan yksi.

Esimerkki: sovelletaan tuennalla täydennettyä eliminointialgoritmia matriisiin

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 3 & -1 & -1 \\ 2 & 3 & 1 \end{pmatrix}.$$

Ensimmäisessä vaiheessa on aluksi vaihdettava rivit 1 ja 2 kertomalla permutaatiomatriisilla

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{jolloin} \quad P_1 A = \begin{pmatrix} 3 & -1 & -1 \\ 1 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}.$$

Edelleen $M_1 = \begin{pmatrix} 1 & & \\ -\frac{1}{3} & 1 & \\ -\frac{2}{3} & & 1 \end{pmatrix}$ ja $A^{(2)} = M_1 P_1 A = \begin{pmatrix} 3 & -1 & -1 \\ & \frac{4}{3} & \frac{7}{3} \\ & \frac{11}{3} & \frac{5}{3} \end{pmatrix}$.

LU-hajotelman matriisilla U on sama ylin vaakarivi kuin matriisilla $P_1 A$,

jolloin $u_{11} = 3$, $u_{12} = -1$ ja $u_{13} = -1$. Matriisin L ensimmäisen sarakkeen alkiot ovat alustavasti $l_{21} = m_{21} = \frac{1}{3}$ ja $l_{31} = m_{31} = \frac{2}{3}$.

Toisessa eliminointivaiheessa on aluksi vaihdettava matriisin $A^{(2)}$ alemmat vaakarivit, sillä $\frac{11}{3} > \frac{4}{3}$. Samalla on muistettava vaihtaa myös matriisin L vastaavilla vaakariveillä sijaitsevat ensimmäisen sarakkeen subdiagonaaliset alkiot (vertaa luvun 4.1 esimerkki), minkä jälkeen $l_{21} = \frac{2}{3}$ ja $l_{31} = \frac{1}{3}$. Vastaava permutaatiomatriisi on

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{ja} \quad P_2 A^{(2)} = \begin{pmatrix} 3 & -1 & -1 \\ & \frac{11}{3} & \frac{5}{3} \\ & \frac{4}{3} & \frac{7}{3} \end{pmatrix}.$$

Toisessa eliminointivaiheessa käytetään kerrointa $m_{32} = \frac{4}{11}$, ja LU-hajotelmassa on nyt

$$P = P_2 P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \; L = \begin{pmatrix} 1 & & \\ \frac{2}{3} & 1 & \\ \frac{1}{3} & \frac{4}{11} & 1 \end{pmatrix} \text{ ja } U = \begin{pmatrix} 3 & -1 & -1 \\ & \frac{11}{4} & \frac{5}{3} \\ & & \frac{57}{33} \end{pmatrix}.$$

Osittaistuennan asemasta käytetään toisinaan niinsanottua *täydellistä tuentaa*, jossa rivienvaihdon ohella suoritetaan myös mahdollinen sarakkeidenvaihto siten, että tukialkioksi tulee itseisarvoltaan suurin luvuista $a_{ij}^{(k)}$, missä $i \geq k$ ja $j \geq k$.

5 Eliminointi lohkomatriiseilla

Pari laskusääntöä:
Yksikkömatriisille I pätee $IC = C$ aina, kun tulo IC on määritelty. Yleisemmin: olkoon I_l tyyppiä $l \times l$ oleva yksikkömatriisi ja B_{22} mielivaltainen $m \times m$-matriisi. Kerrottaessa jokin matriisi C vasemmalta $(l+m) \times (l+m)$-matriisilla

$$B = \begin{pmatrix} I_l & 0 \\ 0 & B_{22} \end{pmatrix} \tag{5.1}$$

tulon BC l ensimmäistä vaakariviä ovat samat kuin matriisissa C. Tämä johtuu siitä, että

$$BC = \begin{pmatrix} I_l & 0 \\ 0 & B_{22} \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ B_{22}C_{21} & B_{22}C_{22} \end{pmatrix} \tag{5.2}$$

aina, kun matriisin C lohkojen dimensiot ovat yhteensopivia siten, että tulot $B_{22}C_{21}$ ja $B_{22}C_{22}$ on määritelty.

Gaussin eliminointimenetelmässä esiintyvät matriisit M_k ovat muotoa (5.1) arvolla $l = k - 1$. Kerrottaessa matriisi

$$A^{(k)} = P_k M_{k-1} P_{k-1} \cdots M_1 P_1 A \tag{5.3}$$

vasemmalta matriisilla M_k säilyvät siis $k-1$ ensimmäistä vaakariviä muuttumattomina. Kirjoitettaessa $A^{(k)}$ lohkomuodossa kuten matriisi C kaavassa (5.2) nähdään, että arvolla $l = k - 1$ $C_{21} = 0$, jolloin myös vastaava lohko tulossa $BC = M_k A^{(k)}$ on nollamatriisi.

Eliminointialgoritmi ja LU-hajotelma voidaan yleistää lohkomatriiseille, jotka ovat tyyppiä $(l+m) \times (l+m)$. Olkoon

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \tag{5.4}$$

Tällainen matriisi A, jossa A_{11} on *säännöllinen* $l \times l$-matriisi, voidaan saattaa *lohkolliseen yläkolmiomuotoon* kertomalla vasemmalta *lohkollisella alkeismatriisilla* $M_1 = I - U_1 V_1^T$, missä $U_1 = \begin{pmatrix} 0 & 0 \\ A_{21} & A_{11}^{-1} \end{pmatrix}$ ja $V_1 = \begin{pmatrix} I_l \\ 0 \end{pmatrix}$ ovat $(l+m) \times l$-matriiseja. M_1 voidaan kirjoittaa lohkomuodossa

$$M_1 = \begin{pmatrix} I_l & 0 \\ 0 & I_m \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ A_{21} & A_{11}^{-1} \end{pmatrix} (I_l \; 0) = \begin{pmatrix} I_l & 0 \\ -A_{21}A_{11}^{-1} & I_m \end{pmatrix}.$$

Edelleen

$$M_1 A = \begin{pmatrix} I_l & 0 \\ -A_{21}A_{11}^{-1} & I_m \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & C \end{pmatrix}, \tag{5.5}$$

missä $C = A_{22} - A_{21}A_{11}^{-1}A_{12}$ vastaa eliminointialgoritmin matriisia

$$\begin{pmatrix} a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \ddots & \vdots \\ a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix}$$

kaavassa (3.6). Matriisi A_{11} vastaa eliminoinnin tukialkiota.

6 Symmetriset matriisit

6.1 LDL^T-hajotelma

Olkoon A symmetrinen $n \times n$-matriisi. Oletetaan, että A on säännöllinen ja että sovellettaessa Gaussin eliminointimalgoritmia matriisiin A ei esiinny

rivienvaihtoja. Tällöin matriisilla A on LU-hajotelma $A = LU$ ja matriisin A symmetrisyyden nojalla

$$LU = A = A^T = (LU)^T = U^T L^T.$$

Merkitsemällä $D = \operatorname{diag}(u_{ii})$ tästä seuraa

$$A = \left(U^T D^{-1}\right)\left(DL^T\right), \tag{6.1}$$

missä alakolmiomatriisin $U^T D^{-1}$ diagonaaliset alkiot ovat ykkösiä. Näin ollen (6.1) on myös matriisin A LU-hajotelma, joten LU-hajotelman yksikäsitteisyyden nojalla $L = U^T D^{-1}$ ja siis

$$A = LDL^T. \tag{6.2}$$

Tämä on symmetrisen matriisin A LDL^T-hajotelma.

Säännöllisellä symmetrisellä matriisilla ei aina ole LDL^T-hajotelmaa. Olkoon esimerkiksi $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Jos matriisilla A olisi muotoa (6.2) oleva hajotelma

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & l_{21} \\ 0 & 1 \end{pmatrix},$$

saataisiin

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} d_1 & d_1 l_{21} \\ d_1 l_{21} & d_1 l_{21} + d_2 \end{pmatrix},$$

josta seuraa keskenään ristiriidassa olevat yhtälöt $d_1 = 0$ ja $d_1 l_{21} = 1$.

Seuraavassa osoitetaan, että LDL^T-hajotelma on olemassa aina kun A on symmetrinen ja positiivisesti definiitti.

Lause 15. *Olkoon A symmetrinen ja positiivisesti definiitti $n \times n$-matriisi. Silloin matriisin A kaikki ominaisarvot ja singulaariarvot ovat positiivisia reaalilukuja. Edelleen, jos matriisista A poistetaan ensimmäinen rivi ja sarake, syntyvä $(n-1) \times (n-1)$-matriisi on myös positiivisesti definiitti.*

Todistus. Olkoon λ jokin matriisin A ominaisarvo ja $\mathbf{x}$ vastaava ominaisvektori. Silloin $A\mathbf{x} = \lambda\mathbf{x}$, joten

$$\mathbf{x}^T A\mathbf{x} = \lambda \mathbf{x}^T \mathbf{x} = \lambda \left\|\mathbf{x}\right\|_2^2.$$

Koska matriisi A on positiivisesti definiitti ja $\mathbf{x} \neq \mathbf{0}$, tästä seuraa $\lambda > 0$.

Matriisin A diagonaalisen alkion a_{ii} ohella tarkastellaan avaruuden $\mathbb{R}^n$ yksikkövektoria $\mathbf{e}_i$, jonka i:s komponentti on 1. Silloin $\mathbf{e}_i^T A \mathbf{e}_i = a_{ii}$, josta päätellään, että $a_{ii} > 0$ (sillä A positiivisesti definiitti).

Olkoon A_{22} se $(n-1) \times (n-1)$-matriisi, joka saadaan poistamalla matriisista A ensimmäinen rivi ja sarake. Olkoon $\mathbf{y}$ mielivaltainen nollasta eroava $(n-1)$-vektori ja olkoon $\mathbf{x} = \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix}$. Koska matriisi A on positiivisesti definiitti, niin $\mathbf{x}^T A \mathbf{x} > 0$. Toisaalta

$$\mathbf{x}^T A \mathbf{x} = \begin{pmatrix} 0 & \mathbf{y}^T \end{pmatrix} \begin{pmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} 0 \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{y}^T \end{pmatrix} \begin{pmatrix} A_{21}^T \mathbf{y} \\ A_{22} \mathbf{y} \end{pmatrix} = \mathbf{y}^T A_{22} \mathbf{y},$$

missä $A_{21} = (a_{21} \ \ldots \ a_{n1})^T$. Näin ollen myös $\mathbf{y}^T A_{22} \mathbf{y} > 0$, joten A_{22} on positiivisesti definiitti. □

6.2 Choleskyn hajotelma

Olkoon edelleen A symmetrinen ja positiivisesti definiitti $n \times n$-matriisi. Matriisin A LDL^T-hajotelma löydetään eliminointialgoritmia muistuttavalla menettelyllä seuraavassa esitetyllä tavalla. Kuten edellä, kirjoitetaan matriisi A lohkomuodossa

$$A = \begin{pmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix}.$$

Gaussin elimointialgoritmi (ilman tuentaa) alkaa kertomalla matriisi A vasemmalta alkeismatriisilla M_1, jolloin

$$M_1 A = \begin{pmatrix} 1 & 0 \\ -\left(\frac{1}{a_{11}}\right) A_{21} & I_{n-1} \end{pmatrix} \begin{pmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & A_{21}^T \\ 0 & A_{22} - \left(\frac{1}{a_{11}}\right) A_{21} A_{21}^T \end{pmatrix}$$

Matriisin A symmetrisyyttä voidaan hyödyntää kertomalla tämän jälkeen oikealta matriisilla M_1^T, jolloin syntyvä matriisi

$$M_1 A M_1^T = \begin{pmatrix} a_{11} & 0 \\ 0 & A_{22} - \frac{1}{a_{11}} A_{21} A_{21}^T \end{pmatrix}.$$

on edelleen symmetrinen ja positiivisesti definiitti. Jos nimittäin $\mathbf{x}$ on mielivaltainen nollasta eroava n-vektori, niin $M_1^T \mathbf{x} \neq \mathbf{0}$ (M_1 säännöllinen) ja siten $\mathbf{x}^T M_1 A M_1^T \mathbf{x} > 0$.

Lauseen 15 nojalla matriisin $M_1 A M_1^T$ lohko $A_{22} - \frac{1}{a_{11}} A_{21} A_{21}^T$on myös symmetrinen ja positiivisesti definiitti, joten seuraaava eliminointivaihe voidaan

aloittaa ilman tuentaa.

Symmetrisyyden säilyttämiseksi kerrotaan jälleen vasemmalta alkeismatriisilla M_2 ja oikealta matriisilla M_2^T, jolloin tulokseksi saadaan symmetrinen ja positiivisesti definiitti matriisi $M_2M_1AM_1^TM_2^T$, jonka kahden ensimmäisen sarakkeen ainoat nollasta eroavat alkiot ovat diagonaalisia.

Näin jatkamalla syntyy lopulta hajotelma

$$MAM^T = D, \tag{6.3}$$

missä $M = M_{n-1} \cdots M_1$ ja D on diagonaalimatriisi, jonka kaikki alkiot ovat positiivisia. Merkitsemällä $L = M^{-1}$ tästä seuraa etsitty matriisin A LDL^T-hajotelma $A = LDL^T$.

Saatu LDL^T-hajotelma voidaan esittää myös muodossa

$$A = LDL^T = LD^{\frac{1}{2}} \cdot D^{\frac{1}{2}}L^T = \bar{L}\bar{L}^T = R^TR,$$

missä $\bar{L} = LD^{\frac{1}{2}}$ ja $R = \bar{L}^T$ on säännöllinen yläkolmiomatriisi. Hajotelma

$$A = R^TR \tag{6.4}$$

on *Choleskyn hajotelma*. Se on olemassa aina, kun matriisi A on symmetrinen ja positiivisesti definiitti.

Yhtälö (6.4) pysyy voimassa, jos matriisin R minkä hyvänsä vaakarivin alkiot muutetaan vastaluvuikseen, mutta tavallisesti matriisi R valitaan kuten yllä siten, että kaikki diagonaaliset alkiot ovat positiivisia.

Choleskyn hajotelman matriisin R alkiot voidaan laskea vaaka- tai pystyriveittäin samaan tapaan kuin matriisien L ja U alkiot matriisin A LU-hajotelmassa (vrt. luku 3.2). Alkioittain kirjoitettuna hajotelma $A = R^TR$ on

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{12} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} r_{11} & & & \\ r_{12} & r_{22} & & \\ \vdots & \vdots & \ddots & \\ r_{1n} & r_{2n} & \dots & r_{nn} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}.$$

Laskettaessa matriisin R alkiot riveittäin nähdään aluksi, että $a_{11} = r_{11}^2$,

joten $r_{11} = \sqrt{a_{11}}$. Neliöjuuri on hyvin määritelty, sillä matriisin A on positiivisesti definiitti, jolloin $a_{11} > 0$ (Lause 15). Muut ensimmäisen vaakarivin alkiot saadaan yhtälöistä

$$a_{12} = r_{11}r_{12},\ \ldots\ ,\ a_{1n} = r_{11}r_{1n}$$

muodossa

$$r_{1j} = \frac{a_{1j}}{r_{11}} \quad (2 \leq j \leq n)\,.$$

Toisella rivillä nähdään aluksi, että $a_{22} = r_{12}^2 + r_{22}^2$, joten $r_{22}^2 = a_{22} - r_{12}^2$ ja $r_{22} = \sqrt{a_{22} - r_{12}^2}$. Muut toisen vaakarivin alkiot saadaan yhtälöistä $a_{2j} = r_{12}r_{1j} + r_{22}r_{2j}$ $(3 \leq j \leq n)$, joista seuraa

$$r_{2j} = \frac{a_{2j} - r_{12}r_{1j}}{r_{22}} \quad (3 \leq j \leq n)\,.$$

Yleisesti, kun matriisin R $k-1$ ensimmäistä vaakariviä on laskettu, määrätään k:nnella rivillä aluksi diagonaalinen alkio yhtälön

$$a_{kk} = r_{1k}^2 + r_{2k}^2 + \ldots + r_{kk}^2 \tag{6.5}$$

perusteella, jolloin $r_{kk}^2 = a_{kk} - r_{1k}^2 - r_{2k}^2 - \ldots - r_{k-1,k}^2$. Sen jälkeen muut k:nnen rivin alkiot saadaan yhtälöstä $a_{kj} = r_{1k}r_{1j} + r_{2k}r_{2j} + \ldots + r_{kk}r_{kj}$ $(k+1 \leq j \leq n)$ muodossa

$$r_{kj} = \frac{a_{kj} - r_{1k}r_{1j} - \ldots - r_{k-1,k}r_{k-1,j}}{r_{kk}}$$

Choleskyn hajotelmaa käyttämällä voidaan symmetrinen lineaarinen systeemi ratkaista normaalia pienemmällä työmäärällä. Hajotelman $A = R^TR$ muodostamiseen tarvitaan nimittäin vain noin $\frac{1}{6}n^3$ floppia eli suunnilleen puolet eliminointialgoritmin vaatimasta työmäärästä. Sen jälkeen systeemi $R^TR\mathbf{x} = \mathbf{b}$ palautuu yhtälöpariksi

$$\begin{cases} R^T\mathbf{y} = \mathbf{b} \\ R\mathbf{x} = \mathbf{y} \end{cases},$$

jossa kerroinmatriisit ovat toistensa transpooseja.

Voidaan osoitaa, että yläoleva algoritmi yhtälöryhmän $A\mathbf{x} = \mathbf{b}$ ratkaisemiseksi Choleskyn hajotelman avulla *ilman tuentaa* on peräytyvästi numeerisesti stabiili, mikäli A ei ole kovin häiriöaltis. Olkoon esimerkiksi

$$A = \begin{pmatrix} 0.0001 & 0.01 \\ 0.01 & 100 \end{pmatrix}$$

Vaikka tukialkio a_{11} on kovin pieni, ei Choleskyn hajotelman matriisissa

$$R = \begin{pmatrix} 0.01 & 1.0000 \\ 0 & 9.9999 \end{pmatrix}$$

esiinny hyvin suuria alkioita. Tämä johtuu yleisesti siitä, että yhtälöiden (6.5) $a_{kk} = r_{1k}^2 + 2_{2k}^2 + \ldots r_{kk}^2$ perusteella $|r_{ik}| \leq \sqrt{a_{kk}}$.

7 QR-hajotelma

7.1 Householderin muunnokset

Liitetään jokaiseen nollasta eroavaan vektoriin $\mathbf{u}$ vastaava *Householderin matriisi*

$$H = I - \frac{2\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|_2^2} = I - \frac{\mathbf{u}\mathbf{u}^T}{\beta},$$

missä $\beta = \frac{1}{2}\|\mathbf{u}\|_2^2$. Tällöin H on symmetrinen alkeismatriisi, jota merkitään myös $H(\mathbf{u})$. Lisäksi H on ortogonaalinen:

$$\begin{aligned} H^T H &= HH^T = H^2 \\ &= \left(I - \frac{\mathbf{u}\mathbf{u}^T}{\beta}\right)\left(I - \frac{\mathbf{u}\mathbf{u}^T}{\beta}\right) \\ &= I - 2\frac{\mathbf{u}\mathbf{u}^T}{\beta} + \frac{\mathbf{u}\mathbf{u}^T\mathbf{u}\mathbf{u}^T}{\beta^2} \\ &= I - 2\frac{\mathbf{u}\mathbf{u}^T}{\beta} + \frac{\|\mathbf{u}\|_2^2\,\mathbf{u}\mathbf{u}^T}{\beta^2} \quad \left(\|\mathbf{u}\|_2^2 = 2\beta\right) \\ &= I - 2\frac{\mathbf{u}\mathbf{u}^T}{\beta} + \frac{2}{\beta}\mathbf{u}\mathbf{u}^T \\ &= I \end{aligned}$$

Matriisi $H(\mathbf{u})$ riippuu vain vektorin $\mathbf{u}$ suunnasta ja $H(\mathbf{u}) \neq I$.

Householderin matriisin määrittelemä lineaarikuvaus on *Householderin muunnos*. Jos kahdella eri vektorilla $\mathbf{a}$ ja $\mathbf{b}$ on sama 2-normi, on aina olemassa Householderin muunnos H siten, että $H\mathbf{a} = \mathbf{b}$. H löydetään ratkaisemalla vektorin $\mathbf{u}$ suhteen yhtälö

$$H\mathbf{a} = \left(I - \frac{\mathbf{u}\mathbf{u}^T}{\beta}\right)\mathbf{a} = \mathbf{b}. \tag{7.1}$$

Kirjoittamalla (7.1) muotoon

$$\left(-\frac{\mathbf{u}^T\mathbf{u}}{\beta}\right)\mathbf{u} = \mathbf{b} - \mathbf{a}$$

nähdään, että vektorin $\mathbf{u}$ on oltava vektorin $\mathbf{b}-\mathbf{a}$ suuntainen: $\mathbf{u} = \gamma(\mathbf{b}-\mathbf{a})$. Kääntäen, jos $\mathbf{u} = \gamma(\mathbf{b}-\mathbf{a})$, missä $\gamma \in \mathbb{R}\setminus\{0\}$, niin

$$\begin{aligned}
\beta &= \frac{1}{2}\|\mathbf{u}\|_2^2 \\
&= \frac{1}{2}\gamma^2\|\mathbf{b}-\mathbf{a}\|_2^2 \\
&= \frac{\gamma^2}{2}(\mathbf{b}-\mathbf{a})^T(\mathbf{b}-\mathbf{a}) \\
&= \frac{\gamma^2}{2}\left(\mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{a} - \mathbf{a}^T\mathbf{b} + \mathbf{a}^T\mathbf{a}\right) \\
&= \frac{\gamma^2}{2}\left(\underbrace{\|\mathbf{b}\|_2^2}_{=\|\mathbf{a}\|_2^2} + \|\mathbf{a}\|_2^2 - 2\mathbf{b}^T\mathbf{a}\right) \\
&= \gamma^2\left(\|\mathbf{a}\|_2^2 - \mathbf{b}^T\mathbf{a}\right)
\end{aligned}$$

ja siis

$$H\mathbf{a} = \left(I - \frac{\mathbf{u}\mathbf{u}^T}{\beta}\right)\mathbf{a} = \mathbf{a} - \frac{1}{\beta}\left(\mathbf{u}^T\mathbf{a}\right)\mathbf{u}.$$

Tässä $\mathbf{u}^T\mathbf{a} = \gamma\left(\mathbf{b}^T - \mathbf{a}^T\right)\mathbf{a} = \gamma\left(\mathbf{b}^T\mathbf{a} - \|\mathbf{a}\|_2^2\right) = -\frac{\beta}{\gamma}$, joten

$$H\mathbf{a} = \mathbf{a} - \frac{1}{\beta}\left(-\frac{\beta}{\gamma}\right)\gamma(\mathbf{b}-\mathbf{a}) = \mathbf{a} + (\mathbf{b}-\mathbf{a}) = \mathbf{b}.$$

Mielivaltaisen vektorin $\mathbf{c}$ kuva Householderin muunnoksessa H saadaan vähentämällä vektorista $\mathbf{c}$ eräs vektorin $\mathbf{u}$ suuntainen vektori:

$$H\mathbf{c} = \left(I - \frac{\mathbf{u}\mathbf{u}^T}{\beta}\right)\mathbf{c} = \mathbf{c} - \left(\frac{\mathbf{u}^T\mathbf{c}}{\beta}\right)\mathbf{u}. \tag{7.2}$$

Jos jokin vektorin $\mathbf{u}$ komponenteista häviää, ei vastaava vektorin $\mathbf{c}$ komponentti siis muutu kuvauksessa $\mathbf{c} \mapsto H\mathbf{c}$. Edelleen $H\mathbf{c} = \mathbf{c}$, mikäli $\mathbf{u}^T\mathbf{c} = 0$.

Olkoon esimerkiksi $\mathbf{a} = \begin{pmatrix} \frac{9}{2} \\ -1 \end{pmatrix}$ ja $\mathbf{b} = \begin{pmatrix} \frac{7}{2} \\ -3 \end{pmatrix}$, jolloin $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2$. Etsitään Householderin muunnos H siten, että $H\mathbf{a} = \mathbf{b}$. Ylläolevan perusteella

riittää, että $\mathbf{u}$ on vektorin $\mathbf{b}-\mathbf{a} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$ suuntainen. Valitsemalla $\mathbf{u} = \mathbf{b}-\mathbf{a}$ saadaan $\|\mathbf{u}\|_2 = \sqrt{5}$ ja

$$H = I - \frac{2\mathbf{u}\mathbf{u}^T}{\|\mathbf{u}\|_2^2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{2}{5}\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ -\frac{4}{5} & -\frac{3}{5} \end{pmatrix}.$$

Selvästi $H\mathbf{a} = \mathbf{b}$.

7.2 QR-hajotelma

Jokaista säännöllistä matriisia A kohden voidaan löytää Householderin matriisit $H_1, H_2, \ldots, H_{n-1}$ siten, että

$$H_{n-1} \cdots H_2 H_1 A = R \tag{7.3}$$

on säännöllinen yläkolmiomatriisi.

Matriisien $H_1, \ldots, H_{n-1}$ konstruointi muistuttaa eliminointialgoritmia sikäli, että matriisin A kertominen matriisilla H_1 hävittää ensimmäisen sarakkeen subdiagonaaliset alkiot ja kerrottaessa edelleen vasemmalta matriisilla H_2 häviävät toisen sarakkeen subdiagonaaliset alkiot ja niin edelleen.

Jos $H_1 = I - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\beta_1}$, niin matriisin $H_1 A$ ensimmäinen sarake on

$$H_1\mathbf{a}_1 = \left(I - \frac{\mathbf{u}_1\mathbf{u}_1^T}{\beta_1}\right)\mathbf{a}_1 = \mathbf{a}_1 - \frac{\mathbf{u}_1^T\mathbf{a}_1}{\beta_1}\mathbf{u}_1.$$

Tämän sarakkeen subdiagonaaliset alkiot ovat nollia, mikäli $H_1\mathbf{a}_1$ on avaruuden $\mathbb{R}^n$ yksikkövektorin $\mathbf{e}_1$ suuntainen:

$$H_1\mathbf{a}_1 = r_{11}\mathbf{e}_1 = \begin{pmatrix} r_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{7.4}$$

Luvun 7.1 perusteella (7.4) toteutuu, jos $|r_{11}| = \|\mathbf{a}_1\|_2$ ja $\mathbf{u}_1$ on vektorin $r_{11}\mathbf{e}_1 - \mathbf{a}_1$ suuntainen, seuraavassa valitaan

$$\mathbf{u}_1 = \mathbf{a}_1 - r_{11}\mathbf{e}_1 = \begin{pmatrix} a_{11} - r_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}.$$

Skalaari r_{11} voi olla joko positiivinen tai negatiivinen, mutta merkitsevien numeroiden kumoutumisen välttämiseksi r_{11} valitaan yleensä erimerkkiseksi kuin a_{11}. Poikkeustapauksen muodostaa tilanne, jossa $\mathbf{a}_1$ on alunperin jo vektorin $\mathbf{e}_1$ suuntainen.

Merkitään $A^{(2)} = H_1 A$, jolloin $A^{(2)}$ on muotoa

$$A^{(2)} = H_1 A = \begin{pmatrix} r_{11} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix}.$$

Olkoon esimerkiksi $A = \begin{pmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & -1 & -1 \end{pmatrix}$ (vrt. luvun 3.1 esimerkki). Silloin

$$\|\mathbf{a}_1\|_2 = \sqrt{14} = |r_{11}| \quad \text{ja} \quad \mathbf{u}_1 = \begin{pmatrix} 1+\sqrt{14} \\ 2 \\ 3 \end{pmatrix}.$$

Neljän numeron tarkkuudella

$$H_1 = \begin{pmatrix} -0.2673 & -0.5345 & -0.8018 \\ -0.5345 & 0.7745 & -0.3382 \\ -0.8018 & -0.3382 & 0.4927 \end{pmatrix}$$

$$A^{(2)} = H_1 A = \begin{pmatrix} -3.742 & -1.069 & -0.2673 \\ 0 & 2.127 & 0.04368 \\ 0 & -2.309 & -2.434 \end{pmatrix}.$$

Seuraava Householderin matriisi H_2 pyritään valitsemaan siten, että matriiseilla $A^{(2)}$ ja $H_2 A^{(2)}$ on sama ensimmäinen rivi ja sarake ja siten, että toisen sarakkeen subdiagonaaliset alkiot matriisissa $A^{(3)} = H_2 A^{(2)}$ ovat nollia. Jos $H_2 = I - \frac{\mathbf{u}_2 \mathbf{u}_2^T}{\beta_2}$, niin matriisin $H_2 A^{(2)}$ toinen sarake

$$H_2 \mathbf{a}_2^{(2)} = \left(I - \frac{\mathbf{u}_2 \mathbf{u}_2^T}{\beta_2} \right) \mathbf{a}_2^{(2)} = \mathbf{a}_2^{(2)} - \frac{\mathbf{u}^T \mathbf{a}_2^{(2)}}{\beta_2} \mathbf{u}_2.$$

Subdiagonaaliset alkiot häviävät, jos $H_2\mathbf{a}_2^{(2)}$ on muotoa

$$H_2\mathbf{a}_2^{(2)} = \begin{pmatrix} a_{12}^{(2)} \\ r_{22} \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Tähän päästään valitsemalla r_{22} siten, että $r_{22}^2 + \left(a_{12}^{(2)}\right) = \left\|\mathbf{a}_2^{(2)}\right\|_2^2$, jolloin $\mathbf{u}_2$ voi olla esimerkiksi

$$\mathbf{u}_2 = \mathbf{a}_2^{(2)} - \begin{pmatrix} a_{12}^{(2)} \\ r_{22} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ a_{22}^{(2)} - r_{22} \\ a_{32}^{(2)} \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix}.$$

Koska vektorin $\mathbf{u}_2$ ensimmäinen komponentti on 0, niin matriiseilla $A^{(2)}$ ja $H_2A_2^{(2)}$ on sama ensimmäinen vaakarivi (kts. kaavaa (7.2) seuraava huomautus). Myös ensimmäiset sarakkeet ovat samat, sillä $\mathbf{u}_2^T\mathbf{a}_1^{(2)} = 0$.

Prosessia jatkamalla saadaan lopulta hajotelma (7.3). Koska matriisit $H_1, \ldots, H_{n-1}$ ovat symmetrisiä ja ortogonaalisia, niiden tulo

$$Q = H_1 \cdots H_{n-1}$$

on ortogonaalinen ja

$$Q^T = H_{n-1}^T \cdots H_1^T = H_{n-1} \cdots H_1.$$

Siten (7.3) voidaan esittää muodossa

$$Q^T A = R$$

ja kertomalla vasemmalta matriisilla Q saadaan matriisin A *QR-hajotelma*

$$A = QR.$$

QR-hajotelman laskemiseen tarvitaan noin $\frac{2}{3}n^3$ floppia; työmäärä on siis kaksinkertainen eliminointialgoritmiin verrattuna. Etuna QR-hajotelmalla on sen numeerinen stabiilisuus: kerrottaessa Householderin matriisilla sarakkeiden 2-normit säilyvät eikä kerroinmatriisin alkioissa tapahdu pahanlaatuista

kasvua (vrt. luku 4.2).

Lineaarisen systeemin $A\mathbf{x} = \mathbf{b}$ ratkaiseminen QR-hajotelman avulla tapahtuu kirjoittamalla systeemi $QR\mathbf{x} = \mathbf{b}$ muotoon

$$R\mathbf{x} = Q^T\mathbf{b}. \tag{7.5}$$

Aluksi lasketaan $Q^T\mathbf{b}$, jonka jälkeen systeemi (7.5) ratkeaa takaisinsijoituksella.

Systeemille (1.2) saadaan neljän numeron tarkkuudella

$$R = \begin{pmatrix} -3.742 & -1.069 & -0.2673 \\ & -3.140 & -1.820 \\ & & -1.617 \end{pmatrix}, \quad Q^T\mathbf{b} = \begin{pmatrix} -6.682 \\ 1.320 \\ -1.617 \end{pmatrix}.$$

Takaisinsijoituksella saadaan tarkka ratkaisu $\mathbf{x} = (2, -1, 1)^T$.

Yksityiskohtainen virheanalyysi osoittaa, että systeemin $A\mathbf{x} = \mathbf{b}$ ratkaiseminen QR-hajotelman avulla on yleisessä tapauksessa peräytyvän virheanalyysin kannalta stabiilimpaa kuin LU-hajotelmaa käytettäessä.

8 Pienimmän neliösumman ratkaisu

8.1 Ratkaisun ominaisuuksia

Tarkastellaan lineaarista systeemiä $A\mathbf{x} = \mathbf{b}$, missä

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{ja} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

Jos systeemi $A\mathbf{x} = \mathbf{b}$ ei ole ratkeava, vektoria $\mathbf{b}$ ei voida esittää lineaarikombinaationa matriisin A sarakkeista. Tällöin voidaan tyytyä etsimään sellaista vektoria $\mathbf{x}$, että vektori $A\mathbf{x}$ on mahdollisimman lähellä vektoria $\mathbf{b}$ siten, että

$$\|\mathbf{b} - A\mathbf{x}\|_2 \tag{8.1}$$

on mahdollisimman pieni. Jos (8.1) saavuttaa minimin jollakin vektorilla $\mathbf{x}$, saa miniminsä myös lausekkeen neliö:

$$\|\mathbf{b} - A\mathbf{x}\|_2^2 = \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij}x_j \right)^2.$$

Vektori $\rho = \mathbf{b} - A\mathbf{x}$ on minimointiprobleeman *residuaali*.

Palautetaan mieleen maaliavaruus $R(A)$ ja nolla-avaruus $\mathcal{N}\left(A^T\right)$, jotka ovat toistensa ortogonaalisia komplementteja: jokainen avaruuden $\mathbb{R}^m$ vektori $\mathbf{c}$ voidaan esittää yksikäsitteisellä tavalla kahden ortogonaalisen vektorin summana muodossa $\mathbf{c} = \mathbf{c}_R + \mathbf{c}_{\mathcal{N}}$, missä $\mathbf{c}_R \in R(A)$ ja $\mathbf{c}_{\mathcal{N}} \in \mathcal{N}\left(A^T\right)$.

Kirjoitetaan vektoreille $\mathbf{b}$ ja ρ vastaavat esitykset:

$$\mathbf{b} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}} \quad \text{ja} \quad \rho = \rho_R + \rho_{\mathcal{N}}.$$

Yhtälöstä $\rho = \mathbf{b} - A\mathbf{x}$ seuraa tällöin

$$\rho_R + \rho_{\mathcal{N}} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}} - A\mathbf{x} \quad \text{eli} \quad \rho_R - \mathbf{b}_R + A\mathbf{x} = \mathbf{b}_{\mathcal{N}} - \rho_{\mathcal{N}}.$$

Tässä vasen puoli kuuluu avaruuteen $R(A)$ ja oikea puoli avaruuteen $\mathcal{N}\left(A^T\right)$. Koska $R(A) \cap \mathcal{N}\left(A^T\right) = \{\mathbf{0}\}$, päätellään siis

$$\rho_R = \mathbf{b}_R - A\mathbf{x} \quad \text{ja} \quad \rho_{\mathcal{N}} = \mathbf{b}_{\mathcal{N}}. \tag{8.2}$$

Residuaalin ρ pituuden laskemiseksi todetaan, että

$$\begin{aligned} \|\rho\|_2^2 = \rho^T\rho &= \left(\rho_R^T + \rho_{\mathcal{N}}^T\right)(\rho_R + \rho_{\mathcal{N}}) \\ &= \rho_R^T\rho_R + \underbrace{\rho_R^T\rho_{\mathcal{N}}}_{=0} + \underbrace{\rho_{\mathcal{N}}^T\rho_R}_{=0} + \rho_{\mathcal{N}}^T\rho_{\mathcal{N}} \\ &= \|\rho_R\|_2^2 + \|\rho_{\mathcal{N}}\|_2^2 \end{aligned}$$

(sillä ρ_R ja $\rho_{\mathcal{N}}$ ovat ortogonaaliset, jolloin $\rho_R^T\rho_{\mathcal{N}} = \rho_{\mathcal{N}}^T\rho_R = 0$).

Yhtälöiden (8.1) ja (8.2) perusteella saadaan siten

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|_2^2 = \|\rho\|_2^2 &= \|\rho_R\|_2^2 + \|\rho_{\mathcal{N}}\|_2^2 \\ &= \|\mathbf{b}_R - A\mathbf{x}\|_2^2 + \|\mathbf{b}_{\mathcal{N}}\|_2^2 \\ &\geq \|\mathbf{b}_{\mathcal{N}}\|_2^2 \end{aligned}$$

ja yhtäsuuruus on voimassa täsmälleen silloin, kun $\mathbf{b}_R - A\mathbf{x} = \mathbf{0}$.

Koska $\mathbf{b}_R \in R(A)$, systeemi $A\mathbf{x} = \mathbf{b}_R$ on aina ratkeava. Näin ollen on olemassa $\mathbf{x}$ siten, että $\mathbf{b}_R - A\mathbf{x} = \mathbf{0}$, jolloin lauseke (8.1) saavuttaa pienimmän arvonsa $\|\mathbf{b}_{\mathcal{N}}\|_2$. Pienin arvo saavutetaan täsmälleen silloin, kun $\rho_R = \mathbf{b}_R - A\mathbf{x} = \mathbf{0}$ eli täsmälleen silloin, kun $\rho \in \mathcal{N}\left(A^T\right)$. Tästä seuraa

Lause 16. *Seuraavat kolme ehtoa ovat yhtäpitäviä*

(*i*) *Lauseke* (8.1) *saa pienimmän mahdollisen arvon.*

(*ii*) $A^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{0}$ (8.3)

(*iii*) $A\boldsymbol{x} = \boldsymbol{b}_R$ (8.4)

Minimointiprobleeman ratkaisu $\mathbf{x}$ on yksikäsitteinen, jos ja vain jos systeemillä (8.4) on täsmälleen yksi ratkaisu eli jos matriisin A sarakkeet ovat lineaarisesti riippumattomat.

Olkoon esimerkiksi

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{ja} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -5 \end{pmatrix}. \tag{8.5}$$

Esitys $\mathbf{b} = \mathbf{b}_R + \mathbf{b}_{\mathcal{N}}$ löydetään kuten Osan I luvun 2.2 viimeisessä esimerkissä:

$$\mathbf{b}_R = \begin{pmatrix} -1 \\ 2 \\ -3 \end{pmatrix} \quad \text{ja} \quad \mathbf{b}_{\mathcal{N}} = \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix}.$$

Koska matriisin A sarakkeet ovat lineaarisesti riippumattomat, systeemin (8.4) ratkaisu on yksikäsitteinen. Ratkaisemalla (8.4) saadaan $\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$ ja vastaava residuaali $\rho = \mathbf{b} - A\mathbf{x} = \begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix}$.

Pienimmän neliösumman ratkaisua voidaan käyttää *lineaaristen mallien* ratkaisemiseen.

Esimerkki: Radioaktiivisten aineiden seoksessa on kolmea eri isotooppia, joiden määrät x_1, x_2 ja x_3 ovat tuntemattomia. Hajoamislain mukaan havaittavan säteilyn määrä ajan funktiona on muotoa

$$x_1 e^{-\lambda_1 t} + x_2 e^{-\lambda_2 t} + x_3 e^{-\lambda_3 t}, \tag{8.6}$$

missä vakiot λ_1, λ_2 ja λ_3 riippuvat isotooppien puoliintumisajoista. Käytännössä säteilymittarin lukemat eivät ole tarkalleen lausekkeen (8.6) mukaisia, mutta suoritettassa m havaintoa hetkillä $t_1, \ldots, t_m$ saadaan sarja lukemia $b_1, \ldots, b_m$, joille pätee likimain

$$b_i \approx x_1 e^{-\lambda_1 t_i} + x_2 e^{-\lambda_2 t_i} + x_3 e^{-\lambda_3 t_i} = \sum_{j=1}^{3} a_{ij} x_j, \quad (1 \le i \le m) \tag{8.7}$$

missä $a_{ij} = e^{-\lambda_j t_i}$.

Mittaustulosten perusteella voidaan arvioida isotooppien (tuntemattomia) määriä x_1, x_2 ja x_3 seuraavasti: Pyritään määräämään x_1, x_2 ja x_3 siten, että kaavat (8.7) ovat voimassa mahdollisimman tarkkaan pienimmän neliösumman mielessä. Toisin sanoen, etsitään luvut x_1, x_2 ja x_3 siten, että

$$\sum_{i=1}^{n} \left(b_i - \sum_{j=1}^{3} a_{ij} x_j \right)$$

on mahdollisimman pieni. Tämän minimointiprobleeman ratkaisu on systeemin $A\mathbf{x} = \mathbf{b}$ pienimmän neliösumman ratkaisu.

9 Normaaliyhtälöt

Kirjoittamalla yhtälö (8.4) muotoon

$$A^T A\mathbf{x} = A^T \mathbf{b} \tag{9.1}$$

saadaan niinsanotut *normaaliyhtälöt*. Systeemi (9.1) on aina ratkeava vaikka kerroinmatriisi $A^T A$ olisi singulaarinen. Lauseen 16 nojalla jokainen systeemin (9.1) ratkaisu $\mathbf{x}$ on yhtälöryhmän $A\mathbf{x} = \mathbf{b}$ pienimmän neliösumman ratkaisu. Kääntäen: jokainen pienimmän neliösumman ratkaisu toteuttaa normaaliyhtälöt (9.1).

Lemma 17. *Matriisi $A^T A$ on positiivisesti definiitti, jos ja vain jos matriisi A on täyttä sarakeastetta eli matriisin A sarakkeet ovat lineaarisesti riippumattomat.*

Todistus.

$$\begin{aligned} A^T A \text{ on positiivisesti definiitti} &\Leftrightarrow \mathbf{x}A^T A\mathbf{x} > 0, \text{ aina kun } \mathbf{x} \neq \mathbf{0} \\ &\Leftrightarrow \|A\mathbf{x}\|_2^2 > 0, \text{ aina kun } \mathbf{x} \neq \mathbf{0} \\ &\Leftrightarrow A\mathbf{x} \neq \mathbf{0}, \text{ aina kun } \mathbf{x} \neq \mathbf{0} \\ &\Leftrightarrow A \text{ on täyttä sarakeastetta} \end{aligned}$$

□

Jos matriisi A on täyttä sarakeastetta, voidaan pienimmän neliösumman ratkaisu siis löytää seuraavalla algoritmilla:

1. Muodostetaan normaaliyhtälöiden kerroinmatriisi $A^T A$ ja vektori $A^T \mathbf{b}$.

2. Lasketaan Choleskyn hajotelma $A^TA = R^TR$, missä R on yläkolmiomatriisi.

3. Ratkaistaan kolmiomuotoiset systeemit $R^T\mathbf{y} = A^T\mathbf{b}$ ja $R\mathbf{x} = \mathbf{y}$. Jälkimmäisen kolmiomuotoisen systeemin ratkaisu on pienimmän neliösumman ratkaisu.

Olkoon esimerkiksi matriisi A ja vektori $\mathbf{b}$ kuten (8.5). Silloin matriisi A on täyttä sarakeastetta, ja

$$A^TA = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{ja} \quad A^T\mathbf{b} = \begin{pmatrix} 1 \\ -4 \end{pmatrix}.$$

Matriisin A^TA Choleskyn hajotelma on $A^TA = R^TR$, missä

$$R = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 1 \\ & \sqrt{3} \end{pmatrix} \quad \text{ja} \quad R^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & \\ 1 & \sqrt{3} \end{pmatrix}.$$

Systeemin $R^T\mathbf{y} = A^T\mathbf{b}$ ratkaisu saadaan eteenpäinsijoituksella:

$$\begin{cases} \frac{1}{\sqrt{2}} 2y_1 = 1 \\ \frac{1}{\sqrt{2}} \left(y_1 + \sqrt{3}y_2\right) = -4 \end{cases} \Rightarrow \; y_1 = \frac{1}{\sqrt{2}} \; \Rightarrow \; y_2 = -3\sqrt{\frac{3}{2}}.$$

Systeemi $R\mathbf{x} = \mathbf{y}$ on nyt

$$\begin{cases} \frac{1}{\sqrt{2}} \left(2x_1 + x_2\right) = \frac{1}{\sqrt{2}} \\ \sqrt{\frac{3}{2}} x_2 = -3\sqrt{\frac{3}{2}} \end{cases} \Rightarrow \; x_2 = -3 \; \Rightarrow \; x_1 = 2$$

Tulos $\mathbf{x} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$ on sama kuin luvun 8 esimerkissä.

Pienimmän neliösumman ratkaisun etsiminen normaaliyhtälöiden avulla saattaa olla numeerisesti epästabiilia, mikäli matriisi A^TA on häiriöaltis. Sen vuoksi probleema ratkaistaan monesti käyttämällä yleistettyä QR-hajotelmaa.