1

Ortogonalizacija

Osnovna tema ovog poglavlja je ortogonalnost vektora u prostoru $\mathbb{R}^m$ ili $\mathbb{C}^m$. Stoga ćemo koristiti isključivo euklidsku normu

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}} = \sqrt{\sum_{i=1}^{m} x_i^2}, \quad \mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{m} x_i y_i,$$

na $\mathbb{R}^m$, odnosno

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}} = \sqrt{\sum_{i=1}^{m} |x_i|^2}, \quad \mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^{m} x_i \overline{y}_i,$$

na $\mathbb{C}^m$. Algoritmi bazirani na ortogonalizaciji jednako djeluju u realnom i kompleksnom prostoru s tom razlikom da treba koristiti odgovarajući skalarni produkt. Jedna od razlika koja se može pojaviti je pitanje predznaka. U realnom slučaju predznak broja može biti ili +1 ili -1, dok u kompleksnom slučaju svaki broj z sa svojstvom $|z| = 1$ može imati ulogu predznaka (kompleksni predznak).

1.1 Gram–Schmidtova ortogonalizacija

Neka je zadan skup linearno nezavisnih vektora

$$\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m (\text{ili } \ \mathbb{C}^m)$$

pri čemu evidentno mora biti $m \geq n$. Gram-Schmidtovim postupkom na prirodan način se konstruira ortonormirani skup vektora

$$\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n \in \mathbb{R}^m (\text{ili } \ \mathbb{C}^m)$$

koji razapinju isti potprostor u $\mathbb{R}^m$ ($\mathbb{C}^m$). Postupak je sljedeći:

$$\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \\
\mathbf{q}_2 &= \frac{\overline{\mathbf{a}}_2}{\|\overline{\mathbf{a}}_2\|}, \quad \overline{\mathbf{a}}_2 = \mathbf{a}_2 - (\mathbf{q}_1 \cdot \mathbf{a}_2)\mathbf{q}_1, \\
\mathbf{q}_3 &= \frac{\overline{\mathbf{a}}_3}{\|\overline{\mathbf{a}}_3\|}, \quad \overline{\mathbf{a}}_3 = \mathbf{a}_3 - (\mathbf{q}_1 \cdot \mathbf{a}_3)\mathbf{q}_1 - (\mathbf{q}_2 \cdot \mathbf{a}_3)\mathbf{q}_2, \\
&\cdots \\
\mathbf{q}_n &= \frac{\overline{\mathbf{a}}_n}{\|\overline{\mathbf{a}}_n\|}, \quad \overline{\mathbf{a}}_n = \mathbf{a}_n - \sum_{i=1}^{n-1} (\mathbf{q}_i \cdot \mathbf{a}_n)\mathbf{q}_i.
\end{aligned}$$

Postupak počinje normiranjem prvog vektora. U i-tom koraku od vektora $\mathbf{a}_i$ oduzima njegovu projekciju na prvih $i-1$ vektora: $\mathbf{q}_1, \ldots, \mathbf{q}_{i-1}$. Time $\mathbf{q}_i$ postaje ortogonalan na sve prethodne vektore.

Lema 1.1 Gram-Schmidtov postupak primijenjen na linearno nezavisan skup vektora

$$\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\} \subset \mathbb{R}^m(\mathbb{C}^m)$$

daje ortonormiran skup vektora $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ za koji je

$$\mathcal{L}(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_i) = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_i)$$

za svako $i = 1, 2, \ldots, n$.

Dokaz. Dokaz je prilično evidentan, ali da bismo ga strogo proveli moramo koristiti matematičku indukciju. Za svako $k \geq 1$ dokazujemo istovremeno da skupovi $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k\}$ i $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ razapinju isti podprostor te da je skup $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k\}$ ortonormiran.

Za $k = 1$ tvrdnja je evidentan (baza indukcije). Pretpostavimo da je

$$\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k-1}\}$$

skup ortonormiranih vektora koji razapinje isti potprostor kao i $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}\}$. Za sve $j < k$ imamo

$$\begin{aligned}
\overline{\mathbf{a}}_k \cdot \mathbf{q}_j &= \mathbf{a}_k \cdot \mathbf{q}_j - \sum_{i=1}^{k-1} (\mathbf{q}_i \cdot \mathbf{a}_k)\mathbf{q}_i \cdot \mathbf{q}_j \\
&= \mathbf{a}_k \cdot \mathbf{q}_j - (\mathbf{q}_j \cdot \mathbf{a}_k) = 0.
\end{aligned}$$

Pokažimo da $\overline{\mathbf{a}}_k \neq 0$. U suprotnom bi bilo

$$\mathbf{a}_k = \sum_{i=1}^{k-1} (\mathbf{q}_i \cdot \mathbf{a}_k)\mathbf{q}_i \in \mathcal{L}(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{k-1}) = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1})$$

što je kontradikcija s pretpostavkom linearne nezavisnosti. (Uočite da smo koristili pretpostavku indukcije.) Time vidimo da je $\mathbf{q}_k$ dobro definiran, jedinične norme i ortogonalan na sve $\mathbf{q}_j$ za $j \leq 1 < k$. Stoga je skup $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k\}$ ortonormiran.

Nadalje, vrijedi $\mathbf{a}_k \in \mathcal{L}(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k)$ budući da je

$$\mathbf{a}_k = \|\overline{\mathbf{a}}_k\|\mathbf{q}_k + \sum_{i=1}^{k} (\mathbf{q}_i \cdot \mathbf{a}_n)\mathbf{q}_i,$$

pa prema pretpostavci indukcije vrijedi

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k) \subset \mathcal{L}(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k).$$

Zbog linearne nezavisnosti vektora $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ mora vrijediti

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k) = \mathcal{L}(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k)$$

što dokazuje lemu. Q.E.D.

Gram-Schmidtova postupak može se primijeniti i na prebrojiv skup linearno nezavisnih vektora čime dobivamo prebrojiv skup ortonormiranih vektora. Pri tome za beskonačan skup vektora kažemo da je linearno nezavisan ako je svaki njegov konačan podskup linearno nezavisan.

Algoritam Gram-Schmidtove ortogonalizacije ima ovaj oblik:

for $j = 1, 2, \ldots, n$ **do**
  for $i = 1, 2, \ldots, j-1$ **do**
    $r_{ij} = \mathbf{q}_i \cdot \mathbf{a}_j$
  end for
  $\overline{\mathbf{a}}_j = \mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij}\mathbf{q}_i$
  $r_{jj} = \|\overline{\mathbf{a}}_j\|$
  $\mathbf{q}_j = \overline{\mathbf{a}}_j / r_{jj}$
end for

Algoritam je prirodno formirati tako da radi na ulaznom skupu vektora koje pretvara u ortonormirane. Na taj način dolazimo do Algoritma 1.

Algoritam 1 ne koristi se u praksi budući da je numerički **nestabilan**. Problem predstavlja dokidanje značajnih znamenaka pri oduzimanju komponenti u smjeru prethodnih ortogonalnih vektora.

Stabilnost algoritma može se poboljšati ako se odmah nakon formiranja vektora $\mathbf{q}_i$ eliminira komponenta u njegovom smjeru u svim vektorima $\mathbf{a}_{i+1}, \ldots, \mathbf{a}_n$. U slučaju četiri vektora to bi funkcioniralo na sljedeći način:

$$\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \\
\mathbf{a}'_2 &= \mathbf{a}_2 - (\mathbf{q}_1 \cdot \mathbf{a}_2)\mathbf{q}_1, \\
\mathbf{a}'_3 &= \mathbf{a}_3 - (\mathbf{q}_1 \cdot \mathbf{a}_3)\mathbf{q}_1, \\
\mathbf{a}'_4 &= \mathbf{a}_4 - (\mathbf{q}_1 \cdot \mathbf{a}_4)\mathbf{q}_1.
\end{aligned}$$

Algorithm 1 Klasična Gram-Schmidtova ortogonalizacija

Ulaz: $n=$ broj vektora
$\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ linearno nezavisni vektori
for $j = 1, 2, \ldots, n$ **do**
  $\mathbf{v} = \mathbf{a}_j$
  for $i = 1, 2, \ldots, j-1$ **do**
    $r = \mathbf{a}_i \cdot \mathbf{v}$
    $\mathbf{a}_j = \mathbf{a}_j - r\mathbf{a}_i$
  end for
  $r = \|\mathbf{a}_j\|$
  $\mathbf{a}_j = \mathbf{a}_j / r$
end for
Izlaz: $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ ortonormirani vektori.

Nakon toga možemo izračunati

$$\mathbf{q}_2 = \frac{\mathbf{a}_2'}{\|\mathbf{a}_2'\|},$$

te

$$\begin{aligned} \mathbf{a}_3'' &= \mathbf{a}_3' - (\mathbf{q}_2 \cdot \mathbf{a}_3')\mathbf{q}_2, \\ \mathbf{a}_4'' &= \mathbf{a}_4' - (\mathbf{q}_2 \cdot \mathbf{a}_4')\mathbf{q}_2. \end{aligned}$$

Sada je

$$\mathbf{q}_3 = \frac{\mathbf{a}_3''}{\|\mathbf{a}_3''\|},$$

i konačno imamo

$$\mathbf{a}_4''' = \mathbf{a}_4'' - (\mathbf{q}_3 \cdot \mathbf{a}_4'')\mathbf{q}_3, \quad \mathbf{q}_4 = \frac{\mathbf{a}_4'''}{\|\mathbf{a}_4'''\|}.$$

U aritmetici beskonačne točnosti, zbog ortogonalnosti vektora $\mathbf{q}_i$, novi skalarni produkti jednaki su starim. Na primjer,

$$\mathbf{q}_3 \cdot \mathbf{a}_4'' = \mathbf{q}_3 \cdot (\mathbf{a}_4' - (\mathbf{q}_2 \cdot \mathbf{a}_4')\mathbf{q}_2) = \mathbf{q}_3 \cdot \mathbf{a}_4' = \mathbf{q}_3 \cdot (\mathbf{a}_4 - (\mathbf{q}_1 \cdot \mathbf{a}_4)\mathbf{q}_1) = \mathbf{q}_3 \cdot \mathbf{a}_4.$$

U aritmetici konačne točnosti evidentno se radi o dva različita algoritma. Novi algoritam možemo zapisati na sljedeći način:

for $i = 1, 2, \ldots, n$ **do**
  $r_{ii} = \|\mathbf{a}_i\|$
  $\mathbf{q}_i = \mathbf{a}_i / r_{ii}$
  for $j = i+1, \ldots, n$ **do**
    $r_{ij} = \mathbf{q}_i \cdot \mathbf{a}_j$
    $\mathbf{a}_j = \mathbf{a}_j - r_{ij}\mathbf{q}_i$
  end for
end for

Cijeli se algoritam može zapisati bez pomoćnih vektora kao u Algoritmu 2.

Algorithm 2 Modificirana Gram-Schmidtova ortogonalizacija

Ulaz: n= broj vektora
  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ linearno nezavisni vektori
for $i = 1, 2, \dots, n$ **do**
  $r = \|\mathbf{a}_i\|$
  $\mathbf{a}_i = \mathbf{a}_i / r$
  for $j = i + 1, \dots, n$ **do**
    $r = \mathbf{a}_j \cdot \mathbf{a}_i$
    $\mathbf{a}_j = \mathbf{a}_j - r\mathbf{a}_i$
  end for
end for
Izlaz: $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ ortonormirani vektori.

1.2 QR faktorizacija

Skup vektora prirodno je pamtiti u obliku matrice. Ako su $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ vektori u $\mathbb{R}^m$ (ili $\mathbb{C}^m$) onda ih smještavamo u stupce matrice A dimenzije $m \times n$. Ako je skup vektora linearno nezavisan, onda je nužno $m \geq n$ pa imamo općenito pravokutnu matricu koja ima više redaka nego stupaca.

Na isti način poredamo ortonormirane vektore $\mathbf{q}_i$, dobivene Gram-Schmidtovim postupkom, u stupce matrice koju ćemo nazvati Q, iste dimenzije $m \times n$. Uvedimo oznake

$$r_{ij} = \mathbf{q}_i \cdot \mathbf{a}_j \quad \text{za } i \leq j, \tag{1.1}$$

Vidimo da formule Gram-Schmidtove ortogonalizacije mogu biti zapisane u obliku

$$\begin{aligned}
\mathbf{a}_1 &= r_{11}\mathbf{q}_1 \\
\mathbf{a}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2, \\
\mathbf{a}_3 &= r_{13}\mathbf{q}_1 + r_{23}\mathbf{q}_2 + r_{33}\mathbf{q}_3, \\
&\cdots \\
\mathbf{a}_n &= r_{1n}\mathbf{q}_1 + r_{2n}\mathbf{q}_2 + \cdots + r_{nn}\mathbf{q}_n,
\end{aligned}$$

što vodi na sljedeći odnos između matrica A i Q:

$$\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} \\
a_{21} & a_{22} & \cdots & a_{2n} \\
a_{31} & a_{32} & \cdots & a_{3n} \\
a_{41} & a_{42} & \cdots & a_{4n} \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \cdots & \vdots \\
\vdots & \vdots & \cdots & \vdots \\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
=
\begin{bmatrix}
q_{11} & q_{12} & \cdots & q_{1n} \\
q_{21} & q_{22} & \cdots & q_{2n} \\
q_{31} & q_{32} & \cdots & q_{3n} \\
q_{41} & q_{42} & \cdots & q_{4n} \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \cdots & \vdots \\
\vdots & \vdots & \cdots & \vdots \\
q_{m1} & q_{m2} & \cdots & q_{mn}
\end{bmatrix}
\begin{bmatrix}
r_{11} & r_{12} & \cdots & r_{1n} \\
 & r_{22} & \cdots & r_{2n} \\
 & & \ddots & \vdots \\
 & & & r_{nn}
\end{bmatrix}$$

gdje je $(a_{ki})_{k=1}^m = \mathbf{a}_i$ i $(q_{ki})_{k=1}^m = \mathbf{q}_i$.

Faktorizacija matrice A dimenzije $m \times n$, $m \geq n$, oblika

$$A = QR$$

gdje je Q matrica dimenzije $m \times n$ s ortonormiranim stupcima, a R gornja trokutasta matrica dimenzije $n \times n$, naziva se reducirana QR faktorizacija matrice A. Uočimo da stupci matrice Q čine ortonormiranu bazu u slici matrice A.

Potpuna QR faktorizacija matrice A [1] je faktorizacija oblika

$$A = QR$$

gdje je Q matrica dimenzije $m \times m$ s ortonormiranim stupcima (unitarna matrica, odn. ortogonalna u realnom slučaju), a R gornja trokutasta matrica dimenzije $m \times n$. Potpuna QR faktorizacija se dobiva iz reducirane QR faktorizacije tako da se matrici Q doda $m - n$ ortonormiranih stupaca koji s ostalim stupcima čine ortonormiranu bazu u $\mathbb{R}^m$ (odnosno $\mathbb{C}^m$), a matrici R se dodaju nul-reci kako bi se dopunila do matrice dimenzije $m \times n$.

Primjer 1.1 Jedan jednostavan primjer reducirane QR faktorizacije je

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Odgovarajuća potpuna QR faktorizacija bi bila

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Teorem 1.1 Svaka matrica $A \in \mathbb{R}^{m \times n}$ (odnosno $A \in \mathbb{C}^{m \times n}$), $m \geq n$, ima QR faktorizaciju.

Dokaz. Ako je matrica punog ranga, onda je egzistencija faktorizacije dana Gram-Schmidtovom ortogonalizacijom. Ako matrica nije punog ranga, onda je moguće u tijeku ortogonalizacije dobiti nul vektor. U tom je slučaju dovoljno nul vektor zamijeniti s bilo kojim ortonormiranim vektorom, ortogonalnim na prethodno generirane vektore. Q.E.D.

Teorem 1.2 Svaka matrica punog ranga $A \in \mathbb{R}^{m \times n}$ (odnosno $A \in \mathbb{C}^{m \times n}$), $m \geq n$, ima jedinstvenu reduciranu QR faktorizaciju za koju je $r_{jj} > 0$.

[1] Odnosno, jednostavno QR faktorizacija.

Dokaz. Pokazuje se na osnovu Gram-Schmidtove ortogonalizacije. Jedina sloboda koju imamu pri generiranju ortogonalnih vektora je biranje predznaka novo generiranog vektora, ali to smo fiksirali zahtijevom $r_{jj} > 0$. Q.E.D.

QR faktorizacija može poslužiti za rješavanje jednadžbi $A\mathbf{x} = \mathbf{b}$. Iz $QR\mathbf{x} = \mathbf{b}$ dobivamo ovaj postupak. Prvo treba izračunati $\mathbf{f} = Q^*\mathbf{b}$, a zatim riješiti sustav s gornjom trokutastom matricom $R\mathbf{x} = \mathbf{f}$.

Dva su osnovna načina kako se računa QR faktorizacija: jedan je pomoću Housholderovih transformacija, a drugi pomoću Givensonovih rotacija.

1.3 Householderova triangulacija

Računanju QR faktorizacije možemo pristupiti kao problemu svođenja matrice na trokutast formu pomoću ortogonalnih tranformacija. Preciznije, treba naći n ortogonalnih (unitarnih) matrica $Q_1, \ldots, Q_n$ takvih da je

$$Q_n \cdots Q_2 Q_1 A = R,$$

gdje je R gornja trokutasta matrica. Tada je $A = QR$ pri čemu je $Q^* = Q_n \cdots Q_2 Q_1$.

Svaka ortogonalna matrica kojom množimo matricu A ima zadatak (slično kao u Gaussovim eliminacijama) poništiti elemente jednog stupca matrice ispod dijagonale: matrica Q_i će poništavati elemente i-tog stupca,

Taj proces možemo zamisliti na sljedeći način:

$$\underset{A}{\begin{bmatrix} \mathsf{x} & \mathsf{x} & \mathsf{x} \\ \mathsf{x} & \mathsf{x} & \mathsf{x} \\ \mathsf{x} & \mathsf{x} & \mathsf{x} \\ \mathsf{x} & \mathsf{x} & \mathsf{x} \\ \mathsf{x} & \mathsf{x} & \mathsf{x} \end{bmatrix}} \xrightarrow{Q_1} \underset{Q_1A}{\begin{bmatrix} \mathbf{x} & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \end{bmatrix}} \xrightarrow{Q_2} \underset{Q_2Q_1A}{\begin{bmatrix} \mathbf{x} & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \\ 0 & 0 & \mathbf{x} \\ 0 & 0 & \mathbf{x} \\ 0 & 0 & \mathbf{x} \end{bmatrix}} \xrightarrow{Q_3} \underset{Q_3Q_2Q_1A}{\begin{bmatrix} \mathbf{x} & \mathbf{x} & \mathbf{x} \\ 0 & \mathbf{x} & \mathbf{x} \\ 0 & 0 & \mathbf{x} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}$$

Matrice Q_i su dimenzije $m \times m$, a matrica R je dimenzije $m \times n$, kao i A (potpuna QR faktorizacija). Iz gornje slike možemo uočiti da Q_i mora djelovati samo na dijelu matrice u kojem još nisu poništeni elementi i ne smije "dirati" već generirane nule. Stoga Q_i mora ostavljati na miru gornju lijevu podmatricu dimenzije $(i-1) \times (i-1)$ što znači da mora imati odgovarajuću blok strukturu. U gornjem primjeru Q_2 i Q_3 moraju biti oblika

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \\ 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \\ 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \\ 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} & \mathsf{x} \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} \\ 0 & 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} \\ 0 & 0 & \mathsf{x} & \mathsf{x} & \mathsf{x} \end{bmatrix}.$$

Matrica Q_2, pri množenju s lijeva, ostavlja na miru prvi redak i nule u prvom stupcu. Matrica Q_3 ostavlja na miru prva dva retka i elemente prva dva stupca.

Općenito će matrica Q_k imati oblik

$$Q_i = \begin{bmatrix} I & 0 \\ 0 & F \end{bmatrix}$$

gdje je I jedinična matrica dimenzije $(i-1) \times (i-1)$, a F ortogonalna (unitarna) matrica dimenzije $(m-i+1) \times (m-i+1)$. Očito je tada i cijela matrica ortogonalna (unitarna).

Matrica F djeluje samo na podmatrici dimenzije $(m-i+1) \times (m-i+1)$ i njen zadatak je poništiti elemente prvog stupca podmatrice ispod dijagonalnog (dakle prvog). Efekt tog preslikavanja je

$$\mathbf{x} = \begin{bmatrix} \mathsf{x} \\ \mathsf{x} \\ \mathsf{x} \\ \mathsf{x} \\ \mathsf{x} \end{bmatrix} \xrightarrow{F} \begin{bmatrix} \pm\|\mathbf{x}\| \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \pm\|\mathbf{x}\|\mathbf{e}_1.$$

Uočimo da zbog ortogonalnosti matrice norma vektora mora ostati ista, pa je stoga prvi element transformiranog vektora jednak normi originalnog. Predznak možemo po volji odabrati. U kompleksnom slučaju na mjestu predznaka $\pm$ može stajati bilo koji kompleksan broj modula jedan (kompleksan predznak).

Do matrice F koja vrši traženu transformaciju možemo doći na sljedeći način:

Uzmimo radi određenosti predznak $+$, tj. neka je $F\mathbf{x} = \|\mathbf{x}\|\mathbf{e}_1$. Pogledajmo vektor $\mathbf{v} = F\mathbf{x} - \mathbf{x} = \|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}$. Projektor P na ravninu okomitu na taj vektor je

$$P = I - \frac{\mathbf{v} \otimes \mathbf{v}}{\|\mathbf{v}\|^2}$$

gdje smo za $\mathbf{a}, \mathbf{b} \in \mathbb{R}^k$ sa $\mathbf{a} \otimes \mathbf{b} \in \mathbb{R}^{k \times k}$ označili tenzorski produkt vektora definiran formulom

$$\forall \mathbf{y} \in \mathbb{R}^k, \quad (\mathbf{a} \otimes \mathbf{b})\mathbf{y} = (\mathbf{b} \cdot \mathbf{y})\mathbf{a}.$$

U kanonskoj bazi imamo

$$(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i b_j.$$

U slučaju kompleksnog prostora imamo $(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i \overline{b}_j$.

Da bismo preslikali $\mathbf{x}$ u $\|\mathbf{x}\|\mathbf{e}_1$ trebamo umjesti projektora uzeti reflektor u odnosu na istu ravninu, tj. trebamo udvostručiti korak u smjeru projektora (vidi sliku). Time dolazimo do preslikavanja

$$F = I - 2\frac{\mathbf{v} \otimes \mathbf{v}}{\|\mathbf{v}\|^2}, \quad \mathbf{v} = \|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}.$$

Ono djeluje po formuli

$$F\mathbf{y} = \mathbf{y} - 2\frac{\mathbf{v} \cdot \mathbf{y}}{\|\mathbf{v}\|^2}\mathbf{v},$$

i lako se provjerava da je

$$\begin{aligned}
F\mathbf{x} &= \mathbf{x} - 2\frac{(\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}) \cdot \mathbf{x}}{\|\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}\|^2}(\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}) \\
&= \mathbf{x} - 2\frac{\|\mathbf{x}\|x_1 - \|\mathbf{x}\|^2}{2\|\mathbf{x}\|^2 - 2\|\mathbf{x}\|x_1}(\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}) \\
&= \mathbf{x} + (\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}) = \|\mathbf{x}\|\mathbf{e}_1.
\end{aligned}$$

Lako je provjeriti da je matrica F istovremeno simetrična i ortogonalna (hermitska i unitarna u kompleksnom slučaju), tj. (realni slučaj)

$$F = F^\tau, \quad FF^\tau = I.$$

Do istih rezultata dolazimo ako uzmemo da je $F\mathbf{x} = -\|\mathbf{x}\|\mathbf{e}_1$. U tom slučaju treba uzeti $\mathbf{v} = -\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}$. Pitanje je koji je od ta dva izbora bolji, odnosno koji je od ta dva izbora stabilniji u odnosu na greške zaokruživanja?

Mi ne želimo da vektor $\pm\|\mathbf{x}\|\mathbf{e}_1$ bude suviše blizak vektoru $\mathbf{x}$ budući da računamo s vektorom $\mathbf{v}$ koji je njihova razlika. Stoga ćemo uzeti predznak različit od predznaka x_1, tj.

$$\mathbf{v} = -\mathrm{sgn}(x_1)\|\mathbf{x}\|\mathbf{e}_1 - \mathbf{x}$$

pa je

$$F\mathbf{x} = -\mathrm{sgn}(x_1)\|\mathbf{x}\|\mathbf{e}_1.$$

Pogledajmo sada kako izgleda i-ti korak Householderove triangulacije. Matrica Q_i s kojom množimo matricu A ima oblik

$$Q_i = \begin{bmatrix} I_{i-1} & 0 \\ 0 & I_{m-i+1} - \mathbf{v}^i \otimes \mathbf{v}^i / c_i \end{bmatrix}$$

gdje je I_j jedinična matrica dimenzije $j \times j$, $\mathbf{v}_i \in \mathbb{R}^{m-j+1}$ vektor dan formulom

$$\mathbf{v}^i = \begin{bmatrix} \mathrm{sgn}(a_{ii})\|\hat{\mathbf{a}}_i\| + a_{ii} \\ a_{i+1i} \\ \vdots \\ a_{mi} \end{bmatrix}, \quad \hat{a}_i = \begin{bmatrix} a_{ii} \\ a_{i+1i} \\ \vdots \\ a_{mi} \end{bmatrix}, \quad c_i = \frac{\|\mathbf{v}^i\|^2}{2} = \|\hat{\mathbf{a}}_i\|(\|\hat{\mathbf{a}}_i\| + |a_{ii}|).$$

Primijetimo u ovim formulama da su a_{ij} elementi matrice $Q_{i-1} \cdots Q_1 A$ te da smo promijenili predznak vektora $\mathbf{v}$, što ne utječe na Householderovu transformaciju.

Nakon množenja s matricom Q_i dobivamo matricu sljedećeg oblika.

$$\begin{array}{c} \\ 1 \\ \\ \\ i \\ \\ \\ \\ m \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & & & i & & j & & n \end{array} \\ \begin{pmatrix} x & \cdots & x & x & \cdots & x & \cdots & x \\ & \ddots & \vdots & \vdots & & \vdots & & \vdots \\ & & x & x & \cdots & x & \cdots & x \\ & & & a_{ii} & \cdots & a_{ij} & \cdots & a_{in} \\ & & & 0 & \cdots & a_{\cdot j} & \cdots & a_{\cdot n} \\ & & & 0 & \cdots & a_{\cdot j} & \cdots & a_{\cdot n} \\ & & & \vdots & \cdots & \vdots & \cdots & \vdots \\ & & & 0 & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix} \end{array}$$

Elementi u recima 1 do $i-1$ označeni su x i oni se pri množenju ne mijenjaju. Isto tako se ne mijenju nule koje se nalaze ispod dijagonale u stupcima 1 do $i-1$. Kao efekt množenja javljaju se nule ispod dijagonale u i-tom stupcu. Element a_{ii} dan je formulom

$$a_{ii} = -\mathrm{sgn}(a_{ii})\|\hat{\mathbf{a}}_i\|$$

dok se elementi u donjem desnom bloku dobivaju po formulama

$$a_{kj} = a_{kj} - (\mathbf{v}^i \cdot \hat{\mathbf{a}}_j) v_k^i / c_i$$

za $k = i, \ldots, m$ i $j = i+1, \ldots, n$.

Faktore Q_i nije potrebno formirati. Dovoljno je pamtiti vektore $\mathbf{v}^i$ i brojeve c_i. Tada je za svaki vektor $\mathbf{x} \in \mathbb{R}^m$

$$Q_i\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_i - (\mathbf{v}^i \cdot \mathbf{x})v_i^i/c_i \\ \vdots \\ x_m - (\mathbf{v}^i \cdot \mathbf{x})v_m^i/c_i \end{bmatrix}.$$

Transponirana matrica Q_i^τ (adjungirana Q_i^*) jednaka je Q_i.

Algoritam ćemo organizirati tako da transformiramo ulaznu matricu dok u gornjem trokutu ne dobijemo matricu R. Vektore $\mathbf{q}^i$ ćemo pamtiti u stupcima matrice A, analogno kao kod Gaussovih eliminacija: i-ti vektor pamtimo u i-tom stupcu, ispod dijagonale. Pri tome nam nedostaje jedno mjesto pa ćemo uvesti poseban vektor $\mathbf{d}$ u kojem ćemo pamtiti dijagonalu matrice R. Na izlazu će donji trokut matrice (uključivši dijagonalu) sardžavati vektore $\mathbf{q}^i$ dok će strogi gornji trokut sadržavati matricu R bez dijagonale, koja je u vektoru $\mathbf{d}$. Brojeve c_i pamtit ćemo u vektoru $\mathbf{c}$. Rezultat je dan u Algoritmu 3.

Nakon QR faktorizacije potrebni su nam algoritmi koji će računti produkte $Q\mathbf{x}$ i $Q^\tau\mathbf{x}$. Algoritam faktorizacije daje nam rastav

$$Q_n \cdots Q_2Q_1A = R,$$

pa je stoga $Q^\tau = Q_n \cdots Q_2Q_1$. Zbog $Q_i^\tau = Q_i$ imamo $Q = Q_1Q_2 \cdots Q_n$ pa vidimo da će množenje s $Q\mathbf{x}$ i $Q^\tau\mathbf{x}$ biti realizirano s algoritmima u kojima se razlikuje samo poredak transformacija.

Algoritam mnozenja $Q^\tau\mathbf{x}$ ima oblik

for $i = 1, 2, \ldots, n$ **do**
  for $j = i, \ldots, m$ **do**
    $x_j = x_j - (\mathbf{v}^i \cdot \mathbf{x})v_j^i/c_i$
  end for
end for

dok algoritam za $Q\mathbf{x}$ ima oblik

for $i = n, n-1, , \ldots, 1$ **do**
  for $j = i, \ldots, m$ **do**
    $x_j = x_j - (\mathbf{v}^i \cdot \mathbf{x})v_j^i/c_i$
  end for
end for

Uzimajući u obzir kako su sačuvani vektori $\mathbf{v}^i$ nakon QR faktorizacije dobivamo Algoritam 4 i Algoritam 5.

Algorithm 3 QR faktorizacija Householderovim transformacijama

Ulaz: m = broj redaka matrice,
n = broj stupaca matrice $n \leq m$,
$A = (a_{ij})$ matrica reda $m \times n$
// Po svim stupcima matrice
for $i = 1, 2, \ldots, n$ **do**
  $\sigma = 0$
  for $j = i, \ldots, m$ **do**
    $\sigma = \sigma + a_{ij}^2$
  end for
  $\sigma = \mathrm{sqrt}(\sigma)\mathrm{sgn}(a_{ii})$
  $a_{ii} = a_{ii} + \sigma$
  $d_i = -\sigma$
  $c_i = \sigma a_{ii}$
  // po stupcima desno od i-tog
  for $j = i + 1, \ldots, n$ **do**
    $\gamma = 0$ // skalarni produkt stupaca
    for $k = i, \ldots, m$ **do**
      $\gamma = \gamma + a_{ki} a_{kj}$
    end for
    // primjena Householderove refleksije
    for $k = i, \ldots, m$ **do**
      $a_{kj} = a_{kj} - \gamma a_{ki} / c_i$
    end for
  end for
end for
Izlaz: Matrica A, vektori **d** i **c**.

Algorithm 4 Množenje $Q^\tau \mathbf{x}$ nakon QR faktorizacije

Ulaz: m = broj redaka matrice,
n = broj stupaca matrice $n \leq m$,
$A = (a_{ij})$ matrica reda $m \times n$ nakon QR faktorizacije
$\mathbf{c} = (c_i)$ vektor dobiven QR faktorizacijom

```
for i = 1, 2, ..., n do
  σ = 0
  for j = i, ..., m do
    σ = σ + A_ji x_j
  end for
  for j = i, ..., m do
    x_j = x_j − σ A_ji / c_i
  end for
end for
```

Izlaz: vektor $\mathbf{x}$.

Algorithm 5 Množenje $Q\mathbf{x}$ nakon QR faktorizacije

Ulaz: m = broj redaka matrice,
n = broj stupaca matrice $n \leq m$,
$A = (a_{ij})$ matrica reda $m \times n$ nakon QR faktorizacije
$\mathbf{c} = (c_i)$ vektor dobiven QR faktorizacijom

```
for i = n, n − 1, ..., 1 do
  σ = 0
  for j = i, ..., m do
    σ = σ + A_ji x_j
  end for
  for j = i, ..., m do
    x_j = x_j − σ A_ji / c_i
  end for
end for
```

Izlaz: vektor $\mathbf{x}$.

1.4 Singularne vrijednosti matrice

Svaka (realna ili) kompleksna matrica $A \in \mathbb{C}^{m\times n}$ predstavlja jedan linearan operator iz $\mathbb{C}^n$ u $\mathbb{C}^m$. Prepostavimo da u prostoru $\mathbb{C}^n$ imamo jednu ortonormiranu bazu $\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$, a u $\mathbb{C}^m$ neku drugu ortonormiranu bazu $\{\mathbf{u}^1, \ldots, \mathbf{u}^m\}$. Pomoću te dvije baze možemo formirati matrice ranga jedan $\mathbf{u}^j \otimes \mathbf{v}^i : \mathbb{C}^n \to \mathbb{C}^m$, definirane formulom

$$(\mathbf{u}^j \otimes \mathbf{v}^i)_{k,l} = u^j_k \overline{v^i_l},$$

odnosno, za svaki vektor $\mathbf{x} \in \mathbb{C}^n$, $(\mathbf{u}^j \otimes \mathbf{v}^i)\mathbf{x} = (\mathbf{x} \cdot \mathbf{v}^i)\mathbf{u}^j$, gdje točka označava skalarni produkt[2] u $\mathbb{C}^n$. Hermitsko adjungiranje matrica vrši se jednostavnom zamjenom poretka vektora:

$$(\mathbf{u}^j \otimes \mathbf{v}^i)^H = (\mathbf{v}^i \otimes \mathbf{u}^j).$$

Općenitije, izborom brojeva $\sigma_1, \ldots, \sigma_r$, gdje je $r \le \min(n, m)$, možemo formirati matricu

$$A = \sum_{j=1}^{r} \sigma_j (\mathbf{u}^j \otimes \mathbf{v}^j) \tag{1.2}$$

Ta matrica ima sljedeća svojstva: za $k = 1, \ldots, r$

$$A\mathbf{v}^k = \sigma_k \mathbf{u}^k, \quad A^H \mathbf{u}^k = \sigma_k \mathbf{v}^k. \tag{1.3}$$

Odavde vidimo da je

$$\mathcal{N}(A) = \mathcal{L}(\mathbf{v}^{r+1}, \ldots, \mathbf{v}^n), \quad \mathcal{R}(A) = \mathcal{L}(\mathbf{u}^1, \ldots, \mathbf{u}^r)$$
$$\mathcal{R}(A^H) = \mathcal{L}(\mathbf{v}^1, \ldots, \mathbf{v}^r), \quad \mathcal{N}(A^H) = \mathcal{L}(\mathbf{u}^{r+1}, \ldots, \mathbf{u}^m)$$

i dobivamo sljedeće odnose:

$$\mathcal{N}(A)^\perp = \mathcal{R}(A^H), \quad \mathcal{R}(A)^\perp = \mathcal{N}(A^H).$$

Nadalje, matrice $AA^H \in \mathbb{C}^{m\times m}$, $A^H A \in \mathbb{C}^{n\times n}$ imaju svojstvo

$$A^H A \mathbf{v}^k = \sigma_k^2 \mathbf{v}^k, \quad AA^H \mathbf{u}^k = \sigma_k^2 \mathbf{u}^k, \tag{1.4}$$

za $k = 1, \ldots, r$. Dakle, σ_k^2 su svojstvene vrijednosti matrica $A^H A$ i AA^H, a $\mathbf{v}^k$ i $\mathbf{u}^k$ su pripadni svojstveni vektori.

Prirodno je sada postaviti pitanje da li svaku matricu $A \in \mathbb{C}^{m\times n}$ možemo napisati u obliku (1.2), odnosno možemo li naći ortogonalne vektore $\mathbf{v}^k$ i $\mathbf{u}^k$ te realne brojeve σ_k takve da (1.2) vrijedi. Iz jednakosti (1.4) vidimo kako bismo trebali definirati vektore. Matrice $AA^H \in \mathbb{C}^{m\times m}$, $A^H A \in \mathbb{C}^{n\times n}$ su hermitske i zbog

$$AA^H \mathbf{u} \cdot \mathbf{u} = \|A^H \mathbf{u}\|_2^2 \ge 0, \quad A^H A \mathbf{u} \cdot \mathbf{u} = \|A\mathbf{u}\|_2^2 \ge 0,$$

[2]Skalarni produkt je linearan u prvom argumentu i antilinearan u drugom.

pozitivno semidefinitne. Stoga obje matrice imaju ortonormiranu bazu svojstvenih vektora, a pripadne svojstvene vrijednosti su im realne i nenegativne.

Označimo svojstvene vektore matrice $A^H A$ sa $\mathbf{v}^k \in \mathbb{C}^n$, $k = 1, \dots, n$, a pripadne svojstvene vrijednsti sa σ_k^2, $k = 1, \dots, n$, što možemo zbog nenegativnosti svojstvenih vrijednosti. Radi određenosti indeksirajmo svojstvene parove tako da bude[3]

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0, \tag{1.5}$$

pri čemu evidentno svaku svojstvenu vrijednost ponavljamo onoliko puta kolika joj je kratnost. Uočimo da je

$$\|A\mathbf{v}^k\|_2^2 = A\mathbf{v}^k \cdot A\mathbf{v}^k = \mathbf{v}^k \cdot A^H A\mathbf{v}^k = \sigma_k^2 \mathbf{v}^k \cdot \mathbf{v}^k = \sigma_k^2 \tag{1.6}$$

i stoga je

$$A\mathbf{v}^k = 0 \quad \text{za } k = r+1, \dots, n. \tag{1.7}$$

Definirajmo vektore

$$\mathbf{u}^k = \frac{1}{\sigma_k} A\mathbf{v}^k \in \mathbb{C}^m, \quad k = 1, \dots, r. \tag{1.8}$$

Zbog

$$\mathbf{u}^k \cdot \mathbf{u}^l = \frac{1}{\sigma_k \sigma_l} A\mathbf{v}^k \cdot A\mathbf{v}^l = \frac{1}{\sigma_k \sigma_l} A^H A\mathbf{v}^k \cdot \mathbf{v}^l = \frac{\sigma_k}{\sigma_l} \mathbf{v}^k \cdot \mathbf{v}^l = \frac{\sigma_k}{\sigma_l} \delta_{k,l}$$

oni čine ortonormiran skup vektora. Time su i linearno nezavisni pa je stoga sigurno $r \leq m$. Dopunimo skup $\{\mathbf{u}^1, \dots, \mathbf{u}^r\}$ na neki način s vektorima $\mathbf{u}^{r+1}, \dots, \mathbf{u}^m$ do ortonormirane baze u $\mathbb{C}^m$. Na taj način definirani vektori $\mathbf{u}^k$ čine svojstvene vektore matrice AA^H, sa svojstvenim vrijednostima σ_k^2. Zaista, za $1 \leq k \leq r$ imamo

$$AA^H \mathbf{u}^k = \frac{1}{\sigma_k} A(A^H A)\mathbf{v}^k = \sigma_k A\mathbf{v}^k = \sigma_k^2 \mathbf{u}^k.$$

Za $r+1 \leq k \leq m$ i za proizvoljno $\mathbf{v}^j \in \mathbb{C}^n$, $j = 1, \dots, r$, imamo

$$A^H \mathbf{u}^k \cdot \mathbf{v}^j = \mathbf{u}^k \cdot A\mathbf{v}^j = \sigma_j \mathbf{u}^k \cdot \mathbf{u}^j = 0$$

zbog ortogonalnosti, odnosno $A^H \mathbf{u}^k \perp \mathcal{L}(\mathbf{v}^1, \dots, \mathbf{v}^r)$. Prema tome vidimo da vrijedi inkluzija $A^H \mathbf{u}^k \subset \mathcal{L}(\mathbf{v}^{r+1}, \dots, \mathbf{v}^n)$, pa prema (1.7) imamo $AA^H \mathbf{u}^k = 0$ za sve $r+1 \leq k \leq m$.

Proizvoljan $\mathbf{x} \in \mathbb{C}^n$ možemo raspisati po vektorima baze:

$$\mathbf{x} = \sum_{j=1}^{n} (\mathbf{x} \cdot \mathbf{v}^j) \mathbf{v}^j.$$

[3] σ_k su pozitivni korijeni svojstvenih vrijednosti.

Imamo

$$A\mathbf{x} = \sum_{j=1}^{r} (\mathbf{x} \cdot \mathbf{v}^j) A\mathbf{v}^j = \sum_{j=1}^{r} \sigma_j (\mathbf{x} \cdot \mathbf{v}^j) \mathbf{u}^j = \sum_{j=1}^{r} \sigma_j (\mathbf{u}^j \otimes \mathbf{v}^j) \mathbf{x}.$$

Uočimo još da su uz uvjet (1.5) brojevi σ_i jedistveno određeni.

Teorem 1.3 Za svaku matricu $A \in \mathbb{C}^{m\times n}$ postoji ortonormirana baza $\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ u $\mathbb{C}^n$ i ortonormirana baza $\{\mathbf{u}^1, \ldots, \mathbf{u}^m\}$ u $\mathbb{C}^m$ te realni brojevi (1.5) takvi da vrijedi (1.2). Brojeve σ_i nazivamo singularnim vrijednostima matrice A, a vektore $\mathbf{v}^i$ i $\mathbf{u}^j$ desnim, odn. lijevim singularnim vektorima. Singularne vrijednosti su jedinstvene.

Formula (1.2) inducira faktorizaciju matrice A koju nazivamo SVD (eng. *singular-value decomposition*). Uvest ćemo matrice

$$U = [\mathbf{u}^1 \cdots \mathbf{u}^m] \in \mathbb{C}^{m\times m}, \quad V = [\mathbf{v}^1 \cdots \mathbf{v}^n] \in \mathbb{C}^{n\times n},$$
$$\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots) \in \mathbb{R}^{m\times n},$$

gdje vektori predstavljaju stupce matrice. Σ je pravokutna dijagonalna matrica koja na glavnoj dijagonali ima vrijednosti σ_1 do σ_r, a ostatak glavne dijagonale su nule. Sada

$$A\mathbf{v}^k = \sigma_k \mathbf{u}^k, \quad k = 1, \ldots, r,$$

dobiva matrični oblik $AV = U\Sigma$, ili

$$A = U\Sigma V^H. \tag{1.9}$$

To je SVD faktorizacija. U njoj se uvijek poredak stupaca u matricama U i V definira tako da vrijedi (1.5). SVD ima i svoj reducirani oblik u kojem uzimamo matricu $\hat{U} \in \mathbb{C}^{m\times r}$ koja se sastoji od prvih r stupaca matrice U te kvadratnu matricu $\hat{\Sigma} \in \mathbb{R}^{r\times r}$, koja ima samo pozitivne singularne vrijednosti, i matricu $\hat{V} \in \mathbb{C}^{n\times r}$ koja se sastoji od prvih r stupaca matrice V:

$$A = \hat{U}\hat{\Sigma}\hat{V}^H.$$

Teorem 1.4 Neka je σ_1 najveća singularna vrijednost matrice $A \in \mathbb{C}^{m\times n}$. Tada je

$$\|A\|_2 = \sigma_1.$$

Dokaz. Matrice U i V su unitarne pa imamo

$$\begin{aligned}
\|A\|_2 &= \sup\{\|A\mathbf{x}\|_2 \colon \|\mathbf{x}\|_2 = 1\} = \sup\{\|U\Sigma V^H\mathbf{x}\|_2 \colon \|\mathbf{x}\|_2 = 1\} \\
&= \sup\{\|\Sigma V^H\mathbf{x}\|_2 \colon \|\mathbf{x}\|_2 = 1\} = \sup\{\|\Sigma\mathbf{y}\|_2 \colon \|V\mathbf{y}\|_2 = 1\} \\
&= \sup\{\|\Sigma\mathbf{y}\|_2 \colon \|\mathbf{y}\|_2 = 1\} \\
&= \sup\{\sqrt{\sigma_1^2 y_1^2 + \cdots + \sigma_r^2 y_r^2} \colon \|\mathbf{y}\|_2 = 1\} \\
&= \max\{|\sigma_i| \colon i = 1, \ldots, r\} = \sigma_1.
\end{aligned}$$

□

Važnost singularne dekompozicije leži u jednostavnosti kojom se matrica može aproksimirati jednostavnijim matricama. Ideja je da se odbace sve singularne vrijednosti koje su manje od nekog zadanog praga. Tada se mogu odbaciti i pripadni lijevi i desni singularni vektori i dobiva se aproksimacija nižeg ranga. Na primjer, ako matricu

$$A = \sum_{j=1}^{r} \sigma_j (\mathbf{u}^j \otimes \mathbf{v}^j)$$

aproksimiramo s

$$A_p = \sum_{j=1}^{p} \sigma_j (\mathbf{u}^j \otimes \mathbf{v}^j)$$

za $1 \leq p < r$, dobivamo prema Teoremu 1.4 grešku

$$\|A - A_p\|_2 = \sigma_{p+1}.$$

Pomoću singularne dekompozicije možemo definirati pseudo-inverz matrice. Ako je $A = U\Sigma V^H$ onda je prirodno pseudo-inverz definirati formulom

$$A^+ = V\Sigma^+ U^H,$$

gdje je $\Sigma^+ = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots)$. U slučaju da je matrica $A \in \mathbb{C}^{n \times n}$ kvadratna i regularna ($r = n$), onda je lako vidjeti da je $A^+ = A^{-1}$. U drugim slučajevima predstavlja generalizaciju pojma inverza koja je korisna u metodi najmanjih kvadrata.

1.5 Linearan problem najmanjih kvadrata

Zadana je matrica $A = (a_{ij})$ dimenzije $m \times n$ gdje je $m \geq n$ i vektor $\mathbf{b} = (b_i) \in \mathbb{R}^m$. Za matricu A, radi jednostavnosti, pretpostavljamo da je punog ranga.

Treba naći vektor $\mathbf{x} \in \mathbb{R}^n$ za koji je

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \min_{\mathbf{y} \in \mathbb{R}^n} \|A\mathbf{y} - \mathbf{b}\|_2. \tag{1.10}$$

Taj problem nazivamo linearnim problemom najmanjih kvadrata. Ako je $\mathbf{x}$ rješenje te zadaće, onda je i $\mathbf{x} + \mathbf{y}$ također jedno rješenje za svako $\mathbf{y} \in \mathcal{N}(A)$ iz jezgre matrice A.

Uočimo da je funkcija

$$\mathbf{y} \mapsto f(\mathbf{y}) = \|A\mathbf{y} - \mathbf{b}\|_2 \geq 0,$$

neprekidna, odozdo ograničena na $\mathbb{R}^n$. Nadalje ona je dobro definirana i na kvocijentnom prostoru $\mathbb{R}^n/\mathcal{N}(A)$, čiji su elementi klase ekvivalencije $[\mathbf{x}] = \mathbf{x} + \mathcal{N}(A)$, $\|[\mathbf{x}]\|_2 = \inf\{\|\mathbf{x}+\mathbf{y}\|_2 \colon \mathbf{y} \in \mathcal{N}(A)\}$, i na njemu je neprekidna i korercitivna, što znači da vrijedi

$$\lim_{\|[\mathbf{y}]\|_2 \to \infty} \|A\mathbf{y} - \mathbf{b}\|_2 = +\infty.$$

Budući da neprekidna funkcija na kompaktu postiže svoj minimum i maksimum, zaključujemo da funkcija $f([\mathbf{y}])$ postiže svoj minimum na nekoj dovoljno velikoj kugli $\overline{K(0,R)}$ u kvocijentnom prostoru. Time je egzistencija rješenja linearnog problema najmanjih kvadrata dokazan.

Skup svih rješenja karakteriziran je sljedećom lemom.

Lema 1.2 Neka je

$$S = \{\mathbf{x} \in \mathbb{R}^m \colon \|A\mathbf{x} - \mathbf{b}\|_2 = \min\}.$$

Tada je

$$\mathbf{x} \in S \quad \Leftrightarrow \quad A^\tau(\mathbf{b} - A\mathbf{x}) = 0.$$

Dokaz. Neka je $A^\tau(\mathbf{b} - A\mathbf{x}) = 0$. Tada je za svako $\mathbf{y} \in \mathbb{R}^n$

$$\mathbf{b} - A\mathbf{y} = \mathbf{b} - A\mathbf{x} + A(\mathbf{y} - \mathbf{x}),$$

$$\begin{aligned} \|\mathbf{b} - A\mathbf{y}\|_2^2 &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \|A(\mathbf{y} - \mathbf{x})\|_2^2 + 2(\mathbf{b} - A\mathbf{x}) \cdot A(\mathbf{y} - \mathbf{x}) \\ &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \|A(\mathbf{y} - \mathbf{x})\|_2^2 + 2A^\tau(\mathbf{b} - A\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \\ &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \|A(\mathbf{y} - \mathbf{x})\|_2^2 \end{aligned}$$

i stoga je za svako $\mathbf{y} \in \mathbb{R}^m$ vrijedi

$$\|\mathbf{b} - A\mathbf{x}\|_2 \leq \|\mathbf{b} - A\mathbf{y}\|_2.$$

Time je jedan smjer dokazan. Obrat. Pretpostavimo da je

$$A^\tau(\mathbf{b} - A\mathbf{x}) = \mathbf{z} \neq 0.$$

Uzmimo $\mathbf{y} = \mathbf{x} + \varepsilon\mathbf{z}$.

$$\begin{aligned} \|\mathbf{b} - A\mathbf{y}\|_2^2 &= \|\mathbf{b} - A\mathbf{x} - \varepsilon A\mathbf{z}\|_2^2 \\ &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \varepsilon^2\|A\mathbf{z}\|_2^2 - 2\varepsilon(\mathbf{b} - A\mathbf{x}) \cdot A\mathbf{z} \\ &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \varepsilon^2\|A\mathbf{z}\|_2^2 - 2\varepsilon A^\tau(\mathbf{b} - A\mathbf{x}) \cdot \mathbf{z} \\ &= \|\mathbf{b} - A\mathbf{x}\|_2^2 + \varepsilon^2\|A\mathbf{z}\|_2^2 - 2\varepsilon\|\mathbf{z}\|_2^2. \end{aligned}$$

Za $\varepsilon > 0$ dovoljno mali bit će

$$\varepsilon^2\|A\mathbf{z}\|_2^2 - 2\varepsilon\|\mathbf{z}\|_2^2 < 0$$

što daje

$$\|\mathbf{b} - A\mathbf{y}\|_2 < \|\mathbf{b} - A\mathbf{x}\|_2$$

odnosno, $\mathbf{x}$ nije rješenje problema minimizacije. Q.E.D.

Jednadžbu

$$A^\tau A\mathbf{x} = A^\tau \mathbf{b}$$

nazivamo normalna jednadžba. To je jednadžba koju zadovoljavaju stacionarne točke funkcije

$$f(\mathbf{x}) = \frac{1}{2}(A\mathbf{x} - \mathbf{b}) \cdot (A\mathbf{x} - \mathbf{b}).$$

Zaista, deriviranjem dobivamo

$$\nabla f(\mathbf{x}) = A^\tau (A\mathbf{x} - \mathbf{b}).$$

Uočimo da je normalna jednadžba linearan sustav dimenzije $n \times n$ koji je uvijek konsistentan. Naime, za konsistentnost sustava potrebno je da je desna strana u slici operatora, što se za normalnu jednadžbu svodi na $A^\tau \mathbf{b} \in \mathcal{R}(A^\tau A)$, no to je evidentno zadovoljeno na osnovu sljedećeg rezultata iz linearne algebre: kodomena linearnog operatora može prikazati na sljedeći način:

$$\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^\tau).$$

Odavde vidimo da rješenje problema najmanjih kvadrata daje upravo taj rastav. Imamo

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

pri čemu je $\mathbf{b} - A\mathbf{x} \in \mathcal{N}(A^\tau)$ i $A\mathbf{x} \in \mathcal{R}(A)$. Time je ujedno i dokazana egzistencija rješenja problema najmanjih kvadrata.

Lema 1.3 Matrica $A^\tau A$ je pozitivno definitna ako i samo ako su stupci matrice A linearno nezavisni, tj. rang$(A) = n \leq m$.

Dokaz. Ako su stupci linearno nezavisni, onda za svako $\mathbf{x} \in \mathbb{R}^n$ vrijedi

$$\mathbf{x} \neq 0 \quad \Rightarrow \quad A\mathbf{x} \neq 0 \quad \Rightarrow \quad \|A\mathbf{x}\|_2^2 > 0 \quad \Leftrightarrow \quad \mathbf{x} \cdot A^\tau A\mathbf{x} > 0.$$

Obrat. Ako stupci nisu linearno nezavisni, onda postoju neki $\mathbf{x}^0$ za koji je $A\mathbf{x}^0 = 0$ pa kao ogore zaključujemo da je $\mathbf{x}^0 \cdot A^\tau A\mathbf{x}^0 = 0$ odnosno, matrica $A^\tau A$ nije pozitivno definitna. Q.E.D.

Normalnu jednadžbu moguće je rješavati na različite načine. Naizravnija mogućnost je putem Gaussovih eliminacija. Primjenom Gaussovih eleiminacija na matricu $A^\tau A$ dobiva se tzv. faktorizacija Choleskog:

$$A^\tau A = R^\tau R$$

gdje je $R \in \mathbb{R}^{n \times n}$ gornja trokutasta matrica s pozitivnom dijagonalom. Tada je potrebno riješiti sustav

$$R^\tau R\mathbf{x} = A^\tau \mathbf{b}$$

što se svodi na rješavanje dva sustava s trokutastom matricom:

$$R^\tau \mathbf{y} = A^\tau \mathbf{b}, \quad R\mathbf{x} = \mathbf{y}.$$

Ovdje treba primijetiti da je u realnim slučajevima $m >> n$ i da je sustav koji treba riješiti relativno male domenzije $(n \times n)$. S druge strane, potrebno je izvršiti n^2 skalarnih produkata u $\mathbb{R}^m$ da bi se formirala matrica $A^\tau A$.

Druga mogućnost je koristiti QR faktorizaciju. Uzmimo. dakle, da matrica $A \in \mathbb{R}^{m \times n}$ ima potpunu QR faktorizaciju

$$A = Q \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix},$$

gdje je $Q \in \mathbb{R}^{m \times m}$ ortogonalna matrica te

$$\begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \in \mathbb{R}^{m \times n}, \quad \hat{R} \in \mathbb{R}^{n \times n} \text{ gornja trokutasta matrica.}$$

Sada zbog ortogonalnosti matrice Q dobivamo

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|Q \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \mathbf{x} - \mathbf{b}\|_2 = \|Q(\begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \mathbf{x} - Q^\tau \mathbf{b})\|_2 = \| \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \mathbf{x} - Q^\tau \mathbf{b}\|_2.$$

Vidimo da prvo treba izračunati vektor

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix} = Q^\tau \mathbf{b}, \quad \mathbf{d}_1 \in \mathbb{R}^n, \ \ \mathbf{d}_2 \in \mathbb{R}^{m-n},$$

a zatim treba riješiti sustav

$$\hat{R}\mathbf{x} = \mathbf{d}_1.$$

Vrijednost ciljne funkcije u točki minimuma je

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{d}_2\|_2.$$

Ovaj algoritam smo izveli uz pretpostavku da imamo potpunu QR faktorizaciju. Ako imamo samo parcijalnu QR faktorizaciju, dakle matricu $\hat{R}$ i prvih n stupaca matrice Q, što ćemo označiti s $\hat{Q} \in \mathbb{R}^{m \times n}$, onda vidimo da treba računati

$$\mathbf{d} = \hat{Q}^\tau \mathbf{b}, \quad \hat{R}\mathbf{x} = \mathbf{d}.$$

Time dobivamo sljedeći algoritam za rejšavanje linearnog problema najmanjih kvadrata pomoću QR faktorizacije:

1: Izračunati (reduciranu) QR faktorizaciju $A = \hat{Q}\hat{R}$,
2: Izračunati $\mathbf{d} = \hat{Q}^\tau \mathbf{b}$,
3: Riješiti sustav $\hat{R}\mathbf{x} = \mathbf{d}$.

Analogno kao pomoću QR faktorizacije, problem najmanjih kvadrata možemo riješiti pomoću SVD faktorizacije. Pri tome će nam biti dovoljna reducirana faktorizacija. Uzmimo da imamo SVD faktorizaciju $A = U\Sigma V^\tau$ realne matrice A. Tada je zbog ortogonalnosti matrice U:

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|U\Sigma V^\tau \mathbf{x} - \mathbf{b}\|_2 = \|U(\Sigma V^\tau \mathbf{x} - U^\tau \mathbf{b})\|_2 = \|\Sigma V^\tau \mathbf{x} - U^\tau \mathbf{b}\|_2.$$

Uvedimo oznaku $\mathbf{y} = V^\tau \mathbf{x} \in \mathbb{R}^n$ te reducirane faktore $\hat{U} \in \mathbb{R}^{m\times r}$ i $\hat{\Sigma} \in \mathbb{R}^{r\times n}$. Vektor $U^\tau \mathbf{b}$ može se zapisati u obliku

$$U^\tau \mathbf{b} = \begin{bmatrix} \hat{U}^\tau \mathbf{b} \\ \mathbf{d}_2 \end{bmatrix}$$

gdje $\hat{U}^\tau \mathbf{b}$ predstavlja prvih r komponenti vektora, a $\mathbf{d}_2$ preostalih $m - r$ komponenti. Sada je

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|\hat{\Sigma}\mathbf{y} - \hat{U}^\tau \mathbf{b}\|_2^2 + \|\mathbf{d}_2\|_2^2.$$

Rješenje problema najmanjih kvadrata sada dobivamo sljedećom procedurom:

1: Izračunati (reduciranu) SVD faktorizaciju $A = \hat{U}\hat{\Sigma}V^\tau$,
2: Izračunati $\mathbf{d} = \hat{U}^\tau \mathbf{b}$,
3: Riješiti sustav $\hat{\Sigma}\mathbf{y} = \mathbf{d}$,
4: Izračunati $\mathbf{x} = V\mathbf{y}$.

Uočimo da je *rješavanje sustava* trivijalno.