

Bor Plestenjak

Nelinearni problemi lastnih vrednosti

skripta

verzija: 14. april 2010

Kazalo

1	Nelinearni problem lastnih vrednosti	5
1.1	Uvod	5
1.2	Posebne lastnosti	6
1.3	Pomožni rezultati	8
1.4	Algebrajska in geometrijska večkratnost lastne vrednosti	9
1.5	Občutljivost enostavne lastne vrednosti	11
1.6	Rayleighov funkcional	14
1.7	Ortogonalnost lastnih vektorjev	20
2	Numerične metode za polne probleme	22
2.1	Newtonova metoda in inverzna iteracija	22
2.2	Inverzna iteracija z residuali	25
2.3	Zaporedne linearne aproksimacije	26
2.4	Dvostranske različice inverzne iteracije	27
2.5	QR metoda Kublanovskaje	28
2.6	LU metoda Bohteta	29
2.7	Zajamčena iteracija	30
3	Polinomski (kvadratni) problem lastnih vrednosti	33
3.1	Uvod	33
3.2	Povezava z diferencialnimi enačbami	35
3.3	Smithova forma	37
3.4	Linearizacija	38
3.5	Kvadratni problem lastnih vrednosti	40
3.6	Hiperboličen in nadkritično dušen QEP	41
3.7	Skaliranje in izbira optimalne linearizacije	44
3.8	Numerične metode za polne QEP	46
3.8.1	Simetrične in definitne linearizacije	47
3.8.2	Tridiagonalne matrike	47
3.8.3	Žiroskopski problemi	48
3.8.4	Palindromski polinomski problemi lastnih vrednosti	50
4	Metode podprostorov za kvadratni problem lastnih vrednosti	52
4.1	Metode podprostorov	52
4.1.1	Podprostor Krilova	52
4.1.2	Arnoldijev algoritem	53
4.1.3	Galerkinov pogoj in Ritzeve vrednosti	54
4.1.4	Harmonične Ritzeve vrednosti	55
4.1.5	Lanczosev algoritem	56

4.1.6	Jacobi–Davidsonova metoda	58
4.1.7	Jacobi–Davidsonova metoda s harmoničnimi Ritzevim vrednostmi	60
4.2	Iterativne metode za QEP preko linearizacije	61
4.2.1	Arnoldijeva metoda na $A - \lambda B$	61
4.2.2	SOAR (Second-order Arnoldi)	62
4.2.3	Jacobi–Davidsonova metoda	63
4.2.4	Alternative za Rayleighov kvocient	65
4.2.5	Računanje več lastnih vrednosti	67
5	Metode podprostorov za splošni nelinearni problem	70
5.1	Uvod	70
5.2	Nelinearna Arnoldijeva metoda	71
5.3	Jacobi–Davidsonova metoda	72

Predgovor

Skripta je namenjena predmetu *Nelinearni problemi lastnih vrednosti*, ki se predava na študijski smeri Matematika doktorskega študijskega programa Matematika in fizika na Fakulteti za matematiko in fiziko Univerze v Ljubljani.

Predmet obravnava numerično reševanje nelinearnih problemov lastnih vrednosti, poseben poudarek pa je na polinomskih in kvadratnih problemih lastnih vrednosti. Obravnavane so direktne in iterativne metode, poleg numeričnih metod pa je še polno zanimivih teoretičnih rezultatov, ki bi znali biti zanimivi za študente, ki jih zanima linearna algebra.

Za lažje spremljanje predmeta in skripte je potrebno predznanje numerične linearne algebre, ki ga npr. študenti univerzitetnega študija matematike lahko pridobijo pri izbirnem predmetu *Numerična linearna algebra* v 3. letniku. Pri tem predmetu študenti spoznajo direktne metode za računanje lastnih vrednosti enostavnega problema lastnih vrednosti. Zaželeno, ni pa nujno potrebno, je tudi poznavanje osnovnih iterativnih metod za lastne vrednosti, ki jih študenti magistrskega študija matematike lahko pridobijo pri predmetu *Iterativne numerične metode v linearni algebri*.

Skripta je nastala ob prvem izvajanju predmeta v šolskem letu 2009/10.

izred. prof. dr. Bor Plestenjak

Poglavje 1

Nelinearni problem lastnih vrednosti

1.1 Uvod

Naj bo $T(\lambda)$ matrika velikosti $n \times n$, katere elementi so funkcije parametra λ . Za vsak $\lambda \in \mathbb{C}$ je torej $T(\lambda)$ kompleksna matrika iz $\mathbb{C}^{n \times n}$. Za elemente lahko predpostavimo, da so zadosti krat zvezno odvedljivi.

Če obstajata tak skalar $\lambda \in \mathbb{C}$ in neničelni vektor $x \in \mathbb{C}^n$, da je

$$T(\lambda)x = 0,$$

potem je λ *lastna vrednost*, x pa (*desni*) *lastni vektor*. Podobno je neničelni vektor $y \in \mathbb{C}^n$ *levi lastni vektor*, če je $y^*T(\lambda) = 0$. Pravimo, da je (λ, x, y) *lastna trojica* za T .

Posebni primeri nelinearnih problemov lastnih vrednosti so:

- Navadni linearni problem lastnih vrednosti (EP). Če za matriko $A \in \mathbb{C}^{n \times n}$ definiramo

$$T(\lambda) = \lambda I - A,$$

potem je očitno $T(\lambda)x = 0$ natanko tedaj, ko je $Ax = \lambda x$.

- Posplošeni problem lastnih vrednosti (GEP). Če za matriki $A, B \in \mathbb{C}^{n \times n}$ vzamemo

$$T(\lambda) = \lambda B - A,$$

potem je $T(\lambda)x = 0$ ekvivalentno $Ax = \lambda Bx$.

Pri GEP se lahko zgodi, da sta matriki A in B izbrani tako, da je $\det(T(\lambda)) \equiv 0$. V tem primeru pravimo, da je problem *singularen*. Nasprotno za problem $T(\lambda)x = 0$ pravimo, da je *regularen*, če je $\det(T(\lambda)) \not\equiv 0$. V primeru, ko je problem regularen, so lastne vrednosti rešitve karakteristične enačbe $\det(T(\lambda)) = 0$.

- Kvadratni problem lastnih vrednosti (QEP). Za matrike $M, C, K \in \mathbb{C}^{n \times n}$ definiramo

$$T(\lambda) = \lambda^2 M + \lambda C + K.$$

V primeru, ko je matrika M nesingularna, imamo $2n$ lastnih vrednosti, saj je $\det(T(\lambda))$ v tem primeru polinom stopnje $2n$ z vodilnim koeficientom $\det(M)$.

Problem je sicer nelinearen, a se da linearizirati. Ena izmed možnosti je, da ga prevedemo na GEP oblike

$$\left(\lambda \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} C & K \\ -I & 0 \end{bmatrix} \right) \begin{bmatrix} \lambda x \\ x \end{bmatrix} = 0.$$

- Polinomski problem lastnih vrednosti (PEP). Tu imamo

$$T(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \cdots + \lambda A_1 + A_0.$$

Podobno kot kvadratni problem lastnih vrednosti lahko tudi polinomski problem lastnih vrednosti prevedemo na GEP z matrikama velikosti $(mn) \times (mn)$. Če je A_m nesingularna matrika, potem ima PEP mn lastnih vrednosti.

- Racionalni problem lastnih vrednosti (REP), npr.:

$$T(\lambda) = -K + \lambda M + \sum_{k=1}^m \frac{\lambda}{\sigma_k - \lambda} C_k.$$

Obstaja več vrst REP. Vse lahko prevedemo na polinomski problem lastnih vrednosti, če izraz spravimo na skupni imenovalac. Torej bi tudi REP lahko linearizirali v GEP.

- Iracionalni problem lastnih vrednosti, npr.:

$$T(\lambda) = K - \lambda M + i \sum_{k=1}^m \sqrt{\lambda - \sigma_k^2} W_k.$$

Zgornji problem, kjer so $M, K, W_1, \dots, W_m$ simetrične matrike, pri čemer je K nenegativno, M pa pozitivno definitna, $\sigma_1, \dots, \sigma_m$ pa so nenegativni skalarji, nastopa pri razvoju linearnih pospeševalnikov.

Iracionalnega problema lastnih vrednosti ne moremo linearizirati in zanj lahko rečemo, da je pristno nelinearen.

- Pri študiju diferencialnih enačb s časovnimi zamiki nastopajo nelinearni problemi lastnih vrednosti oblike

$$T(\lambda) = -\lambda I + A_0 + \sum_{k=1}^m A_k e^{-h_k \lambda},$$

kjer so $h_1, \dots, h_m$ pozitivni premiki, $A_0, \dots, A_m$ pa so realne matrike.

1.2 Posebne lastnosti

Pri navadnem problemu lastnih vrednosti imamo polno lepih lastnosti, ki jih pri nelinearnih problemih nimamo več ali pa veljajo le ob zelo močnih predpostavkah. V tem razdelku bomo na kratko predstavili glavne razlike.

Za navadni problem lastnih vrednosti, kjer je matrika velikosti $n \times n$, vemo, da ima natanko n lastnih vrednosti (šteto z večkratnostmi vred). Za splošni nelinearni problem ne vemo nič o številu lastnih vrednosti. Lahko jih ima manj kot n , več kot n , ali pa celo neskončno.

Pri EP nam simetrična (hermitska) oblika matrike prinese realne lastne vrednosti in ortogonalno (unitarno) matriko lastnih vektorjev.

Za nelinearni problem lastnih vrednosti $T(\lambda)$ pravimo, da je:

- a) *realno simetričen*, če je $T(\lambda) \in \mathbb{R}^{n \times n}$ in $T(\lambda) = T(\lambda)^T$ za vsak $\lambda \in \mathbb{R}$;
- b) *kompleksno simetričen*, če je $T(\lambda) = T(\lambda)^T$ za vsak $\lambda \in \mathbb{C}$;
- c) *hermitski*, če je $T(\lambda)^H = T(\bar{\lambda})$ za vsak $\lambda \in \mathbb{C}$.

Medtem, ko ima hermitski EP same realne lastne vrednosti, to že pri GEP ni več res. Za splošni NEP lahko ugotovimo naslednje:

- a) če je problem realno simetričen, potem imamo za $\lambda \in \mathbb{R}$ lastno trojico (λ, x, x) ;
- b) če je problem kompleksno simetričen, potem imamo lastne trojice oblike $(\lambda, x, \bar{x})$.
- c) če je problem hermitski, potem trojice nastopajo v parih (λ, x, y) in $(\bar{\lambda}, y, x)$, v primeru $\lambda \in \mathbb{R}$ pa imamo trojico (λ, x, x) .

Zgled 1.1 Vzemimo $A = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$ in različne matrike B .

- a) Pri $B = \begin{bmatrix} 1 & 3 \\ 3 & -1 \end{bmatrix}$ je $T(\lambda) = \lambda B - A$ realno simetričen GEP. Kljub temu ima konjugirani lastni vrednosti $\lambda_1 = 0.7 + 0.1i$ in $\lambda_2 = 0.7 - 0.1i$.
- b) Pri $B = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$ dobimo kompleksno simetričen GEP, ki pa ni hermitski. Lastni vrednosti sta $\lambda_1 = 1 - 2i$ in $\lambda_2 = -1 - 2i$, torej ne velja niti to, da lastne vrednosti nastopajo v konjugiranih parih.
- c) Pri $B = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$ dobimo hermitski simetrični problem, ki ni kompleksno simetričen. Lastni vrednosti sta $\lambda_1 = \sqrt{5}i$ in $\lambda_2 = -\sqrt{5}i$. Pri hermitskih GEP so lastne vrednosti bodisi realne ali pa nastopajo v konjugiranih parih. $\square$

PEP $T(\lambda) = \lambda^m A_m + \dots + A_0$ je očitno hermitski natanko tedaj, ko so vse matrike $A_0, \dots, A_m$ hermitske. Posebni primer je hermitski GEP $T(\lambda) = \lambda B - A$, za katerega se izkaže, da ima same realne lastne vrednosti natanko tedaj, ko je matrični šop $\lambda B - A$ definiten, oziroma

$$\min\{(z^H A z)^2 + (z^H B z)^2 : z \in \mathbb{C}^n, \|z\| = 1\} > 0.$$

Za definitnost hermitskega GEP zadošča, da je ena izmed matrik A ali B pozitivno definitna.

Pri hermitskem QEP $T(\lambda) = \lambda^2 M + \lambda C + K$ se izkaže, da je zadosten pogoj za realni spekter to, da je problem *hiperboličen*, kar pomeni, da je matrika M pozitivno definitna in

$$(x^H C x)^2 - 4(x^H K x)(x^H M x) > 0$$

za vsak $x \neq 0$.

Omenimo še, da splošni NEP pred računanjem lastnih vrednosti ne moremo transformirati v lepšo obliko, ki bi bila za računanje ugodnejša, kot to lahko naredimo v primeru EP z redukcijo na Hessenbergovo obliko in v primeru GEP z redukcijo na Hessenbergovo-trikotno obliko.

Zaradi tega lahko pričakujemo, da za vsako reševanje sistema z matriko $T(\lambda)$ porabimo $\mathcal{O}(n^3)$ osnovnih računskih operacij.

Schurovo formo lahko pri EP in GEP izkoristimo za to, da v primeru, ko imamo en lastni par, naredimo deflacijo in preostale lastne vrednosti izračunamo iz matrik za eno manjše dimenzije. Tega pri NEP ne moremo narediti, saj ne obstaja nobena posplošitev Schurove forme.

1.3 Pomožni rezultati

Elementi a_{ij} matrike A velikosti $n \times n$ naj bodo zadosti krat zvezno odvedljive funkcije parametra λ , torej

$$A(\lambda) = \begin{bmatrix} a_{11}(\lambda) & \cdots & a_{1n}(\lambda) \\ \vdots & & \vdots \\ a_{n1}(\lambda) & \cdots & a_{nn}(\lambda) \end{bmatrix}.$$

Za odvod potem velja

$$A'(\lambda) = \begin{bmatrix} a'_{11}(\lambda) & \cdots & a'_{1n}(\lambda) \\ \vdots & & \vdots \\ a'_{n1}(\lambda) & \cdots & a'_{nn}(\lambda) \end{bmatrix}.$$

Če sta A in B matriki, katerih elementi so odvisni od parametra λ , potem za produkt velja

$$\frac{d}{d\lambda}(A(\lambda)B(\lambda)) = A'(\lambda)B(\lambda) + A(\lambda)B'(\lambda).$$

Od tod dobimo

$$\frac{d}{d\lambda}(A(\lambda)^2) = A'(\lambda)A(\lambda) + A(\lambda)A'(\lambda)$$

in

$$\frac{d}{d\lambda}(A(\lambda)^k) = \sum_{j=1}^k A(\lambda)^{j-1} A'(\lambda) A(\lambda)^{k-j},$$

saj $A(\lambda)$ in $A'(\lambda)$ v splošnem ne komutirata.

Če je $A(\lambda)$ obrnljiva, potem z odvajanjem zveze $A(\lambda)A^{-1}(\lambda) = I$ dobimo

$$\frac{d}{d\lambda}(A(\lambda)^{-1}) = -A(\lambda)^{-1}A'(\lambda)A(\lambda)^{-1} \quad (1.1)$$

in

$$\frac{d}{d\lambda}(A(\lambda)^{-k}) = -A(\lambda)^{-k} \frac{d}{d\lambda}(A(\lambda)^k) A(\lambda)^{-k}.$$

Lema 1.1 Naj bo A matrika velikosti $n \times n$, katere elementi so dovolj krat zvezno odvedljive funkcije parametra λ . Če velja $a_{ij}(\lambda_0) = 0$ za vse $i, j = 1, \dots, n$, potem za $f(\lambda) = \det(A(\lambda))$ velja

$$f(\lambda_0) = f'(\lambda_0) = \cdots = f^{(n-1)}(\lambda_0) = 0.$$

Dokaz. Dovolj se je spomniti, kako zgleda formula za determinanto matrike. V njej nastopajo vsote produktov po n elementov matrike A . Ker imajo v vsakem produktu vsi faktorji pri λ_0 vrednost 0, so očitno vsi odvodi determinante do vključno $(n - 1)$ -vega enaki 0. ■

V nadaljevanju bomo potrebovali tudi kot med podprostoroma, ki ju razpenjata vektorja u in v . Označimo ga s $\phi(u, v)$, definiran pa je s

$$\cos \phi = \frac{|u^H v|}{\|u\| \|v\|}.$$

Kot rešitev vzamemo kot med 0 in $\pi/2$.

Lema 1.2 Dana sta vektorja $x, y \in \mathbb{R}^n$, taka da je $\|x\| = \|y\| = 1$ in $\|x - y\| < 1$. Potem je

$$\sin \phi \leq \|x - y\| = \sqrt{\frac{2}{1 + \cos \phi}} \sin \phi \leq \sqrt{2} \sin \phi,$$

kjer je $\phi = \phi(x, y)$ kot med $\text{Lin}(x)$ in $\text{Lin}(y)$.

Lema 1.3 Dana sta vektorja $x, y \in \mathbb{C}^n$, taka da je $\|x\| = \|y\| = 1$ in $\phi = \phi(x, y) < \pi/2$. Potem obstaja tak $\alpha \in \mathbb{R}$, da za $\tilde{x} = e^{i\alpha} x$ velja

$$\sin \phi \leq \|\tilde{x} - y\| = \sqrt{\frac{2}{1 + \cos \phi}} \sin \phi \leq \sqrt{2} \sin \phi.$$

1.4 Algebrajska in geometrijska večkratnost lastne vrednosti

Pravimo, da je lastna vrednost λ_0 *algebrajsko enostavna*, če je $f'(\lambda_0) \neq 0$, kjer je $f(\lambda) = \det(T(\lambda))$. Podobno pravimo, da je λ_0 *m -kratna lastna vrednost*, če je $f(\lambda_0) = f'(\lambda_0) = \dots = f^{(m-1)}(\lambda_0)$ in $f^{(m)}(\lambda_0) \neq 0$.

Geometrijska večkratnost lastne vrednosti je enaka dimenziji prostora $\ker(T(\lambda_0))$, torej se ujema z maksimalnim številom neodvisnih desnih lastnih vektorjev. Pri navadnem problemu lastnih vrednosti vemo, da iz algebrajske enostavnosti sledi tudi geometrijska enostavnost. Pri NEP to ni tako očitno, zato bomo to dokazali.

Izrek 1.4 Naj bo λ_0 lastna vrednost nelinearnega problema lastnih vrednosti $T(\cdot) : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$. Potem je njena algebrajska večkratnost večja ali enaka geometrijski večkratnosti.

Dokaz. Naj bo $\dim(\ker(T(\lambda_0))) = r$. To pomeni, da obstajata taki permutacijski matriki P_1 in P_2 , da je

$$P_1 T(\lambda) P_2 = \begin{bmatrix} T_{11}(\lambda) & T_{12}(\lambda) \\ T_{21}(\lambda) & T_{22}(\lambda) \end{bmatrix},$$

kjer je $T_{11}(\lambda_0)$ nesingularna matrika velikosti $(n - r) \times (n - r)$.

Sedaj definiramo $\tilde{T}(\lambda) = P_1 T(\lambda) P_2$. Če je λ iz dovolj majhne okolice λ_0 , potem je matrika $T_{11}(\lambda)$ nesingularna in lahko razcepimo $\tilde{T}(\lambda) = L(\lambda) D(\lambda) U(\lambda)$, kjer so

$$\begin{aligned} L(\lambda) &= \begin{bmatrix} I_{n-r} & 0 \\ T_{21}(\lambda) T_{11}^{-1}(\lambda) & I_r \end{bmatrix}, \\ D(\lambda) &= \begin{bmatrix} T_{11}(\lambda) & 0 \\ 0 & T_{22}(\lambda) - T_{21}(\lambda) T_{11}^{-1}(\lambda) T_{12}(\lambda) \end{bmatrix}, \\ U(\lambda) &= \begin{bmatrix} I_{n-r} & T_{11}^{-1}(\lambda) T_{12}(\lambda) \\ 0 & I_r \end{bmatrix}. \end{aligned}$$

Iz $T(\lambda) = P_1 L(\lambda) D(\lambda) U(\lambda) P_2$ sledi, da je bločno diagonalna matrika $D(\lambda)$ ekvivalentna $T(\lambda)$. Torej je $\text{rang}(D(\lambda_0)) = \text{rang}(T(\lambda_0)) = n - r$. Ker je očitno $\text{rang}(T_{11}(\lambda_0)) = n - r$, mora potem za matriko v spodnjem desnem kotu veljati $T_{22}(\lambda_0) - T_{21}(\lambda_0) T_{11}^{-1}(\lambda_0) T_{12}(\lambda_0) = 0$.

Determinanta $f(\lambda) = \det(T(\lambda))$ se do predznaka ujema z $\det(D(\lambda)) = \det(T_{11}(\lambda)) s(\lambda)$, kjer je $s(\lambda) = \det(T_{22}(\lambda) - T_{21}(\lambda) T_{11}^{-1}(\lambda) T_{12}(\lambda))$. Po lemi 1.1 velja $s(\lambda_0) = s'(\lambda_0) = \dots = s^{(r-1)}(\lambda_0) = 0$, od tod pa sledi $f(\lambda_0) = f'(\lambda_0) = \dots = f^{(r-1)}(\lambda_0) = 0$, kar pomeni, da je algebrajska večkratnost λ_0 vsaj r . ■

Posledica zgornjega izreka je, da je vsaka algebrajsko enostavna lastna vrednost tudi geometrijsko enostavna.

Posledica 1.5 *Naj bo λ_0 algebrajsko enostavna lastna vrednost nelinearnega problema lastnih vrednosti $T(\cdot) : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$, torej $f(\lambda_0) = 0$ in $f'(\lambda_0) \neq 0$, kjer je $f(\lambda) = \det(T(\lambda))$. Potem je λ_0 tudi geometrijsko enostavna lastna vrednost, torej $\dim(\ker(T(\lambda_0))) = 1$.*

Od sedaj naprej bomo pri lastnih vrednostih, ki so algebrajsko enostavne, izpuščali besedo algebrajsko. Enostavna lastna vrednost pomeni, da je algebrajsko in geometrijsko enostavna.

Pomembno je tudi vedeti, v katerem primeru iz geometrijsko enostavne lastne vrednosti sledi, da je lastna vrednost tudi algebrajsko enostavna. O tem govori naslednji izrek.

Izrek 1.6 *Naj bo λ_0 geometrijsko enostavna lastna vrednost nelinearnega problema lastnih vrednosti $T(\cdot) : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$, torej $\dim(\ker(T(\lambda_0))) = 1$. Naj bo x_0 normiran desni lastni vektor, y_0 pa normiran levi lastni vektor. Potem je λ_0 algebrajsko enostavna lastna vrednost natanko tedaj, ko velja*

$$y_0^H T'(\lambda_0) x_0 \neq 0.$$

Dokaz. Definiramo $(n+1) \times (n+1)$ matriko

$$C(\lambda) = \begin{bmatrix} T(\lambda) & y_0 \\ x_0^H & 0 \end{bmatrix},$$

ki je odvisna od parametra λ . Pokažimo, da je $C(\lambda_0)$ nesingularna matrika. Iz enačbe

$$C(\lambda_0) \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} T(\lambda_0) & y_0 \\ x_0^H & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = 0$$

dobimo enačbi

$$\begin{aligned} T(\lambda_0)x + \mu y_0 &= 0 \\ x_0^H x &= 0. \end{aligned}$$

Ko prvo enačbo pomnožimo z y_0^H , dobimo $\mu = 0$, od tod pa sledi $T(\lambda_0)x = 0$. Zaradi $\dim(\ker(T(\lambda_0))) = 1$ mora biti x oblike $x = \alpha x_0$, iz druge enačbe pa potem sledi $\alpha = 0$. Torej $x = 0$, $\mu = 0$ in $C(\lambda_0)$ je nesingularna.

Za λ iz dovolj majhne okolice λ_0 je potem matrika $C(\lambda)$ nesingularna in lahko definiramo

$$\gamma(\lambda) = e_{n+1}^T C(\lambda)^{-1} e_{n+1}. \quad (1.2)$$

Če označimo $f(\lambda) = \det(T(\lambda))$ in $g(\lambda) = \det(C(\lambda))$, potem vemo, da velja

$$f(\lambda) = \gamma(\lambda)g(\lambda).$$

Ko v zadnjo enačbo vstavimo $\lambda = \lambda_0$, dobimo $\gamma(\lambda_0) = 0$.

Z odvajanjem dobimo

$$f'(\lambda_0) = \gamma'(\lambda_0)g(\lambda_0) + \gamma(\lambda_0)g'(\lambda_0) = \gamma'(\lambda_0)g(\lambda_0),$$

saj je $\gamma(\lambda_0) = 0$. Z odvajanjem (1.2) in upoštevanjem formule (1.1) dobimo

$$\begin{aligned} \gamma'(\lambda_0) &= -e_{n+1}^T C(\lambda_0)^{-1} C'(\lambda_0) C(\lambda_0)^{-1} e_{n+1} \\ &= - \begin{bmatrix} y_0 \\ 0 \end{bmatrix}^H \begin{bmatrix} T'(\lambda_0) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ 0 \end{bmatrix} \\ &= -y_0^H T'(\lambda_0) x_0 \neq 0. \end{aligned}$$

Od tod sledi $f'(\lambda_0) \neq 0$ in lastna vrednost λ_0 je algebrsko enostavna. ■

Opomba 1.1 Pogoj iz zadnjega izreka se v primeru EP spremeni v $y_0^H x_0 \neq 0$. To pomeni, da v primeru geometrijsko enostavne lastne vrednosti za levi in desni lastni vektor velja:

- a) če je λ_0 algebrsko večkratna lastna vrednost, sta ortogonalna;
- b) če je λ_0 algebrsko enostavna lastna vrednost, nista ortogonalna.

Po drugi strani pa vemo, da sta levi in desni lastni vektor, ki pripadata različnim lastnim vrednostim, vedno ortogonalna.

1.5 Občutljivost enostavne lastne vrednosti

Pri občutljivosti lastne vrednosti nas zanima, za koliko se lahko spremeni lastna vrednost, če zmotimo matriko. V primeru EP poznamo naslednji izrek.

Izrek 1.7 Naj bo λ_0 enostavna lastna vrednost matrike A . Pripadajoči desni normiran lastni vektor naj bo x_0 , levi pa y_0 . Za vsako dovolj majhno motnjo ΔA ima potem matrika $A + \Delta A$ enolično lastno vrednost $\lambda_0 + \Delta\lambda_0$, za katero velja

$$\Delta\lambda_0 = \frac{y_0^H \Delta A x_0}{y_0^H x_0} + \mathcal{O}(\|\Delta A\|^2).$$

Po normi relativno pogojenostno število za enostavno lastno vrednost je definirano kot

$$\kappa(\lambda_0, A) := \limsup_{\varepsilon \rightarrow 0} \left\{ \frac{|\Delta\lambda_0|}{\varepsilon} : ((A + \Delta A) - (\lambda_0 + \Delta\lambda_0)I)(x_0 + \Delta x_0) = 0, \|\Delta A\| \leq \varepsilon \|E\| \right\}.$$

Tu ponavadi vzamemo $\|E\| = 1$ za absolutne motnje ali pa $\|E\| = \|A\|$ za relativne motnje. Izkaže se, da je

$$\kappa(\lambda_0, A) = \frac{\|E\|}{|y_0^H x_0|},$$

kjer je enakost dosežena pri $\Delta A = \varepsilon \|E\| y_0 x_0^H$.

Opomba 1.2 Navadno smo občutljivost enostavne lastne vrednosti λ_0 označevali kot $1/|y_0^H x_0|$. To dobimo, če vzamemo $\|E\| = 1$.

V primeru simetričnega EP se levi in desni lastni vektorji ujemajo, zato je $y_0^H x_0 = 1$ in vse lastne vrednosti imajo enako absolutno občutljivost 1.

Podobno lahko občutljivost definiramo tudi za GEP in PEP. Pri PEP matrični polinom

$$T(\lambda) = \lambda^m A_m + \lambda^{m-1} A_{m-1} + \cdots + \lambda A_1 + A_0$$

zmotimo v

$$\tilde{T}(\lambda) = \lambda^m (A_m + \Delta A_m) + \lambda^{m-1} (A_{m-1} + \Delta A_{m-1}) + \cdots + \lambda (A_1 + \Delta A_1) + (A_0 + \Delta A_0),$$

kjer velja $\Delta A_j \leq \varepsilon \|E_j\|$ za $j = 0, \dots, m$. Za občutljivost enostavne lastne vrednosti λ_0 potem velja (poglej [18])

$$\begin{aligned} \kappa(\lambda_0) &= \limsup_{\varepsilon \rightarrow 0} \left\{ \frac{|\Delta\lambda_0|}{\varepsilon} : \tilde{T}(\lambda_0 + \Delta\lambda_0)(x_0 + \Delta x_0) = 0, \|\Delta A_i\| \leq \varepsilon \|E_i\| \right\} \\ &= \frac{\sum_{i=0}^m |\lambda_0|^i \|E_i\|}{|y_0^H T'(\lambda_0) x_0|}. \end{aligned} \tag{1.3}$$

Pri NEP se moramo zadeve lotiti drugače, saj ni čisto jasno, kako predstaviti motnjo. Predpostavimo, da lahko zmoten nelinearni problem zapišemo v obliki $\tilde{T}(\lambda, \varepsilon)$, kjer je $\varepsilon \in \mathbb{C}^d$, pri čemer za nezmoten problem velja $T(\lambda) = \tilde{T}(\lambda, 0)$ za vsak $\lambda \in \mathbb{C}$.

Izrek 1.8 Naj bo $\tilde{T} : D \times E \rightarrow \mathbb{C}^{n \times n}$, kjer sta $D \subset \mathbb{C}$ in $E \subset \mathbb{C}^d$, $0 \in E$, odprti množici, zvezno odvedljiva in naj bo $\lambda_0 \in D$ enostavna lastna vrednost $T(\lambda) := \tilde{T}(\lambda, 0)$ s pripadajočim desnim normiranim lastnim vektorjem x_0 in levim y_0 . V dovolj majhni okolici 0 ima $\tilde{T}(\cdot, \varepsilon)$ enostavno lastno vrednost

$$\lambda(\varepsilon) = \lambda_0 + \sum_{k=1}^d \frac{y_0^H \frac{\partial \tilde{T}}{\partial \varepsilon_k}(\lambda_0, 0) x_0}{y_0^H T'(\lambda_0) x_0} \varepsilon_k + \mathcal{O}(\|\varepsilon\|^2). \tag{1.4}$$

Dokaz. Vpeljemo matriko

$$C(\lambda, \varepsilon) = \begin{bmatrix} \tilde{T}(\lambda_0, \varepsilon) & y_0 \\ x_0^H & 0 \end{bmatrix},$$

za katero lahko na podoben način kot v dokazu izreka 1.6 pokažemo, da je nesingularna v dovolj majhni okolici $(\lambda_0, 0)$.

Sedaj definiramo

$$C(\lambda, \varepsilon)^{-1} e_{n+1} = \begin{bmatrix} x(\lambda, \varepsilon) \\ \gamma(\lambda, \varepsilon) \end{bmatrix} \quad \text{in} \quad C(\lambda, \varepsilon)^{-H} e_{n+1} = \begin{bmatrix} y(\lambda, \varepsilon) \\ \nu(\lambda, \varepsilon) \end{bmatrix}.$$

Velja

$$\gamma(\lambda, \varepsilon) = e_{n+1}^T C(\lambda, \mu)^{-1} e_{n+1}. \quad (1.5)$$

V točki $(\lambda_0, 0)$ dobimo $x(\lambda_0, 0) = x_0$, $y(\lambda_0, 0) = y_0$ in $\gamma(\lambda_0, 0) = 0$.

Z odvajanjem zveze (1.5) po spremenljivki λ dobimo

$$\begin{aligned} \frac{\partial \gamma}{\partial \lambda}(\lambda, \varepsilon) &= -e_{n+1}^T C(\lambda, \varepsilon)^{-1} \begin{bmatrix} \frac{\partial \tilde{T}}{\partial \lambda}(\lambda, \varepsilon) & 0 \\ 0 & 0 \end{bmatrix} C(\lambda, \varepsilon)^{-1} e_{n+1} \\ &= -y(\lambda, \varepsilon)^H \frac{\partial \tilde{T}}{\partial \lambda}(\lambda, \varepsilon) x(\lambda, \varepsilon). \end{aligned}$$

V točki $(\lambda_0, 0)$ dobimo $\frac{\partial \gamma}{\partial \lambda}(\lambda_0, 0) = -y_0^H T'(\lambda_0) x_0 \neq 0$, to pa po izreku o implicitni funkciji pomeni, da je za ε z dovolj majhno normo $\|\varepsilon\|$ enačba $\gamma(\lambda, \varepsilon) = 0$ enolično rešljiva z $\lambda = \lambda(\varepsilon)$, da je $\gamma(\lambda(\varepsilon), \varepsilon) \equiv 0$ in $\lambda(0) = \lambda_0$.

Od tod sledi

$$\frac{\partial \gamma}{\partial \lambda} \frac{\partial \lambda}{\partial \varepsilon_k} + \frac{\partial \gamma}{\partial \varepsilon_k} \equiv 0,$$

kar nam da

$$\frac{\partial \lambda}{\partial \varepsilon_k}(0) = -\frac{1}{y_0^H T'(\lambda_0) x_0} \frac{\partial \gamma}{\partial \varepsilon_k}(\lambda_0, 0)$$

za $k = 1, \dots, d$.

Če (1.5) odvajamo po ε_k , dobimo

$$\frac{\partial \gamma}{\partial \varepsilon_k}(\lambda, \varepsilon) = -y(\lambda, \varepsilon)^H \frac{\partial \tilde{T}}{\partial \varepsilon_k}(\lambda, \varepsilon) x(\lambda, \varepsilon),$$

torej

$$\frac{\partial \lambda}{\partial \varepsilon}(0) = -\frac{1}{y_0^H T'(\lambda_0) x_0} y_0^H \begin{bmatrix} \frac{\partial \tilde{T}}{\partial \varepsilon_1}(\lambda_0, 0) x_0 & \dots & \frac{\partial \tilde{T}}{\partial \varepsilon_d}(\lambda_0, 0) x_0 \end{bmatrix}.$$

Po razvoju v vrsto dobimo

$$\lambda(\varepsilon) = \lambda_0 + \frac{\partial \lambda}{\partial \varepsilon}(0) \varepsilon + \mathcal{O}(\|\varepsilon\|^2),$$

kar pa je ravno (1.4). ■

Občutljivost enostavne lastne vrednosti λ_0 je potem enaka

$$\kappa(\lambda_0) = \limsup_{\|\varepsilon\| \rightarrow 0} \frac{|\lambda(\varepsilon) - \lambda_0|}{\|\varepsilon\|_2} = \frac{1}{|y_0^H T'(\lambda_0) x_0|} \left(\sum_{k=1}^d \left| y_0^H \frac{\partial \tilde{T}}{\partial \varepsilon_k}(\lambda_0, 0) x_0 \right|^2 \right)^{1/2}.$$

Zgled 1.2 Za PEP $T(\lambda) = \lambda^m A_m + \dots + \lambda A_1 + A_0$ lahko motnjo zapišemo v obliki

$$\tilde{T}(\lambda, \epsilon) = \lambda^m (A_m + \epsilon_m E_m) + \dots + \lambda (A_1 + \epsilon_1 E_1) + A_0 + \epsilon_0 E_0.$$

Od tod za $k = 0, \dots, m$ sledi

$$\frac{\partial \tilde{T}}{\partial \epsilon_k} = \lambda_k E_k.$$

Če je λ_0 lastna vrednost T , potem za zmoteno lastno vrednost $\tilde{T}$ po (1.4) velja

$$\lambda(\epsilon) = \lambda_0 - \sum_{k=0}^m \frac{y_0^H E_k x_0 \lambda_0^k}{y_0^H T'(\lambda_0) x_0} \epsilon_k + \mathcal{O}(\|\epsilon\|^2).$$

Če izberemo $\|\cdot\|_\infty$, lahko iz zgornje zveze ocenimo

$$\frac{|\lambda(\epsilon) - \lambda_0|}{\|\epsilon\|_\infty} \leq \sum_{k=0}^m \frac{|\lambda_0|^k |y_0 E_k x_0|}{|y_0^H T'(\lambda_0) x_0|} + \mathcal{O}(\|\epsilon\|_\infty^2) \leq \sum_{k=0}^m \frac{|\lambda_0|^k \|E_k\|}{|y_0^H T'(\lambda_0) x_0|} + \mathcal{O}(\|\epsilon\|_\infty^2),$$

kar se ujema z oceno (1.3). □

1.6 Rayleighov funkcional

Naj bo $A \in \mathbb{C}^{n \times n}$. Za $x \in \mathbb{C}^n$, $x \neq 0$, definiramo Rayleighov kvocient kot

$$\rho(A, x) := \frac{x^H A x}{x^H x}.$$

Rayleighov kvocient ima naslednje lepe lastnosti:

- a) $\rho(\alpha A, \beta x) = \alpha \rho(A, x)$ za poljubna $\alpha, \beta \neq 0$, torej je odvisen le od smeri in ne od velikosti vektorja x .
- b) $\rho(A - \sigma I, x) = \rho(A, x) - \sigma$, kar pomeni, da je invarianten za translacije.
- c) Omejenost: za množico

$$W(A) := \{\rho(A, x) : x \in \mathbb{C}^n, x \neq 0\},$$

ki ji pravimo *zaloga vrednosti* oziroma *numerični zaklad* matrike A , velja, da je zaprta, omejena in konveksna.

Če je matrika A hermitska, je Rayleighov kvocient realen in velja $\lambda_n \leq \rho(A, x) \leq \lambda_1$, kjer sta λ_n in λ_1 najmanjša oziroma največja lastna vrednost matrike A .

- d) Če je x_i lastni vektor, ki pripada lastni vrednosti λ_i , je $\rho(A, x_i) = \lambda_i$.
- e) Najboljša aproksimacija: Pri $\sigma = \rho(A, x)$ je dosežen

$$\min_{\sigma \in \mathbb{C}} \|(A - \sigma I)x\|_2,$$

kar pomeni, da je po tem kriteriju $\rho(A, x)$ najboljši približek za lastno vrednost, ki jo lahko dobimo iz približka za lastni vektor x .

- f) Če je A normalna matrika, potem so lastni vektorji stacionarne točke ρ , vrednosti v stacionarnih točkah pa so lastne vrednosti.

Stacionarnost za odvedljivo funkcijo pomeni, da je v stacionarni točki gradient enak 0. V našem primeru bi morali biti vsi parcialni odvodi $\rho(A, x)$ po posameznih komponentah $x_1, \dots, x_n$ vektorja x enaki 0.

V splošnem kompleksnem primeru $\rho(A, x)$ ni odvedljiva, če pa je A simetrična matrika, se lahko omejimo le na realne vektorje x . V tem primeru dobimo

$$\begin{aligned} \frac{\partial \rho(A, x)}{\partial x_k} &= \frac{\frac{\partial(x^T Ax)}{\partial x_k} x^T x - (x^T Ax) \frac{\partial(x^T x)}{\partial x_k}}{(x^T x)^2} \\ &= \frac{2(Ax)_k}{x^T x} - \frac{(x^T Ax) 2x_k}{(x^T x)^2} = \frac{2}{x^T x} (Ax - \rho(A, x)x)_k, \end{aligned}$$

od koder sledi

$$\nabla \rho(A, x) = \frac{2}{x^T x} (Ax - \rho(A, x)x).$$

Če je x lastni vektor, je $Ax = \rho(A, x)x$ in x je očitno stacionarna točka.

V kompleksnem primeru ne moremo odvajati, zato stacionarne točke definiramo drugače.

Definicija 1.9 Naj bo f preslikava iz $\mathbb{C}^n$ v $\mathbb{C}$. Pravimo, da je z stacionarna točka za f , če za Δz iz dovolj majhne okolice 0 velja $f(z + \Delta z) = f(z) + \mathcal{O}(\|\Delta z\|^2)$.

Če je f dovoljkrat odvedljiva, se torej pri razvoju $f(z + \Delta z)$ okrog stacionarne točke z ne pojavijo členi 1. reda, kar je ekvivalentno $\nabla f(z) = 0$.

Za $\rho(A, x) = \frac{x^H Ax}{x^H x}$ lahko izračunamo smerni odvod v smeri v kot

$$\frac{d\rho(A, x)}{dv} = \lim_{t \rightarrow 0} \frac{\rho(A, x + tv) - \rho(A, x)}{t}.$$

Stacionarnost je ekvivalentna temu, da so v točki x vsi smerni odvodi enaki 0.

Izrek 1.10 Če je A normalna matrika, potem so stacionarne točke Rayleighovega kvocienta natanko lastni vektorji matrike A .

Dokaz. Izračunamo smerni odvod. Iz

$$\begin{aligned} \rho(A, x + tv) - \rho(A, x) &= \frac{(x + tv)^H A (x + tv)}{(x + tv)^H (x + tv)} - \frac{x^H Ax}{x^H x} \\ &= t \frac{(v^H Ax + x^H Av + tv^H Av)x^H x - (x^H Ax)(v^H x + x^H v + tv^H v)}{(x + tv)^H (x + tv)x^H x} \end{aligned}$$

sledi

$$\frac{d\rho(A, x)}{dv} = \frac{(v^H Ax + x^H Av)x^H x - (x^H Ax)(v^H x + x^H v)}{(x^H x)^2}$$

$$\begin{aligned}
&= \frac{v^H Ax - \rho(A, x)v^H x + x^H Av - \rho(A, x)x^H v}{x^H x} \\
&= \frac{v^H(A - \rho(A, x)I)x}{x^H x} + \frac{x^H(A - \rho(A, x)I)v}{x^H x}.
\end{aligned}$$

Podobno lahko izpeljemo

$$\frac{d\rho(A, x)}{d(iv)} = -i \frac{v^H(A - \rho(A, x)I)x}{x^H x} + i \frac{x^H(A - \rho(A, x)I)v}{x^H x}.$$

Za stacionarnost mora tako za vse vektorje v veljati

$$\begin{aligned}
v^H(A - \rho(A, x)I)x + x^H(A - \rho(A, x)I)v &= 0 \\
-v^H(A - \rho(A, x)I)x + x^H(A - \rho(A, x)I)v &= 0,
\end{aligned}$$

kar nam da

$$\begin{aligned}
(A - \rho(A, x)I)x &= 0 \\
(A^H - \overline{\rho(A, x)}I)x &= 0.
\end{aligned}$$

To pomeni, da je x lastni vektor za matriki A in A^H , to pa je lastnost, ki velja za normalne matrike. Vsako normalno matriko lahko namreč zapišemo v obliki $A = UDU^H$, kjer je U unitarna, D pa diagonalna matrika. ■

Zgled 1.3 Če matrika A ni normalna, to ne drži več ...

Za splošne matrike lahko za $x, y \neq 0$ definiramo dvostranski Rayleighov kvocient

$$\rho(A, x, y) := \frac{y^H Ax}{y^H x}.$$

Za splošno matriko A lahko pokažemo, da je stacionarna točka (x, y) , kjer je x desni, y pa levi lastni vektor za isto lastno vrednost.

Naj bo $Ax = \lambda x$ in $y^H A = \lambda y^H$. Od tod dobimo

$$\begin{aligned}
\rho(A, x + u, y + v) &- \rho(A, x, y) \\
&= \frac{(v^H Ax + y^H Au + v^H Au)y^H x - (v^H x + y^H u + v^H u)y^H Ax}{(y^H x + v^H x + y^H u + v^H u)y^H x} \\
&= \frac{(v^H Au)y^H x - \lambda(v^H u)y^H x}{(y^H x + v^H x + y^H u + v^H u)y^H x} \\
&= \frac{v^H Au - \lambda v^H u}{y^H x + v^H x + y^H u + v^H u} = \mathcal{O}(\|u\|\|v\|) = \mathcal{O}((\|u\| + \|v\|)^2)
\end{aligned}$$

in tako smo pokazali, da je (x, y) stacionarna točka.

Za NEP je Duffin posplošil Rayleighov kvocient z vpeljavo naslednjega funkcionala.

Definicija 1.11 Naj bo T realen simetričen NEP. Rayleighov funkcional je definiran kot zvezni funkcional p na $\mathbb{R}^n \setminus \{0\}$, ki ima naslednje lastnosti:

- a) $p(u) \in (a, b)$ za vsak $u \in \mathbb{R}^n$, $u \neq 0$,
- b) $p(\gamma u) = p(u)$ za vsak $\gamma \neq 0$,
- c) $\langle T(p(u))u, u \rangle = 0$,
- d) $\langle T'(p(u))u, u \rangle > 0$.

Množico $W(T) = \{p(u) : u \in \mathbb{R}^n, u \neq 0\} \subset \mathbb{R}$ imenujemo zaloga vrednosti T .

Zgled 1.4 Če vzamemo EP, kjer je $T(\lambda) = \lambda I - A$, potem iz točke c) sledi, da je

$$p(u) = \frac{u^H A u}{u^H u}.$$

Pogoj d) je vedno izpolnjen, saj je $u^H u > 0$ za $u \neq 0$. □

Zgled 1.5 Če vzamemo GEP, kjer je $T(\lambda) = \lambda B - A$, potem iz točke c) dobimo

$$p(u) = \frac{u^H A u}{u^H B u}$$

za $u^H B u \neq 0$. Pogoj d) je izpolnjen, če je B hermitska pozitivno definitna matrika. □

Zgled 1.6 Za QEP $T(\lambda) = \lambda^2 M + \lambda C + K$ iz točke c) sledi, da je Rayleighov funkcional rešitev kvadratne enačbe

$$p(u)^2(u^H M u) + p(u)(u^H C u) + u^H K u = 0. \quad (1.6)$$

Pogoju d) pravimo nadkritična dušenost. Za QEP pravimo, da je nadkritično dušen (angl. overdamped), če je hiperboličen (kar pomeni, da so matrike M, C, K hermitske, M je pozitivno definitna in $(x^H C x)^2 > 4(x^H M x)(x^H K x)$ za vsak $x \neq 0$) in je dodatno matrika C pozitivno definitna, K pa nenegativno definitna. Če za takšen QEP izberemo t.i. primarni funkcional, ki je definiran kot večja izmed rešitev kvadratne enačbe (1.6) oziroma

$$p_+(u) = \frac{-u^H C u + \sqrt{(u^H C u)^2 - 4(u^H M u)(u^H K u)}}{2u^H M u},$$

potem velja, da je

$$u^H T'(p_+(u))u = 2p_+(u)(u^H M u) + u^H C u > 0$$

in pogoj d) je izpolnjen. □

Zgled 1.7 Vzemimo NEP oblike $T(\lambda) = \text{diag}(2 - e^\lambda, 1, \dots, 1)$. Edina lastna vrednost tega problema je $\lambda_0 = \ln 2$, lastni vektor pa je $x_0 = e_1$. Iz točke c) sledi, da Rayleighov funkcional zadošča enačbi

$$u^H T(p(u))u = (2 - e^{p(u)})|u_1|^2 + |u_2|^2 + \dots + |u_n|^2 = 0,$$

od koder sledi

$$p(u) = \ln \left(2 + \frac{|u_2|^2 + \dots + |u_n|^2}{|u_1|^2} \right).$$

Če vzamemo $x = [1 \ -1 \ \dots \ -1]^T$, potem dobimo $p(x) = \ln(n+1)$ in

$$\|T(p(x))x\|_2 = \sqrt{n(n-1)} > \|T(\ln 2)x\|_2 = \sqrt{n-1}.$$

Torej Rayleighov funkcional, za razliko od Rayleighovega kvocienta, ne minimizira nujno ostanka. $\square$

Težava z Rayleighovim funkcionalom je, da enačba, ki jo dobimo iz točke c), nima nujno enolične rešitve. Že pri QEP smo videli, da dobimo dve rešitvi. Če želimo kaj povedati o obstoju in enoličnosti Rayleighovega funkcionala, moramo definicijo malo popraviti. Namesto na celotnem območju sedaj definiramo funkcional le lokalno v okolici lastnega vektorja x_0 . Naj bo

$$K_\epsilon(x_0) = \{u \in \mathbb{C}^n : \phi(u, x_0) \leq \epsilon\},$$

kjer je $\phi(u, x_0)$ kot med podprostoroma, ki ju razpenjata u in x_0 .

Naj bo x_0 desni vektor za enostavno lastno vrednost λ_0 . Sedaj definiramo Rayleighov funkcional kot funkcijo iz $K_\epsilon(x_0)$ v $\bar{B}(\lambda_0, \tau_0) = \{z \in \mathbb{C} : |z - \lambda_0| \leq \tau_0\}$ za $\epsilon < \pi/2$ in $\tau_0 > 0$. Pri točki d) zahtevamo namesto pozitivnosti le $\langle T'(p(u))u, u \rangle \neq 0$.

Enostranski Rayleighov funkcional je primeren za hermitske probleme z realnimi lastnimi vrednostmi. V tem primeru je pogoj $\langle T'(p(u))u, u \rangle \neq 0$ izpolnjen za vse vektorje u , ki so dovolj blizu lastnega vektorja x_0 enostavne lastne vrednosti λ_0 .

Naj bo x_0 desni, y_0 pa levi lastni vektor za isto lastno vrednost λ_0 . Za splošni NEP definiramo dvostranski Rayleighov funkcional v okolici (x_0, y_0) z naslednjimi lastnostmi

$$a) \ p : K_\epsilon(x_0) \times K_\epsilon(y_0) \rightarrow \bar{B}(\lambda_0, \tau_0).$$

$$b) \ p(cu, dv) = p(u, v) \text{ za } c, d \neq 0.$$

$$c) \ \langle T(p(u, v))u, v \rangle = 0,$$

$$d) \ \langle T'(p(u, v))u, v \rangle \neq 0,$$

kjer je $\epsilon < \pi/2$ in $\tau_0 > 0$. Pogoj $\epsilon < \pi/2$ nam zagotavlja, da u in v ne moreta biti enaka 0.

Kaj lahko povemo o obstoju Rayleighovega funkcionala in stacionarnosti? Za začetek pogledajmo simetričen realni primer, kjer privzamemo, da imamo realne lastne vrednosti in lastne vektorje. Potem lahko definiramo

$$g(p, u, v) = v^T T(p)u.$$

Z odvajanjem dobimo

$$\frac{\partial g}{\partial p}(p, u, v) = v^T T'(p)u,$$

$$\frac{\partial g}{\partial u}(p, u, v) = v^T T(p),$$

$$\frac{\partial g}{\partial v}(p, u, v) = u^T T(p)^T.$$

Če je λ_0 enostavna lastna vrednost in sta x_0 in y_0 pripadajoča desni in levi lastni vektor, potem je po izreku 1.6

$$\frac{\partial g}{\partial p}(\lambda_0, x_0, y_0) = y_0^T T'(\lambda_0) x_0 \neq 0,$$

od tod pa po izreku o implicitni funkciji sledi, da v okolici (x_0, y_0) obstaja rešitev $p(u, v)$, za katero velja $g(p(u, v), u, v) \equiv 0$ in $p(x_0, y_0) = \lambda_0$. To je ravno iskani Rayleighov funkcional. Z odvajanjem $g(p(u, v), u, v) \equiv 0$ dobimo

$$\frac{\partial g}{\partial p} \frac{\partial p}{\partial u} + \frac{\partial g}{\partial u} \equiv 0,$$

$$\frac{\partial g}{\partial p} \frac{\partial p}{\partial v} + \frac{\partial g}{\partial v} \equiv 0$$

in izpeljemo

$$\frac{\partial p}{\partial u} = -\frac{v^T T(p)}{v^T T'(p) u}$$

$$\frac{\partial p}{\partial v} = -\frac{u^T T(p)^T}{v^T T'(p) u}.$$

Iz razvoja v Taylorjevo vrsto potem sledi

$$p(u + s, v + t) = p(u, v) - \frac{v^T T(p) s}{v^T T'(p) u} - \frac{u^T T(p)^T t}{v^T T'(p) u} + R(s, t), \quad (1.7)$$

kjer je $R(s, t) = \mathcal{O}((\|s\| + \|t\|)^2)$. Tako smo za realni primer pokazali naslednjo lemo, vse kar moramo še narediti za dokaz je, da v zgornji razvoj vstavimo $u = x_0$ in $v = y_0$.

Lema 1.12 *Naj bo T simetričen realen nelinearni problem lastnih vrednosti in $\lambda_0 \in \mathbb{R}$ enostavna lastna vrednost z realnim desnim in levim lastnim vektorjem x_0 in y_0 . Potem je (x_0, y_0) stacionarna točka Rayleighovega funkcionala.*

Kompleksni primer je dosti bolj zapleten, saj ne moremo odvajati zaradi prisotnosti izraza $v^H u$. Z uporabo drugačnih orodij se da pokazati (glej [17]) naslednji izrek.

Izrek 1.13 *Naj bo λ_0 enostavna lastna vrednost nelinearnega problema lastnih vrednosti T , naj bosta x_0 in y_0 pripadajoča desni in levi lastni vektor, in naj bo $T(\lambda)$ holomorfna na odprti okolici množice $\bar{B}(\lambda_0, \tau_0)$ za $\tau_0 > 0$. Potem obstajata konstanti $0 < \tau < \tau_0$ in $0 < \epsilon_0 < \pi/2$, da za vse $(u, v) \in K_{\epsilon_0}(x_0) \times K_{\epsilon_0}(y_0)$ obstaja enolični $p(u, v) \in \bar{B}(\lambda_0, \tau)$, da je $g(p(u, v), u, v) = v^H T(p(u, v)) u \equiv 0$, pri čemer je $p(x_0, y_0) = \lambda_0$, in velja ocena*

$$|p(u, v) - \lambda_0| \leq \frac{8}{3} \frac{\|T(\lambda_0)\|}{|y_0^H T'(\lambda_0) x_0|} \tan \eta \tan \xi,$$

kjer je $\eta = \phi(u, x_0)$ in $\xi = \phi(v, y_0)$.

Pri predpostavkah zgornjega izreka naj bo $(u, v) \in K_{\epsilon_0/2}(x_0) \times K_{\epsilon_0/2}(y_0)$. Potem za vsaka $\|s\| \leq \delta_u \|u\|$ in $\|t\| \leq \delta_v \|v\|$, kjer je

$$0 \leq \delta_u, \delta_v \leq \delta_0 = \frac{\cos(\epsilon_0/2) - \cos(\epsilon_0)}{1 + \cos(\epsilon_0)},$$

velja $(u + s, v + t) \in K_{\epsilon_0}(x_0) \times K_{\epsilon_0}(y_0)$ in obstaja enolični $p(u + s, t + v) \in \bar{B}(\lambda_0, \tau_0)$, za katerega velja

$$p(u + s, t + v) - p(u, v) = -\frac{v^H T(p)s}{v^H T'(p)u} - \frac{t^H T(p)u}{v^H T'(p)u} + R(s, t),$$

kjer je

$$R(s, t) = \mathcal{O}(\delta_u^2 + \delta_v^2).$$

Ker je, ko gre ϵ_0 proti 0, $\delta_0 \approx 3\epsilon_0^2/16$, od tod sledi stacionarnost točke (x_0, y_0) .

V primeru dvostranskega Rayleighovega funkcionala je torej red aproksimacije lastne vrednosti kvadratičen glede na napako aproksimacije za levi in desni lastni vektor.

V primeru enostranskega Rayleighovega funkcionala se da pokazati, da če za enostavni lastni par (λ_0, x_0) velja $x_0^H T'(\lambda_0)x_0 \neq 0$ in je $T(\lambda)$ holomorfna na odprti okolici množice $\bar{B}(\lambda_0, \tau_0)$ za $\tau_0 > 0$, potem obstajata konstanti $0 < \tau < \tau_0$ in $0 < \epsilon_0 < \pi/2$, da za vsak $u \in K_{\epsilon_0}(x_0)$ obstaja enolični $p(u) \in \bar{B}(\lambda_0, \tau)$, da je $g(p(u), u) = u^H T(p(u))u \equiv 0$, pri čemer je $p(x_0) = \lambda_0$, in velja ocena

$$|p(u) - \lambda_0| \leq \frac{10}{3} \frac{\|T(\lambda_0)\|}{|x_0^H T'(\lambda_0)x_0|} \tan \eta,$$

kjer je $\eta = \phi(u, x_0)$. To pomeni, da je red natančnosti enostranskega Rayleighovega funkcionala enak redu natančnosti aproksimacije lastnega vektorja, torej linearen.

Omenimo še posplošitev Rayleighovega kvocienta, ki jo je vpeljal Lancaster v [10].

Definicija 1.14 Naj bo vektor u približek za desni lastni vektor, vektor v približek za levi lastni vektor, λ pa približek za lastno vrednost nelinearnega problema lastnih vrednosti T . Lancasterjev kvocient je definiran kot

$$p_L(\lambda, u, v) = \lambda - \frac{v^H T(\lambda)u}{v^H T'(\lambda)u}.$$

Očitno velja $p_L(\lambda, cu, dv) = p_L(\lambda, u, v)$ za $c, d \neq 0$ in če je λ_0 enostavna lastna vrednost, x_0, y_0 pa desni in levi lastni vektor, potem je (λ_0, x_0, y_0) stacionarna točka za p_L , če ga opazujemo na območju $\bar{B}(\lambda, \tau_0) \times K_{\epsilon_0}(x_0) \times K_{\epsilon_0}(y_0)$ za dovolj majhen $\epsilon_0 > 0$ in ustrezen $\tau_0 > 0$.

Opazimo lahko tudi, da Lancasterjev kvocient ustreza enemu koraku Newtonove metode za reševanje nelinearne enačbe $f(\lambda) := v^H T(\lambda)u = 0$. To je dodatna motivacija, zakaj vzamemo $p_L(\lambda, u, v)$ za približek za lastno vrednost.

1.7 Ortogonalnost lastnih vektorjev

Če imamo simetrično matriko in sta x in y lastna vektorja, ki pripadata različnim lastnim vrednostima, potem je dobro znano dejstvo, da je $y^T x = 0$. Podobno za simetrični GEP $Ax =$

λBx , kjer je matrika B simetrična pozitivno definitna velja $y^T Bx$ za lastna vektorja različnih lastnih vrednosti.

Za splošno matriko prav tako vemo, da če je x desni lastni vektor za lastno vrednost μ , y pa levi lastni vektor za lastno vrednost $\lambda \neq \mu$, potem je $y^H x = 0$, pri splošnem GEP pa podobno dobimo $y^H Bx = 0$.

Ali se da to posplošiti na NEP? Če je T hermitski in so vse lastne vrednosti realne, potem so lastni vektorji različnih lastnih vrednosti ortogonalni v posplošenem skalarnem produktu

$$[x, y] := \begin{cases} \langle \Delta(p(x), p(y))x, y \rangle & \text{za } x, y \neq 0, \\ 0 & \text{za } x = 0 \text{ ali } y = 0, \end{cases}$$

kjer je p Rayleighov funkcional in Δ deljena diferenca, definirana z

$$\Delta(\alpha, \beta) = \begin{cases} \frac{1}{\alpha - \beta}(T(\alpha) - T(\beta)) & \text{za } \alpha \neq \beta \\ T'(\alpha) & \text{za } \alpha = \beta. \end{cases}$$

Za $[\cdot, \cdot]$ ne veljajo vse lastnosti skalarnega produkta. Tako lahko hitro preverimo, da velja:

- a) $[x, y] = \overline{[y, x]}$,
- b) $[x, x] = \langle T'(p(x))x, x \rangle > 0$,
- c) $[\alpha x, y] = \alpha[x, y]$.

Točka b) npr. sledi iz lastnosti d) o nadkritični dušenosti Rayleighovega funkcionala.

Ne velja pa bilinearnost, saj ne velja nujno $[x + y, z] = [x, z] + [y, z]$.

V primeru EP $T(\lambda) = \lambda I - A$ se posplošeni skalarni produkt $[\cdot, \cdot]$ ujema z navadnim skalarnim produktom, saj velja $\Delta(\alpha, \beta) \equiv I$ in $[x, y] = y^H x$. Podobno v primeru GEP $T(\lambda) = \lambda B - A$ velja $\Delta(\alpha, \beta) \equiv B$ in dobimo $[x, y] = y^H Bx$.

Če pa vzamemo nadkritično dušen QEP $T(\lambda) = \lambda^2 M + \lambda C + K$, potem dobimo

$$\Delta(\alpha, \beta) = (\alpha + \beta)M + C,$$

od koder sledi

$$[x, y] = (p(x) + p(y))y^H Mx + y^H Cx.$$

Vidimo, da za razliko od EP in GEP, pri QEP produkt $[\cdot, \cdot]$ ni več bilinearen.

Poglavje 2

Numerične metode za polne probleme

Dana je dovolj gladka preslikava $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$, ki predstavlja nelinearni probleme lastnih vrednosti. Iščemo rešitve $T(\lambda)x = 0$ za $\lambda \in \mathbb{C}$ in $x \in \mathbb{C}^n$, $x \neq 0$.

V tem poglavju bomo pogledali nekaj glavnih numeričnih metod za nelinearne problem lastnih vrednosti, ki so primerne, kadar je n zmerne velikosti in je matrika $T(\lambda)$ polna.

2.1 Newtonova metoda in inverzna iteracija

Če izberemo vektor $v \in \mathbb{C}^n$ in dodamo k enačbi $T(\lambda)x = 0$ še pogoj $v^H x = 1$, potem lahko zapišemo NEP v obliki nelinearnega sistema $n + 1$ enačb za $n + 1$ neznank:

$$F(x, \lambda) := \begin{bmatrix} T(\lambda)x \\ v^H x - 1 \end{bmatrix} = 0.$$

Za reševanje zgornjega sistema uporabimo Newtonovo metodo. Če je (x_k, λ_k) tekoči približek za lastni par, potem po Newtonovi metodi naslednji približek (x_{k+1}, λ_{k+1}) dobimo kot

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} - JF(x_k, \lambda_k)^{-1} F(x_k, \lambda_k),$$

kjer je Jacobijeva matrika enaka

$$JF(x_k, \lambda_k) = \begin{bmatrix} T(\lambda_k) & T'(\lambda_k)x_k \\ v^H & 0 \end{bmatrix}.$$

Tako pridemo do sistema

$$\begin{bmatrix} T(\lambda_k) & T'(\lambda_k)x_k \\ v^H & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)x_k \\ v^H x_k - 1 \end{bmatrix},$$

oziroma

$$\begin{aligned} T(\lambda_k)x_{k+1} &= -(\lambda_{k+1} - \lambda_k)T'(\lambda_k)x_k \\ v^H x_{k+1} &= 1. \end{aligned} \tag{2.1}$$

Iz zveze (2.1) sledi

$$x_{k+1} = -(\lambda_{k+1} - \lambda_k)T(\lambda_k)^{-1}T'(\lambda_k)x_k. \quad (2.2)$$

To pomeni, da iz zgornje enačbe lahko določimo pravo smer vektorja x_{k+1} . Več kot smer tudi ne potrebujemo, saj lahko potem x_{k+1} enostavno pomnožimo s pravim skalarjem, da bo $v^H x_{k+1} = 1$. Za naslednji korak potrebujemo še nov približek za lastno vrednost λ_{k+1} . Če definiramo

$$u_{k+1} := T(\lambda_k)^{-1}T'(\lambda_k)x_k$$

in enačbo (2.2) skalarno pomnožimo z v , potem dobimo

$$v^H x_{k+1} = -(\lambda_{k+1} - \lambda_k)v^H u_{k+1}.$$

Ob predpostavki, da je prejšnji približek za lastni vektor bil normiran, torej $v^H x_k = 1$, dobimo

$$\lambda_{k+1} = \lambda_k - \frac{v^H x_k}{v^H u_{k+1}}.$$

Tako pridemo do prvega algoritma. Čeprav gre za Newtonovo metodo, metodo zaradi podobnostjo s podobnim algoritmom za EP imenujemo *inverzna iteracija*. V nadaljevanju bomo videli, kako so v standardnem primeru EP povezane inverzna iteracija, Newtonova metoda in Rayleighova iteracija. Podobne povezave potem veljajo tudi za NEP.

Algoritem 2.1 Inverzna iteracija

izberi vektor v in tak začetni približek (λ_0, x_0) za lastni par, da je $v^H x_0 = 1$

$k = 0, 1, \dots$

a) reši linearni sistem $T(\lambda_k)u_{k+1} = T'(\lambda_k)x_k$

b) $\lambda_{k+1} = \lambda_k - \frac{v^H x_k}{v^H u_{k+1}}$

c) $x_{k+1} = \frac{1}{v^H u_{k+1}}u_{k+1}$

Izrek 2.1 Naj bo λ_* enostavna lastna vrednost nelinearnega problema lastnih vrednosti T , x_* pa pripadajoči lastni vektor, za katerega velja $v^H x_* \neq 0$. Potem inverzna iteracija za vsak začetni približek, ki je dovolj blizu (x_*, λ_*) skkonvergira proti λ_* , konvergenca pa je kvadratična.

Dokaz. Ker gre za Newtonovo metodo, je za konvergenco (in to vsaj kvadratično), dovolj pokazati, da je Jacobijeva matrika $JF(x_*, \lambda_*)$ nesingularna. Denimo, da obstaja taka točka $(u, \alpha) \neq (0, 0)$, da je $JF(x_*, \lambda_*) \begin{bmatrix} u \\ \alpha \end{bmatrix} = 0$, torej

$$\begin{bmatrix} T(\lambda_*) & T'(\lambda_*)x_* \\ v^H & 0 \end{bmatrix} \begin{bmatrix} u \\ \alpha \end{bmatrix} = 0. \quad (2.3)$$

Iz prve enačbe dobimo

$$T(\lambda_*)u + \alpha T'(\lambda_*)x_* = 0.$$

Ko to skalarno pomnožimo z y_* , dobimo $\alpha y_*^H T'(\lambda_*)x_* = 0$, kar nam da $\alpha = 0$, saj je lastna vrednost enostavna in zato $y_*^H T'(\lambda_*)x_* \neq 0$. Iz prve vrstice enačbe (2.3) lahko potem sklepamo,

da je $u = \beta x_*$, a potem, ko to vstavimo v drugo vrstico (2.3), dobimo $\beta = 0$, torej $u = 0$. To pomeni, da je Jacobijeva matrika $JF(\lambda_*, x_*)$ res nesingularna. $\square$

Pri inverzni iteraciji rešujemo sistem z matriko $T(\lambda)$. Če je λ blizu enostavne lastne vrednosti λ_* , potem lahko pokažemo, da bo vektor $T(\lambda)^{-1}x$ imel močno komponento v smeri lastnega vektorja x_* . Tu se skličemo na naslednjo lemo iz [17].

Lema 2.2 *Naj bo preslikava $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ dvakrat zvezno odvedljiva na $B(\lambda_*, \tau_*)$ za $\tau_* > 0$ in naj bo (λ_*, x_*, y_*) enostavna lastna trojica. Potem obstaja $0 \leq \tau_0 \leq \tau_*$, da za vse $\tau = \lambda - \lambda_*$, $0 \leq |\tau| \leq \tau_0$, velja*

$$T(\lambda)^{-1} = \frac{1}{\lambda - \lambda_*} \left[\frac{x_* y_*^H}{y_*^H T'(\lambda_*) x_*} + \mathcal{O}(\tau) \right].$$

Posledica tega je, da če uporabimo $T(\lambda)^{-1}$ na vektorju z , za katerega velja $y_*^H z \neq 0$, dobimo

$$w = T(\lambda)^{-1}z = \frac{y_*^H z}{(\lambda - \lambda_*) y_*^H T'(\lambda_*) x_*} + \mathcal{O}(1).$$

To pomeni, da ima pri pogoju $y_*^H z \neq 0$ vektor w dominantno komponento v smeri x_* .

Od tod sledi, da je pametno za vektor z izbrati $z = T'(\lambda)x$. Če je namreč (x, λ) blizu (x_*, λ_*) , kjer je zaradi enostavnosti lastne vrednosti $y_*^H T'(\lambda_*) x_* \neq 0$, bo pogoj $y_*^H z \neq 0$ izpolnjen. Iz te razprave sledi dodatna razlaga, zakaj bo vektor u_{k+1} iz točke a) algoritma 2.1 v primeru, ko je λ_k blizu λ_* , zelo dobra aproksimacija za lastni vektor.

Algoritem bi lahko še modificirali. Tako bi lahko npr. namesto fiksnega vektorja v v vsakem koraku uporabili drug vektor za normiranje vektorja x_k . Prav tako bi lahko nov približek za lastno vrednost λ_{k+1} v koraku b) izračunali na kak drug način, lahko bi npr. vzeli kar Rayleighov funkcional, če ga imamo na voljo.

Zgled 2.1 Če vzamemo EP $T(\lambda) = \lambda I - A$, potem v koraku a) algoritma 2.1 rešimo sistem

$$(\lambda_k I - A)u_{k+1} = x_k,$$

od koder sledi

$$Au_{k+1} = \lambda_k u_{k+1} - x_k.$$

Za λ_{k+1} , izračunano v koraku b), potem velja

$$\lambda_{k+1} = \lambda_k - \frac{v^H x_k}{v^H u_{k+1}} = \frac{v^H (\lambda_k u_{k+1} - x_k)}{v^H u_{k+1}} = \frac{v^H A u_{k+1}}{v^H u_{k+1}},$$

kar pa je dvostranski Rayleighov kvocient. Če je A hermitska in vzamemo $v_k = u_k$, potem pridemo do standardne Rayleighove iteracije, za katero vemo, da ima v primeru hermitske matrike kubično konvergenco, sicer pa kvadratično. $\square$

Če imamo na voljo tudi približek za levi lastni vektor y_k , lahko za v_k vzamemo $v_k = T(\lambda_k)^H y_k$. V tem primeru bo pri enostavni lastni vrednosti pogoj $v_k^H x_k \neq 0$ izpolnjen. Tako dobimo

$$\lambda_{k+1} = \lambda_k - \frac{y_k^H T(\lambda_k) x_k}{y_k^H T'(\lambda_k) x_k},$$

kar je ravno Lancasterjev kvocient.

V koraku b) lahko vrednost λ_{k+1} izračunamo iz Rayleighovega funkcionala kot $\lambda_{k+1} = p(u_{k+1})$. Če to naredimo za realni simetrični primer z realnim lastnim parom, dobimo kubično konvergenco.

2.2 Inverzna iteracija z residuali

Velika težava pri inverzni iteraciji za NEP je, da moramo v vsakem koraku za reševanje sistema $T(\lambda_k)u_{k+1} = T'(\lambda_k)x_k$ izračunati LU razcep matrike $T(\lambda_k)$, za kar v splošnem primeru porabimo $\mathcal{O}(n^3)$ operacij.

V primeru EP se temu lahko izognemo, če fiksiramo λ in namesto Rayleighove iteracije delamo le inverzno iteracijo. Konvergenca je sicer samo linearna, a razcep $T(\lambda)$ moramo izračunati le enkrat.

Če fiksiramo λ_0 , potem v primeru EP dobimo za $T(\lambda) = \lambda I - A$ naslednji postopek

$$\begin{aligned} k &= 0, 1, \dots \\ \text{reši } (A - \lambda_0 I)u_{k+1} &= -x_k \\ x_{k+1} &= (1/\|u_{k+1}\|_\infty)u_{k+1} \end{aligned}$$

Zaporedje vektorjev x_k konvergira proti lastnemu vektorju x_* , ki pripada lastni vrednosti λ_* , ki je najbližja λ_0 .

Pri NEP tega ne moremo narediti. Če fiksiramo λ_0 in rešujemo v vsakem koraku sistem

$$T(\lambda_0)u_{k+1} = T'(\lambda_k)x_k,$$

potem zaporedje $\{x_k\}$ po smeri skonvergira k lastnemu vektorju posplošenega problema lastnih vrednosti $T(\lambda_0)x = T'(\hat{\lambda})x$, kjer je $\hat{\lambda}$ limita zaporedja $\{\lambda_k\}$. Iz dobljenega vektorja se ne da rekonstruirati lastne vrednosti začetnega NEP.

Neumaier je v [14] ta problem rešil na naslednji način. Denimo, da je preslikava T dvakrat zvezno odvedljiva. Po algoritmu za inverzno iteracijo dobimo

$$\begin{aligned} x_k - x_{k+1} &= x_k + (\lambda_{k+1} - \lambda_k)T(\lambda_k)^{-1}T'(\lambda_k)x_k \\ &= T(\lambda_k)^{-1}[T(\lambda_k) + (\lambda_{k+1} - \lambda_k)T'(\lambda_k)]x_k \\ &= T(\lambda_k)^{-1}[T(\lambda_{k+1}) + \mathcal{O}(|\lambda_{k+1} - \lambda_k|^2)]x_k \\ &= T(\lambda_k)^{-1}T(\lambda_{k+1})x_k + \mathcal{O}(|\lambda_{k+1} - \lambda_k|^2). \end{aligned}$$

Pri tem pristopu zamenjava λ_k s fiksnim σ ne pokvari konvergence.

Da lahko po zgornji formuli izračunamo $x_k - x_{k+1}$, potrebujemo najprej λ_{k+1} . Neumaier je predlagal naslednji možnosti:

- a) Če je problem hermitski, potem vzamemo $\lambda_{k+1} = p(x_k)$.
- b) Če problem ni hermitski, vzamemo za λ_{k+1} tisto rešitev enačbe

$$v^H T(\sigma)^{-1} T(\lambda_{k+1}) x_k = 0, \tag{2.4}$$

ki je najbližja λ_k . Tu si lahko predstavljamo, da je $v^H T(\sigma)^{-1}$ aproksimacija y_k^H za levi lastni vektor. Če za reševanje enačbe (2.4) uporabimo en korak Newtonove metode z začetnim približkom λ_k , dobimo

$$\lambda_{k+1} = \lambda_k - \frac{y_k^H T(\lambda_k) x_k}{y_k^H T'(\lambda_k) x_k},$$

kar je ravno Lancasterjeva posplošitev Rayleighovega kvocienta.

Algoritem 2.2 Inverzna iteracija z residuali

izberi vektor v in tak začetni približek (σ, x_0) za lastni par, da je $v^H x_0 = 1$
 $k = 0, 1, \dots$

izračunaj λ_{k+1} :

a) če je T hermitski in $\lambda_k \in \mathbb{R}$, potem $\lambda_{k+1} = p(x_k)$, sicer

b) reši $v^H T(\sigma)^{-1} T(\lambda_{k+1}) x_k = 0$.

$$r_k = T(\lambda_{k+1}) x_k$$

$$\text{reši } T(\sigma) \Delta x_k = r_k$$

$$z_{k+1} = x_k - \Delta x_k$$

$$x_{k+1} = \frac{1}{v^H z_{k+1}} z_{k+1}$$

Izrek 2.3 Naj bo λ_* enostavna lastna vrednost nelinearnega problema lastnih vrednosti T , kjer je preslikava $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ dvakrat zvezno odvedljiva, x_* pa naj bo pripadajoči lastni vektor, normiran tako, da je $v^H x_* = 1$. Potem inverzna iteracija z residuali konvergira za vse σ dovolj blizu λ_* in velja

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = \mathcal{O}(|\sigma - \lambda_*|)$$

in

$$|\lambda_{k+1} - \lambda_*| = \mathcal{O}(|\sigma - \lambda_*|^p),$$

kjer je $p = 2$, če je T hermitski, $\lambda_* \in \mathbb{R}$ in vzamemo $\lambda_{k+1} = p(x_k)$, sicer pa je $p = 1$.

2.3 Zaporedne linearne aproksimacije

Naj bo preslikava $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ spet dvakrat zvezno odvedljiva. Denimo, da že imamo približek λ_k za lastno vrednost. Iščemo popravek $\Delta \lambda_k$ in vektor $x \neq 0$, da bo

$$T(\lambda_k - \Delta \lambda_k) x = 0.$$

Če zgornjo enačbo razvijemo v vrsto, dobimo

$$(T(\lambda_k) - \Delta \lambda_k T'(\lambda_k) + \mathcal{O}(|\Delta \lambda_k|^2)) x = 0.$$

Če zanemarimo kvadratne člene, dobimo, da je popravek $\Delta \lambda_k$ lastna vrednost posplošenega problema lastnih vrednosti

$$T(\lambda_k) x = \Delta \lambda_k T'(\lambda_k) x. \quad (2.5)$$

Pri metodi zaporednih linearnih aproksimacij za $\Delta\lambda_k$ vzamemo po absolutni vrednosti najmanjšo lastno vrednost (2.5). V bližini lastne vrednosti je metoda podobna inverzni iteraciji, saj dobimo

$$\frac{1}{\Delta\lambda_k}x = T(\lambda_k)^{-1}T'(\lambda_k)x.$$

Algoritem 2.3 Zaporedne linearne aproksimacije

izberi začetni približek λ_0 za lastno vrednost

$k = 0, 1, \dots$

za $\Delta\lambda_k$ vzemi po absolutni vrednosti najmanjšo lastno vrednost GEP $T(\lambda_k)x = \theta T'(\lambda_k)x$

$\lambda_{k+1} = \lambda_k - \Delta\lambda_k$

Dela v enem koraku je več kot pri prejšnjih dveh metodah, saj moramo v vsakem koraku rešiti GEP, je pa konvergenca zato ponavadi hitrejša.

2.4 Dvostranske različice inverzne iteracije

Iz algoritmov 2.1 in 2.2 lahko preprosto razvijemo dvostranske metode, kjer računamo hkrati tudi približke za levi lastni vektor.

Algoritem 2.4 Dvostranska inverzna iteracija

izberi začetni približek (λ_0, x_0, y_0) za lastno trojico, da je $x_0^H x_0 = y_0^H y_0 = 1$

$k = 0, 1, \dots$

reši $T(\lambda_k)u_{k+1} = T'(\lambda_k)x_k$

reši $T(\lambda_k)^H v_{k+1} = T'(\lambda_k)^H y_k$

$$x_{k+1} = \frac{1}{\|u_{k+1}\|} u_{k+1}$$

$$y_{k+1} = \frac{1}{\|v_{k+1}\|} v_{k+1}$$

izračunaj $\lambda_{k+1} = p(x_{k+1}, y_{k+1})$

Stroški pri dvostranski inverzni iteraciji, ki je zapisana v Algoritmu 2.4, niso bistveno večji kot pri enostranski inverzni iteraciji. Najzahtevnejša operacija, ki jo moramo pri obeh algoritmihi v vsakem koraku izračunati le enkrat, je namreč faktorizacija matrike $T(\lambda_k)$.

Prednost dvostranske metode je, da je (x_*, y_*) stacionarna točka dvostranskega Rayleighovega funkcionala, kar v splošnem za enostranskega ne velja.

Red konvergence dvostranske inverzne iteracije v bližini enostavne lastne vrednosti je kubičen.

Red konvergence dvostranske inverzne iteracije z residuali, ki je zapisana v Algoritmu 2.5, je $(1 + \sqrt{5})/2$, kar pomeni, da je konvergenca superlinearna.

Algoritem 2.5 Dvostranska inverzna iteracija z residuali

izberi vektorja v, w in tak začetni približek (σ, x_0, y_0) za lastno trojico, da je $v^H x_0 = w^H y_0 = 1$
 $k = 0, 1, \dots$

izračunaj $\lambda_{k+1} = p(x_k, y_k)$

$r_k = T(\lambda_{k+1})x_k$

$s_k = T(\lambda_{k+1})^H y_k$

reši $T(\sigma)\Delta x_k = r_k$

reši $T(\sigma)^H \Delta y_k = s_k$

$z_{k+1} = x_k - \Delta x_k$

$u_{k+1} = y_k - \Delta y_k$

$x_{k+1} = \frac{1}{v^H z_{k+1}} z_{k+1}$

$y_{k+1} = \frac{1}{w^H u_{k+1}} u_{k+1}$

2.5 QR metoda Kublanovskaje

Pri metodi Kublanovskaje rešujemo nelinearno enačbo $\det(T(\lambda)) = 0$. Naj bo

$$T(\lambda)P(\lambda) = Q(\lambda)R(\lambda) \quad (2.6)$$

QR razcep matrike $T(\lambda)$ s pivotiranjem, kjer je $P(\lambda)$ permutacijska matrika, $Q(\lambda)$ unitarna matrika, $R(\lambda)$ pa zgornja trikotna matrika. Pri pivotiranju vrstni red stolpcev izbiramo tako, da velja

$$|r_{11}(\lambda)| \geq |r_{22}(\lambda)| \geq \dots \geq |r_{nn}(\lambda)|.$$

Očitno je $|\det(T(\lambda))| = |\det(R(\lambda))|$, lastno vrednost pa imamo takrat, ko je $r_{nn}(\lambda) = 0$.

Definiramo $f(\lambda) := r_{nn}(\lambda)$ in za računanje ničle funkcije f uporabimo tangentno metodo. Pri odvajanju zveze (2.6) predpostavimo, da je $P(\lambda)$ konstantna matrika. Dobimo

$$T'(\lambda)P(\lambda) = Q'(\lambda)R(\lambda) + Q(\lambda)R'(\lambda),$$

od koder sledi

$$R'(\lambda) = Q(\lambda)^H T'(\lambda)P(\lambda) - Q(\lambda)^H Q'(\lambda)R(\lambda)$$

in

$$R'(\lambda)R(\lambda)^{-1} = Q(\lambda)^H T'(\lambda)P(\lambda)R(\lambda)^{-1} - Q(\lambda)^H Q'(\lambda). \quad (2.7)$$

Vsi diagonalni elementi matrike $Q(\lambda)^H Q'(\lambda)$ so enaki 0, saj z odvajanjem $Q(\lambda)^H Q(\lambda) = I$ dobimo

$$Q(\lambda)^H Q'(\lambda) + Q'(\lambda)^H Q(\lambda) = Q(\lambda)^H Q'(\lambda) + (Q(\lambda)^H Q'(\lambda))^H = 0.$$

To pomeni, da je matrika $Q(\lambda)^H Q'(\lambda)$ poševno hermitska in ima ničelne diagonalne elemente.¹

Ker je matrika $R(\lambda)$ zgornja trikotna, je element v spodnjem desnem kotu matrike $R'(\lambda)R(\lambda)^{-1}$ enak $f'(\lambda)/f(\lambda)$, torej iz (2.7) dobimo

$$\frac{f'(\lambda)}{f(\lambda)} = \frac{r'_{nn}(\lambda)}{r_{nn}(\lambda)} = e_n^T Q(\lambda)^H T'(\lambda)P(\lambda)R(\lambda)^{-1} e_n.$$

Algoritem 2.6 QR metoda Kublanovskaje

 izberi začetni približek λ_0 za lastno vrednost
 $k = 0, 1, \dots$

$$T(\lambda_k)P_k = Q_k R_k \text{ (QR razcep s pivotiranjem)}$$

$$\lambda_{k+1} = \lambda_k - \frac{1}{e_n^T Q_k^H T'(\lambda_k) P_k R_k^{-1} e_n}$$

Opazimo lahko, da se približek λ_{k+1} iz Algoritma 2.6 ujema z Lancasterjevo posplošitvijo Rayleighovega kvocienta. Če vzamemo $y_k = Q_k e_n$ in $x_k = P_k R_k^{-1} e_n$, kar si lahko predstavljamo kot tekoča približka za levi in desni lastni vektor, potem velja $\lambda_{k+1} = p_L(\lambda_k, x_k, y_k)$.

2.6 LU metoda Bohteta

Medtem, ko prejšnja metoda temelji na QR razcepu, je Bohte razvil podobno metodo, ki temelji na LU razcepu in je zato cenejša. Naj bo

$$P(\lambda)T(\lambda) = L(\lambda)U(\lambda) \quad (2.8)$$

LU razcep matrike $T(\lambda)$ z delnim pivotiranjem, kjer je $P(\lambda)$ permutacijska matrika, $L(\lambda)$ spodnja trikotna matrika z enicami na diagonali in $U(\lambda)$ zgornja trikotna matrika. Z odvajanjem (2.8), kjer tako kot prej predpostavimo, da je $P(\lambda)$ konstantna matrika, dobimo

$$P(\lambda)T'(\lambda) = M(\lambda)U(\lambda) + L(\lambda)V(\lambda), \quad (2.9)$$

kjer je $M(\lambda) := L'(\lambda)$ spodnja trikotna matrika z ničlami na diagonali, $V(\lambda) := U'(\lambda)$ pa zgornja trikotna matrika. Izkaže se, da se da iz sistema (2.9) v pravilnem vrstnem redu izračunati vse elemente matrik $M(\lambda)$ in $V(\lambda)$.

Če definiramo $f(\lambda) := \det(T(\lambda))$, potem velja

$$\begin{aligned} f(\lambda) &= \det(P(\lambda)) \prod_{i=1}^n u_{ii}(\lambda) \\ f'(\lambda) &= \det(P(\lambda)) \sum_{j=1}^n v_{jj}(\lambda) \prod_{\substack{i=1 \\ i \neq j}}^n u_{ii}(\lambda), \end{aligned}$$

od koder sledi

$$\frac{f'(\lambda)}{f(\lambda)} = \sum_{j=1}^n \frac{v_{jj}(\lambda)}{u_{jj}(\lambda)}.$$

Za iskanje ničle funkcije f uporabimo tangentno metodo. Bohtetova metoda je prikazana v algoritmu 2.7.

¹To velja le za realne matrike, pri kompleksnih bi še vedno lahko bili diagonalni elementi matrike $Q(\lambda)^H Q'(\lambda)$ čisto imaginarni. Kako se lahko znebimo tega?

Algoritem 2.7 LU metoda Bohteta

 izberi začetni približek λ_0
 $k = 0, 1, \dots$
 $P_k T(\lambda_k) = L_k U_k$ (LU razcep z delnim pivotiranjem)

 iz $P_k T'(\lambda_k) = M_k U_k + L_k V_k$ izračunaj M_k in V_k

$$\lambda_{k+1} = \lambda_k - \frac{1}{\sum_{j=1}^n \frac{v_{jj}(\lambda)}{u_{jj}(\lambda)}}$$

2.7 Zajamčena iteracija

V posebnih primerih lahko minimaks izrek posplošimo na nelinearne probleme lastnih vrednosti in tako dobimo možnost, da izračunamo direktno le lastno vrednost z izbranim indeksom.

Denimo, da imamo za hermitski NEP podan Rayleighov funkcional z naslednjimi lastnosti:

- a) $p : \mathbb{C}^n \setminus \{0\} \rightarrow J \subset \mathbb{R}$, kjer je J odprt interval, lahko tudi neskončen,
- b) $p(\gamma u) = p(u)$ za vsak $\gamma \neq 0$,
- c) $u^H T(p(u))u = 0$,
- d) $u^H T'(p(u))u > 0$.

Predpostavimo še, da ima za vsak $x \in \mathbb{C}^n \setminus \{0\}$ enačba $x^H T(\lambda)x = 0$ največ eno rešitev $\lambda \in J$.

Lema 2.4 Naj bo $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ hermitski nelinearni problem lastnih vrednosti z realnimi lastnimi vrednostmi in naj za Rayleighov funkcional veljajo zgoraj našteje predpostavke. Naj bo $\lambda \in W(T)$ in $x \neq 0$. Potem velja

$$p(u) \begin{cases} < \\ = \\ > \end{cases} \lambda \iff u^H T(\lambda)u \begin{cases} > \\ = \\ < \end{cases} 0.$$

Dokaz. Zveza $p(u) = \lambda \iff u^H T(\lambda)u = 0$ sledi iz same definicije Rayleighovega funkcionala in dejstva, da ima enačba $u^H T(\lambda)u = 0$ največ eno rešitev v J .

Zaradi $u^H T'(p(u))u > 0$ je funkcija $u^H T(\sigma)u$, kjer je u fiksni, pri $\sigma = p(u)$ naraščajoča. Od tod sledi $p(u) > \lambda \iff u^H T(\lambda)u < 0$ in $p(u) < \lambda \iff u^H T(\lambda)u > 0$. ■

Če je funkcional p z zgornjimi lastnostmi od a) do d) definiran za vse neničelne vektorje iz $\mathbb{C}^n$, potem pravimo, da je p nadkritično dušen (angl. overdamped). V tem primeru velja naslednji izrek.

Izrek 2.5 (Rogers) Naj bo $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ hermitski nelinearni problem lastnih vrednosti z realnimi lastnimi vrednostmi in naj bo p nadkritično dušen Rayleighov funkcional. Potem obstaja n lastnih vrednosti problema T , urejenih kot $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ (kjer lastne vrednosti štejemo z geometrijskimi večkratnostmi) in n linearno neodvisnih lastnih vektorjev $x_1, \dots, x_n$, da je $T(\lambda_j)x_j = 0$ za $j = 1, \dots, n$. Za lastne vrednosti velja minimaks lastnost

$$\lambda_j = \min_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=j}} \max_{\substack{x \in V \\ x \neq 0}} p(x)$$

za $j = 1, \dots, n$.

Opomba 2.1 Problem T ima lahko več kot n lastnih vrednosti. Zgornji izrek govori le o podmnožici n lastnih vrednosti. Tako npr. pri nadkritično dušenem QEP obstajata dve skupini lastnih vrednosti (primarna in sekundarna), vsaki pa pripada n linearno neodvisnih lastnih vektorjev.

Kadar tak funkcional ne obstaja na celotni množici $\mathbb{C}^n \setminus \{0\}$, lahko še lokalno vedno posplošimo nekaj lastnosti. Denimo, da funkcional z lastnostmi a) do d) obstaja na odprti podmnožici $D \subset \mathbb{C}^n$. Potem lahko definiramo indeks lastne vrednosti na naslednji način.

Vemo, da je λ lastna vrednost NEP T natanko tedaj, ko je $\mu = 0$ lastna vrednost EP $T(\lambda)x = \mu x$. Ker je $T(\lambda)$ hermitska matrika, so njene lastne vrednosti realne in obstaja tak indeks m , da je

$$0 = \max_{\substack{V \subset \mathbb{C}^n \\ \dim(V)=m}} \min_{\substack{x \in V \\ x \neq 0}} \frac{x^H T(\lambda) x}{x^H x}.$$

Ta m proglasimo za indeks lastne vrednosti λ in pravimo, da je λ m -ta lastna vrednost T . Ekvivalentno je, da je λ m -ta lastna vrednost T natanko tedaj, ko so lastne vrednosti $T(\lambda)$ oblike $\mu_n \leq \dots \leq \mu_m = 0 \leq \dots \leq \mu_1$.

Opomba 2.2 Če vzamemo hermitsko matriko A , potem smo navajeni njene lastne vrednosti označevati tako, da velja $\lambda_n \leq \dots \leq \lambda_1$. Pri NEP so indeksi ravno obrnjeni, saj velja $\lambda_1 \leq \dots \leq \lambda_n$, to pa zato, ker obravnavamo $T(\lambda) = \lambda I - A$.

Izrek 2.6 (Voss in Werner) Za vsak $m \in \{1, \dots, n\}$ ima nelinearni problem lastnih vrednosti T kvečjemu eno m -to lastno vrednost na J , za katero velja

$$\lambda_m = \min_{\substack{V \subset \mathbb{C}^n, V \cap D \neq \emptyset \\ \dim(V)=m}} \sup_{\substack{x \in V \cap D \\ x \neq 0}} \frac{x^H T(\lambda) x}{x^H x}. \quad (2.10)$$

Obratno, če je

$$\lambda = \inf_{\substack{V \subset \mathbb{C}^n, V \cap D \neq \emptyset \\ \dim(V)=m}} \sup_{\substack{x \in V \cap D \\ x \neq 0}} \frac{x^H T(\lambda) x}{x^H x} \in J,$$

potem je λ m -ta lastna vrednost in velja (2.10). Minimum je dosežen pri invariantnem podprostoru $T(\lambda_m)$, ki pripada m po absolutni vrednosti največjim lastnim vrednostim, supremum pa je dosežen za vsak lastni vektor, ki ustreza lastni vrednosti 0.

Algoritem 2.8 Zajamčena iteracija

izberi začetni približek λ_0 za lastno vrednost

$k = 0, 1, \dots$

določi lastni vektor x_k , ki pripada m -ti največji lastni vrednosti $T(\lambda_k)$

reši $x_k^H T(\lambda_{k+1}) x_k = 0$ za λ_{k+1}

Tako dobimo algoritem 2.8. En korak metode si lahko predstavljamo kot navadno iteracijo $\lambda_{k+1} = g(\lambda_m)$, kjer je iteracijska funkcija g definirana kot $g(\lambda) = p(x_m(\lambda))$, kjer je $x_m(\lambda)$ m -ti

lastni vektor. Očitno je λ_m (če obstaja) fiksna točka, saj je $T(\lambda_m)x_m(\lambda_m) = 0$, od tod pa sledi $p(x_m(\lambda_m)) = \lambda_m$.

Za to iteracijo se da pokazati, da je $g'_m(\lambda_m) = 0$, kar pomeni, da je konvergenca vsaj kvadratična. Če pa je $T'(\lambda)$ pozitivno definitna matrika za vsak $\lambda \in J$ in namesto lastnega vektorja, ki pripada m -to lastni vrednosti $T(\lambda_k)$ vzamemo lastni vektor, ki pripada m -ti lastni vrednosti $T(\lambda_k)x = \mu T'(\lambda_k)x$, potem dobimo celo kubično konvergenco.

Poglavje 3

Polinomski (kvadratni) problem lastnih vrednosti

3.1 Uvod

Pri polinomskem problemu lastnih vrednosti (PEP) je dan matrični polinom $P(\lambda) = A_0 + \lambda A_1 + \dots + \lambda^m A_m$, kjer so $A_0, \dots, A_m$ matrike velikosti $n \times n$. Iščemo tak skalar $\lambda_0 \in \mathbb{C}$ in neničelni vektor $x_0 \in \mathbb{C}^n$, da je $P(\lambda_0)x_0 = 0$.

PEP P je regularen, če njegov karakteristični polinom, definiran z $g(\lambda) := \det(P(\lambda))$, ni identično enak 0. Ničle polinoma g so končne lastne vrednosti PEP P . Če jih je manj kot mn , jih do mn dopolnimo z neskončnimi lastnimi vrednostmi. Neskončne lastne vrednosti ustrezajo ničelnim lastnim vrednostim *vzvratnega polinoma* $P_R(\lambda) := \lambda^m P(1/\lambda) = \lambda^m A_0 + \lambda^{m-1} A_1 + \dots + A_m$, pojavijo pa se le, če je matrika A_m nesingularna.

Regularni PEP P ima tako mn (končnih ali neskončnih) lastnih vrednosti.

Zgled 3.1 Naslednji primer kvadratnega problema lastnih vrednosti je povzet iz [19].

Če vzamemo $Q(\lambda) = \lambda^2 M + \lambda C + K$, kjer so

$$M = \begin{bmatrix} 0 & 6 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & -6 & 0 \\ 2 & -7 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

potem je $\det Q(\lambda) = -6\lambda^5 + 11\lambda^4 - 12\lambda^3 + 12\lambda^2 - 6\lambda + 1$ in problem je regularen. Lastni pari so

k	1	2	3	4	5	6
λ_k	$1/3$	$1/2$	1	i	$-i$	∞
x_k	$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Pet lastnih vrednosti je končnih, ena pa neskončna. Opazimo, da imata različni lastni vrednosti lahko isti lastni vektor, kar pri standardnem in posplošenem lastnem problemu seveda ni možno.

Lahko se zgodi, da si največ m različnih lastnih vrednosti deli isti lastni vektor. Namreč, če je x lastni vektor, potem je vsaj ena izmed rešitev enačbe $x^H P(\lambda)x = 0$, ki ima največ m rešitev, enaka lastni vrednosti.

Če niso vse lastne vrednosti polenostavne (imajo geometrijsko večkratnost enako algebrski), potem se srečamo z Jordanovimi verigami.

Definicija 3.1 Množica vektorjev $x_0, x_1, \dots, x_k$, kjer je $x_0 \neq 0$, za katero velja

$$\sum_{j=0}^p \frac{1}{j!} P^{(j)}(\lambda_0) x_{p-j} = 0$$

za $p = 0, 1, \dots, k$, je Jordanova veriga dolžine $k + 1$ za lastno vrednost λ_0 .

Nakaj prvih zvez (za $p = 0, 1, 2$) je

$$\begin{aligned} P(\lambda_0)x_0 &= 0 \\ P(\lambda_0)x_1 + P'(\lambda_0)x_0 &= 0 \\ P(\lambda_0)x_2 + P'(\lambda_0)x_1 + \frac{1}{2!}P''(\lambda_0)x_0 &= 0. \end{aligned}$$

Zgled 3.2 Če vzamemo navadni problem lastnih vrednosti $P(\lambda) = \lambda I - A$, potem se Jordanove verige P ujemajo z Jordanovimi verigami matrike A , saj mora za korenske in lastne vektorje veljati $P(\lambda_0)x_0 = 0$ in $(\lambda_0 I - A)x_k + x_{k-1} = 0$ oziroma $(A - \lambda_0 I)x_k = x_{k-1}$ za $k > 1$.

Pri Jordanovi verigi PEP vektorji $x_0, \dots, x_{k-1}$ niso nujno linearno neodvisni, lahko se celo zgodi, da je kateri izmed vektorjev $x_1, \dots, x_{k-1}$ enak 0.

Zgled 3.3 Naj bo $P(\lambda) = \begin{bmatrix} \lambda^2 & -\lambda \\ 0 & \lambda^2 \end{bmatrix}$. Potem je $\det(P(\lambda)) = \lambda^4$ in edina lastna vrednost je $\lambda_0 = 0$.

Vsak neničelen vektor iz $\mathbb{C}^2$ je lastni vektor. Naj bo lastni vektor $x_0 = [x_{01} \ x_{02}]^T$. Če obstaja, potem mora prvi korenski vektor x_1 zadoščati enačbi $P(\lambda_0)x_1 + P'(\lambda_0)x_0 = 0$ oziroma

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} = 0.$$

Če je $x_{02} = 0$, potem je x_{01} poljuben in x_1 je lahko poljuben vektor. Če pa je $x_{02} \neq 0$, potem vektor x_1 ne obstaja.

Drugi korenski vektor x_2 (sedaj lahko že predpostavimo, da je $x_{02} = 0$) mora zadoščati $P(\lambda_0)x_2 + P'(\lambda_0)x_1 + \frac{1}{2}P''(\lambda_0)x_0 = 0$ oziroma

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{01} \\ 0 \end{bmatrix} = 0.$$

Od tod sledi $x_{01} = x_{12}$, medtem ko sta x_{11} in x_2 poljubna.

Za tretji korenski vektor dobimo enačbo

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{31} \\ x_{32} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} = 0.$$

Od tod sledi $x_{12} = x_{01} \neq 0$, v primeru $x_{12} = 0$ pa sistem ni rešljiv.

Torej ima PEP verige treh dolžin:

- a) verige dolžine 1 vsebujejo vektor $x_0 = \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$, kjer x_{01} in x_{02} nista oba enaka 0,
- b) verige dolžine 2 vsebujejo vektor $x_0 = \begin{bmatrix} x_{01} \\ 0 \end{bmatrix}$ in poljuben vektor x_1 ,
- c) verige dolžine 3 vsebujejo vektorja $x_0 = \begin{bmatrix} x_{01} \\ 0 \end{bmatrix}$, $x_1 = \begin{bmatrix} x_{11} \\ x_{01} \end{bmatrix}$ in poljuben vektor x_2 .

3.2 Povezava z diferencialnimi enačbami

Polinomski problemi lastnih vrednosti so povezani s sistemi linearnih diferencialnih enačb. Denimo, da je dana diferencialna enačba

$$L\left(\frac{d}{dt}\right)u(t) = f(t),$$

kjer je

$$L\left(\frac{d}{dt}\right)u(t) = \sum_{k=0}^m A_k u^{(k)}(t).$$

Pri tem je $u(t) \in \mathbb{C}^n$, matrike $A_0, \dots, A_m$ pa so velikosti $n \times n$.

Diferencialna enačba je povezana s polinomskim problemom lastnih vrednosti

$$L(\lambda)x = \sum_{k=0}^m \lambda^k A_k x = 0.$$

V primeru, ko so vse lastne vrednosti PEP L enostavne, lahko splošno rešitev homogene enačbe

$$L\left(\frac{d}{dt}\right)u(t) = 0 \text{ zapišemo v obliki}$$

$$u_h(t) = \sum_{k=1}^{mn} \alpha_k e^{\lambda_k t} x_k,$$

kjer je $L(\lambda_k)x_k = 0$ in je α_k poljubna konstanta za $k = 1, \dots, mn$.

Če je f oblike $f(t) = e^{i\omega_0 t} f_0$, kar ustreza vsiljenemu nihanju, potem je partikularna rešitev podana z

$$u_p(t) = e^{i\omega_0 t} \sum_{k=1}^{mn} \frac{y_k^H f_0}{i\omega_0 - \lambda_k} x_k,$$

kjer so $y_1, \dots, y_{mn}$ ustrezni levi lastni vektorji. V primeru, ko se $i\omega_0$ približa eni izmed lastnih vrednosti λ_k , lahko člen $\frac{y_k^H f_0}{i\omega_0 - \lambda_k}$ postane zelo velik in pride do t.i. resonance.

Če vse lastne vrednosti niso enostavne, potem v rešitvi nastopajo vektorji, ki tvorijo Jordanove verige.

Izrek 3.2 *Vektorska funkcija*

$$u(t) = \left(\frac{t^k}{k!} x_0 + \frac{t^{k-1}}{(k-1)!} x_1 + \cdots + x_k \right) e^{\lambda_0 t},$$

kjer je $x_0 \neq 0$, je rešitev diferencialne enačbe $L\left(\frac{d}{dt}\right) u(t) = 0$ natanko tedaj, ko velja

$$\sum_{j=0}^p \frac{1}{j!} L^{(j)}(\lambda_0) x_{p-j} = 0$$

za $p = 0, \dots, k$.

Vidimo, da vektorji $x_0, \dots, x_k$, ki nastopajo v zgornjem izreku, tvorijo Jordanovo verigo za lastno vrednost λ_0 . Naslednji izrek nam da novo lastnost takšnih vektorjev.

Izrek 3.3 *Vektorji $x_0, \dots, x_{k-1}$ tvorijo Jordanovo verigo za polinomski problem lastnih vrednosti L za lastno vrednost λ_0 natanko tedaj, ko je $x_0 \neq 0$ in*

$$A_0 X_0 + A_1 X_0 J_0 + \cdots + A_m X_0 J_0^m = 0,$$

kjer je $X_0 = [x_0 \cdots x_{k-1}]$ in je J_0 Jordanova kletka velikosti $k \times k$ oblike

$$J_0 = \begin{bmatrix} \lambda_0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix}.$$

Predpostavimo, da je matrika A_m nesingularna. Iz vseh Jordanovih verig lahko potem sestavimo bločno diagonalno $mn \times mn$ matriko $J = \text{diag}(J_1, \dots, J_l)$, kjer je J_i Jordanova kletka velikosti $n_i \times n_i$ in je $n_1 + \cdots + n_l = mn$. Iz lastnih in korenskih vektorjev sestavimo matriko X velikosti $n \times mn$, da je $X = [X_1 \cdots X_l]$, kjer stolpci matrike $X_i = [x_0^i \cdots x_{n_i-1}^i]$ sestavljajo Jordanovo verigo, ki pripada kletki J_i . Tako dobimo *Jordanov par* (J, X) za PEP L . Velja, da je $mn \times mn$ matrika

$$\begin{bmatrix} X \\ XJ \\ \vdots \\ XJ^{m-1} \end{bmatrix}$$

nesingularna in velja

$$\sum_{k=0}^m A_k X J^k = 0.$$

Če definiramo še $mn \times n$ matriko

$$Y = \begin{bmatrix} X \\ \vdots \\ XJ^{m-2} \\ XJ^{m-1} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I \end{bmatrix} A_m^{-1},$$

dobimo *Jordanovo trojico* (X, J, Y) , kjer so v matriki Y leve Jordanove verige.

Splošna rešitev homogene enačbe $L\left(\frac{d}{dt}\right)u(t) = 0$ ima obliko

$$u_h(t) = Xe^{Jt}a, \quad (3.1)$$

kjer je $a \in \mathbb{C}^{mn}$ vektor poljubnih konstant. Rešitev partikularne enačbe $L\left(\frac{d}{dt}\right)u(t) = f(t)$ je

$$u_p(t) = Xe^{Jt} \int_0^t e^{-Js} Y f(s) ds$$

in splošna rešitev je

$$u(t) = Xe^{Jt} \left(a + \int_0^t e^{-Js} Y f(s) ds \right).$$

Omenimo še, da za homogeno diferencialno enačbo $L\left(\frac{d}{dt}\right)u(t) = 0$ pravimo, da je *stabilna*, če gre rešitev $u(t)$ v limiti, ko gre t proti neskončnosti, vedno proti 0. Iz (3.1) velja, da je to natanko tedaj, ko za vse lastne vrednosti λ PEP L velja $\operatorname{Re}(\lambda) < 0$, kar pomeni, da vse lastne vrednosti PEP ležijo v strogi levi polravnini kompleksne ravnine. V primeru, ko za vse lastne vrednosti velja $\operatorname{Re}(\lambda) \leq 0$, obenem pa so vse lastne vrednosti λ , za katere velja $\operatorname{Re}(\lambda) = 0$, polenostavne, ostane rešitev $u(t)$ omejena, ko gre t proti neskončnosti. V takem primeru pravimo, da je diferencialna enačba *šibko stabilna*.

Če so vse lastne vrednosti PEP L enostavne, potem za vsak λ , ki ni lastna vrednost L , velja

$$L(\lambda)^{-1} = X(\lambda I - J)^{-1} Y^H = \sum_{i=1}^{mn} \frac{x_i y_i^H}{\lambda - \lambda_i},$$

kar je posplošitev leme 2.2, ki velja za splošen nelinearni problem lastnih vrednosti.

3.3 Smithova forma

Za matrična polinoma P in Q iste dimenzije pravimo, da sta *ekvivalentna*, če je

$$P(\lambda) = E(\lambda)Q(\lambda)F(\lambda),$$

kjer sta $E(\lambda)$ in $F(\lambda)$ matriki s konstantno neničelno determinanto, katerih elementi so funkcije parametra λ . Od tod sledi, da se ničle polinoma $\det(P(\lambda))$ ujemajo z ničlami polinoma $\det(Q(\lambda))$. Če sta dodatno matriki $E(\lambda)$ in $F(\lambda)$ neodvisni od λ , pa pravimo, da sta polinoma P in Q strogo ekvivalentna.

Smithov izrek nam zagotavlja, da je vsak matrični polinom P ekvivalenten polinomu, kjer so vse matrike diagonalne. Za vsak P namreč obstajata λ -matriki $E(\lambda)$ in $F(\lambda)$ s konstantno neničelno determinanto, da je

$$P(\lambda) = E(\lambda)D(\lambda)F(\lambda),$$

kjer je $D(\lambda) = \operatorname{diag}(d_1(\lambda), \dots, d_r(\lambda), 0, \dots, 0)$ taka diagonalna $n \times n$ matrika, da so $d_1, \dots, d_r$ monični polinomi (to pomeni, da imajo vodilni koeficient 1) in da d_i deli d_{i+1} za $i = 1, \dots, r-1$. Indeks r je enak maksimalnemu rangi $L(\lambda)$ po vseh $\lambda \in \mathbb{C}$. Za tako definiramo matriko

$D(\lambda)$, ki jo imenujemo *Smithova forma*, velja, da je enolična. Polinomom $d_1, \dots, d_r$ pravimo *invariantni polinomi*. V primeru, ko ima PEP P same enostavne lastne vrednosti, je $d_1(\lambda) = \dots = d_{n-1}(\lambda) = 1$ in $d_n(\lambda) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_{mn})$. Če pa vse lastne vrednosti niso enostavne, potem lahko pišemo

$$d_i(\lambda) = (\lambda - \lambda_1)^{\alpha_{i1}} \cdots (\lambda - \lambda_k)^{\alpha_{ik}},$$

kjer so $\lambda_1, \dots, \lambda_k$ vse različne lastne vrednosti PEP P . Faktorju $(\lambda - \lambda_j)^{\alpha_{ij}}$, kjer je $j = 1, \dots, k$ in $i = 1, \dots, n$, pravimo *elementarni delitelj*. Očitno velja $\alpha_{ij} \leq \alpha_{i+1,j}$ in

$$\sum_{i=1}^n \sum_{j=1}^k \alpha_{ij} = s,$$

kjer je s stopnja polinoma $\det(P(\lambda))$. Indeksi α_{ij} se ujemajo z dolžinami Jordanovih verig.

3.4 Linearizacija

Definicija 3.4 Pravimo, da je $mn \times mn$ matrični šop $A + \lambda B$ *linearizacija matričnega polinoma* $P(\lambda)$, če je

$$\begin{bmatrix} P(\lambda) & \\ & I_{(m-1)n} \end{bmatrix} = E(\lambda)(A + \lambda B)F(\lambda)$$

in sta $E(\lambda)$ in $F(\lambda)$ matriki velikosti $mn \times mn$ s konstantno neničelno determinanto.

Če je $A + \lambda B$ linearizacija PEP P , potem se lastne vrednosti PEP P ujemajo z lastnimi vrednostmi matričnega šopa $A + \lambda B$.

Standardni linearizaciji sta prva in druga spremljevalna forma, ki imata obliko

$$\begin{aligned} C_1(\lambda) &= \lambda \begin{bmatrix} A_m & & & \\ & I & & \\ & & \ddots & \\ & & & I \end{bmatrix} + \begin{bmatrix} A_{m-1} & A_{m-2} & \cdots & A_0 \\ -I & 0 & & \\ & \ddots & \ddots & \\ & & -I & 0 \end{bmatrix} \\ C_2(\lambda) &= \lambda \begin{bmatrix} A_m & & & \\ & I & & \\ & & \ddots & \\ & & & I \end{bmatrix} + \begin{bmatrix} A_{m-1} & -I & & \\ A_{m-2} & 0 & \ddots & \\ \vdots & & \ddots & -I \\ A_0 & & & 0 \end{bmatrix}. \end{aligned}$$

Če je PEP P regularen, potem sta poljubni linearizaciji strogo ekvivalentni, to pa zato, ker sta matrična polinoma ekvivalentna natanko tedaj, ko imata enako Smithovo formo.

Prva spremljevalna forma C_1 ima to lastnost, da če je x lastni vektor PEP P za lastno vrednost λ , potem je

$$\begin{bmatrix} \lambda^{m-1}x \\ \lambda^{m-2}x \\ \vdots \\ x \end{bmatrix}$$

desni lastni vektor za C_1 in lastno vrednost λ . To pomeni, da iz desnega lastnega vektorja C_1 lahko enostavno dobimo desni lastni vektor za P . Podobno za drugo spremljevalno formo C_2 velja, da, če je y levi lastni vektor za PEP P , potem je

$$\begin{bmatrix} \bar{\lambda}^{m-1} y \\ \bar{\lambda}^{m-2} y \\ \vdots \\ y \end{bmatrix}$$

levi lastni vektor za C_2 . Prva in druga spremljevalna forma imata tako zelo uporabno lastnost, da poleg lastnih vrednosti iz linearizacije lahko preberemo tudi lastne vektorje.

Pri linearizaciji C_1 rešujemo splošen problem lastnih vrednosti, ki ga lahko zapišemo v obliki

$$C_1(\lambda)(\Lambda \otimes x) = \begin{bmatrix} P(\lambda)x \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

kjer je

$$\Lambda = \begin{bmatrix} \lambda^{m-1} \\ \lambda^{m-2} \\ \vdots \\ 1 \end{bmatrix}.$$

Ker za vsak vektor x velja $C_1(\lambda)(\Lambda \otimes x) = e_1 \otimes P(\lambda)x$, velja enakost

$$C_1(\lambda)(\Lambda \otimes I) = e_1 \otimes P(\lambda).$$

Posplošitev so linearizacije oblike $L(\lambda) = \lambda X + Y$, kjer velja

$$L(\lambda)(\Lambda \otimes x) = v \otimes P(\lambda)$$

za nek neničlen vektor $v \in \mathbb{C}^m$. Tovrstne linearizacije tvorijo prostor $\mathbb{L}_1$. Izkaže se, da v ne more biti poljuben. Pogoji za to, da je L res linearizacija P je, da nobena lastna vrednost P ni ničla polinoma

$$p(\lambda; v) = \sum_{i=1}^m v_i \lambda^{m-i},$$

kjer v primeru $v_1 = 0$ privzamemo, da je tudi ∞ ničla polinoma $p(\lambda; v)$. To pomeni, da so za izbrani matrični polinom P skoraj vsi vektorji v dobri.

Zgled 3.4 Dan je kvadratni matrični polinom $P(\lambda) = \lambda^2 A_2 + \lambda A_1 + A_0$. Matrični šop

$$L(\lambda) = \lambda X + Y = \lambda \begin{bmatrix} 2A_2 & A_2 \\ A_2 & A_1 \end{bmatrix} + \begin{bmatrix} 2A_1 - A_2 & 2A_0 \\ 0 & A_0 \end{bmatrix}$$

je linearizacija polinoma P iz prostora $\mathbb{L}_1$, če nobena lastna vrednost P ni enaka $-1/2$. Velja namreč

$$L(\lambda)(\Lambda \otimes I) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \otimes P(\lambda)$$

in $p(x; \begin{bmatrix} 2 \\ 1 \end{bmatrix}) = 2x + 1$.

Podobno so v prostoru $\mathbb{L}_2$ posplošitve linearizacije C_2 . Linearizacija $L(\lambda) = \lambda X + Y$ je iz prostora $\mathbb{L}_2$, če velja

$$(\Lambda^T \otimes I)L(\lambda) = w^T \otimes P(\lambda)$$

za nek neničelni vektor $w \in \mathbb{C}^m$.

Zgled 3.5 *Obstajajo tudi linearizacije, ki niso ne vrste $\mathbb{L}_1$ ne vrste $\mathbb{L}_2$. Tako je npr.*

$$L(\lambda) = \lambda \begin{bmatrix} 0 & A_3 & 0 \\ I & A_2 & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} -I & 0 & 0 \\ 0 & A_1 & A_0 \\ 0 & -I & 0 \end{bmatrix}$$

linearizacija kubičnega matričnega polinoma $P(\lambda) = \lambda^3 A_3 + \lambda^2 A_2 + \lambda A_1 + A_0$, a ni ne tipa $\mathbb{L}_1$ ne tipa $\mathbb{L}_2$.

Pomembne so tudi linearizacije polinoma P , ki spadajo v oba razreda. Označimo $\mathbb{DL}(P) = \mathbb{L}_1(P) \cap \mathbb{L}_2(P)$, kjer je $\mathbb{L}_1(P)$ množica vseh linearizacij polinoma P tipa $\mathbb{L}_1$ in $\mathbb{L}_2(P)$ množica vseh linearizacij polinoma P tipa $\mathbb{L}_2$. Seveda ni nujno, da je množica $\mathbb{DL}(P)$ neprazna.

3.5 Kvadratni problem lastnih vrednosti

Polinomski problem lastnih vrednosti najpogosteje nastopa v obliki kvadratnega problema (QEP), katerega splošna oblika je

$$Q(\lambda) = \lambda^2 M + \lambda C + K,$$

kjer so M, C, K dane matrike velikosti $n \times n$.

Če definiramo $Q(S) = MS^2 + CS + K$ za $S \in \mathbb{C}^{n \times n}$, potem pravimo, da je S *desna solventa*, če je $Q(S) = 0$. Podobno je *leva solventa* rešitev matrične enačbe $S^2 M + SC + K = 0$. Če je S solventa, potem velja

$$Q(\lambda) = (\lambda M + MS + C)(\lambda I - S),$$

kar pomeni, da so lastne vrednosti QEP Q unija lastnih vrednosti matrike S in lastnih vrednosti posplošenega problema lastnih vrednosti $-(MS + C)x = \lambda Mx$. Če torej na nek način pridemo do solventa, lahko potem lastne vrednosti QEP izračunamo z znanimi metodami za navaden in posplošen problem lastnih vrednosti.

Obstoj solventa za splošen QEP ni zagotovljen. Naslednja lema nam pove, kako se da solvento skonstruirati iz zadostno linearno neodvisnih lastnih vektorjev, ni pa to edina možnost za solvento.

Lema 3.5 *Če so $w_1, \dots, w_n$ linearno neodvisni lastni vektorji kvadratnega problema lastnih vrednosti Q z lastnimi vrednostmi $\lambda_1, \dots, \lambda_n$, potem je $S = W\Lambda W^{-1}$ solventa za Q , kjer je $W = [w_1, \dots, w_n]$ in $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$.*

Za linearizacijo QEP se pogosto uporabljata naslednji pridruženi obliki

$$\left(\begin{bmatrix} 0 & N \\ -K & -C \end{bmatrix} - \lambda \begin{bmatrix} N & 0 \\ 0 & M \end{bmatrix} \right) \begin{bmatrix} x \\ \lambda x \end{bmatrix} = 0 \quad (3.2)$$

$$\left(\begin{bmatrix} -K & 0 \\ 0 & N \end{bmatrix} - \lambda \begin{bmatrix} C & M \\ N & 0 \end{bmatrix} \right) \begin{bmatrix} x \\ \lambda x \end{bmatrix} = 0 \quad (3.3)$$

ki obe spadata v razred $\mathbb{L}_1$. Tu je N poljubna nesingularna matrika. Izberemo lahko npr. $N = I$, $N = \|M\|I$, $N = \|K\|I$, $N = -K$ (v (3.2), če so matrike hermitske) ali $N = M$ (v (3.3), če so matrike hermitske).

3.6 Hiperboličen in nadkritično dušen QEP

V tem razdelku bomo obravnavali poseben razred QEP, ki se v praksi velikokrat pojavi, kjer so matrike M, K, C hermitske, matrika M pa je še pozitivno definitna. Za poljuben neničelen $x \in \mathbb{C}^n$ lahko definiramo

$$m(x) = x^H M x, \quad k(x) = x^H K x, \quad c(x) = x^H C x.$$

Definicija 3.6 Dan je kvadratni problem lastnih vrednosti $Q(\lambda) = \lambda^2 M + \lambda C + K$, kjer so matrike M, K, C hermitske, matrika M pa je pozitivno definitna. Problem Q je

- hiperboličen, če za vsak $x \neq 0$ velja $c(x)^2 > 4m(x)k(x)$,
- eliptičen, če za vsak $x \neq 0$ velja $c(x)^2 < 4m(x)k(x)$,
- nadkritično dušen, če je hiperboličen in je matrika C pozitivno definitna, matrika K pa nenegativno definitna.

Če je x lastni vektor, potem vemo, da je vsaj ena izmed rešitev kvadratne enačbe $x^H Q(\lambda)x = 0$ enaka lastni vrednosti, ki pripada x . Od tod sledi, da so vse lastne vrednosti hiperboličnega QEP realne, medtem ko eliptičen QEP nima nobene realne lastne vrednosti. Podobno vidimo, da so vse lastne vrednosti nadkritično dušenega QEP negativne.

Izrek 3.7 Naj bo kvadraten problem lastnih vrednosti $Q(\lambda) = \lambda^2 M + \lambda C + K$ hiperboličen. Potem obstaja tak premik θ , da je kvadratni problem lastnih vrednosti $\tilde{Q}(\lambda) := Q(\lambda + \theta)$ nadkritično dušen.

Dokaz.

$$Q(\lambda + \theta) = \lambda^2 M + \lambda(C + 2\theta M) + K + \theta C + \theta^2 M = \lambda^2 \tilde{M} + \lambda \tilde{C} + \tilde{K}.$$

Ker je M pozitivno definitna, bosta očitno za dovolj velik θ obe matriki $\tilde{C}$ in $\tilde{K}$ pozitivno definitni. ■

Zgornji izrek nam omogoča, da nekatere rezultate dokažemo za nadkritično dušene QEP, potem pa jih s premikom prestavimo na hiperbolične QEP.

Predpostavimo, da je QEP nadkritično dušen. Potem lahko definiramo t.i. *primarni* in *sekundarni Rayleighov funkcional*

$$p_1(x) = \frac{-c(x) + d(x)}{2m(x)}, \quad p_2(x) = \frac{-c(x) - d(x)}{2m(x)},$$

kjer je $d(x) = \sqrt{c(x)^2 - 4m(x)k(x)}$. Očitno za vsak $x \neq 0$ velja $p_2(x) < p_1(x)$.

Če je x lastni vektor QEP za lastno vrednost λ , potem pravimo, da je x *primarni lastni vektor*, če je $\lambda = p_1(x)$, oziroma *sekundarni*, če je $\lambda = p_2(x)$.

Hitro lahko vidimo, da za vsak $x \neq 0$ velja

$$\begin{aligned} 2p_1(x)m(x) + c(x) &= d(x) > 0 \\ 2p_2(x)m(x) + c(x) &= -d(x) < 0, \end{aligned}$$

oziroma

$$x^H Q'(p_1(x))x > 0, \quad x^H Q'(p_2(x))x < 0,$$

od koder sledi, da so vse lastne vrednosti nadkritično dušenih (in zaradi tega tudi hiperboličnih) QEP polenostavne.

Lema 3.8 *A in B naj bosta hermitski matriki. Če obstajata vektorja z_1, z_2 , da je*

$$z_1^H A z_1 > 0, \quad z_2^H A z_2 < 0, \quad z_1^H B z_1 = z_2^H B z_2 = 0,$$

potem obstaja tak vektor $z \neq 0$, da je $z^H A z = z^H B z = 0$.

Dokaz. Definiramo $C = A + iB$. Za matriko C potem velja $z_1^H C z_1 > 0$ in $z_2^H C z_2 < 0$.

To pomeni, da numerični zaklad matrike C vsebuje vsaj eno pozitivno in eno negativno lastno vrednost. Zaradi konveksnosti mora $W(C)$ potem vsebovati tudi točko 0, torej obstaja tak vektor z , da je $z^H C z = 0$, to pa pomeni $z^H A z = z^H B z = 0$. ■

Lema 3.9 *Če je λ lastna vrednost nadkritično dušenega kvadratnega problema lastnih vrednosti, potem so pripadajoči lastni vektorji bodisi vsi primarni bodisi vsi sekundarni.*

Dokaz. Denimo, da sta x in y taka lastna vektorja, da je $p_1(x) = \lambda = p_2(y)$. Za poljuben σ je potem $\xi(\sigma) = (1 - \sigma)x + \sigma y$ tudi lastni vektor za lastno vrednost λ , torej velja $m(\xi(\sigma))\lambda^2 + c(\xi(\sigma))\lambda + k(\xi(\sigma)) = 0$. Funkcija

$$f(\sigma) := 2m(\xi(\sigma))\lambda + c(\xi(\sigma)) = \pm d(\xi(\sigma))$$

je zvezna in ima pri $\sigma = 0$ negativno, pri $\sigma = 1$ pa pozitivno vrednost. Zaradi tega mora obstajati tak σ , da je $f(\sigma) = \pm d(\xi(\sigma)) = 0$. To pa ni možno, saj je zaradi nadkritične dušenosti $d(x) > 0$ za vsak $x \neq 0$ in prišli smo do protislovja. ■

Tako lahko vsako lastno vrednost nadkritično dušenega QEP proglasimo za primarno ali za sekundarno. Za QEP Q definiramo zaklad vrednosti $W_i(Q) = \{p_i(x) : x \neq 0\}$ za $i = 1, 2$. Če je $\alpha < \beta$ in $\alpha, \beta \in W_i(Q)$, potem je hitro očitno, da velja $[\alpha, \beta] \subset W_i(Q)$. Tako lahko ugotovimo, da obstajata para $\alpha_1 < \beta_1$ in $\alpha_2 < \beta_2$, da je $W_1(Q) = [\alpha_1, \beta_1]$ in $W_2(Q) = [\alpha_2, \beta_2]$.

Lema 3.10 *Naj bo Q nadkritično dušen kvadratni problem lastnih vrednosti. Potem za zaklada vrednosti primarnega in sekundarnega Rayleighovega funkcionala velja $W_1(Q) \cap W_2(Q) = \emptyset$.*

Dokaz. Denimo, da obstaja $\alpha \in W_1(Q) \cap W_2(Q)$. Torej obstajata vektorja x in y , da velja

$$x^H Q(\alpha)x = 0 \quad y^H Q(\alpha)y = 0 \tag{3.4}$$

$$x^H Q'(\alpha)x > 0 \quad y^H Q'(\alpha)y < 0. \tag{3.5}$$

Po lemi 3.8 potem obstaja tak vektor z , da je $z^H Q(\alpha)z = 0$ in $z^H Q'(\alpha)z = 0$. To pa je protislovje, saj je potem α večkratna ničla primarnega ali sekundarnega Rayleighovega funkcionala, kar zaradi nadkritične dušenosti ni možno. ■

Izrek 3.11 (Markus [11]) Naj bo $Q(\lambda) = \lambda^2 M + \lambda C + K$ hermitski kvadratni problem lastnih vrednosti in M pozitivno definitna. Potem je Q hiperboličen natanko tedaj, ko obstaja tak μ , da je matrika $Q(\mu)$ negativno definitna.

Dokaz. Denimo, da je QEP Q hiperboličen. Po izreku 3.7 obstaja tak premik θ , da je kvadratni problem lastnih vrednosti $\tilde{Q}(\lambda) := Q(\lambda + \theta)$ nadkritično dušen. Potem za zaklada vrednosti velja $W_1(\tilde{Q}) = [\alpha_1, \beta_1]$, $W_2(\tilde{Q}) = [\alpha_2, \beta_2]$, kjer je $\beta_2 < \alpha_1$. Za poljuben $\beta_2 < \gamma < \alpha_1$ je matrika $\tilde{Q}(\gamma)$ negativno definitna. Velja namreč, da za vsak $x \neq 0$ velja $x^H \tilde{Q}(\gamma)x \neq 0$, predznak pa mora biti za vse vektorje enak. Ker je za $|\eta| \gg 0$ matrika $\tilde{Q}(\eta)$ očitno pozitivno definitna in ima kvadratna enačba $x^H \tilde{Q}(\alpha)x = 0$ dve realni rešitvi, eno na $W_1(\tilde{Q})$ in drugo na $W_2(\tilde{Q})$, je $\tilde{Q}(\gamma)$ negativno definitna, to pa pomeni, da je $Q(\gamma + \theta)$ negativno definitna in lahko vzamemo $\mu = \gamma + \theta$.

Predpostavimo, da je matrika $Q(\mu)$ negativno definitna. Ker je za $|\eta| \gg 0$ matrika $Q(\eta)$ pozitivno definitna, ima za vsak $x \neq 0$ enačba $x^H Q(\alpha)x = 0$ dve realni rešitvi. Torej za vsak $x \neq 0$ velja $c(x)^2 > 4m(x)k(x)$, to pa je ravno definicija hiperboličnosti. ■

Posledica 3.12 Naj bo $Q(\lambda) = \lambda^2 M + \lambda C + K$ hiperbolični kvadratni problem lastnih vrednosti. Potem ima Q n primarnih in n sekundarnih lastnih vrednosti.

Dokaz. Ker je Q hiperboličen, obstaja tak μ , da je $Q(\mu)$ negativno definitna matrika. Vemo, da sta za dovolj velik $M > 0$ matriki $Q(-M)$ in $Q(M)$ pozitivno definitni. Ker je $Q(\lambda)$ za vsak $\lambda \in \mathbb{R}$ hermitska matrika, so lastne vrednosti $Q(\lambda)$ analitične funkcije λ . Od tod sledi, da ima Q n primarnih lastnih vrednosti na intervalu (μ, M) in n sekundarnih lastnih vrednosti na intervalu $(-M, \mu)$. ■

Lastne vrednosti lahko uredimo po velikosti in jih oštevilčimo tako, da velja $\lambda_{2n} \leq \dots \leq \lambda_{n+1} < \lambda_n \leq \dots \leq \lambda_1$, pri čemer so $\lambda_1, \dots, \lambda_n$ primarne, $\lambda_{n+1}, \dots, \lambda_{2n}$ pa sekundarne lastne vrednosti.

Lema 3.13 Naj bo Q hiperbolični kvadratni problem lastnih vrednosti. Potem so lastni vektorji, ki pripadajo n primarnim lastnim vrednostim, linearno neodvisni. Enako velja za lastne vektorje sekundarnih lastnih vrednosti.

Dokaz. Predpostavimo lahko, da je problem tak, da so vse lastne vrednosti enostavne. Sedaj uporabimo indukcijo. Naj bodo lastni vektorji $x_1, \dots, x_{k-1}$ za $\lambda_1, \dots, \lambda_{k-1}$ linearno neodvisni, sedaj pa jim dodamo še lastni vektor x_k za λ_k , kjer je $\lambda_k < \lambda_{k-1}$. Denimo, da je x_k linearna kombinacija $x_1, \dots, x_{k-1}$. Potem bi moralo veljati $\lambda_{k-1} \leq p_1(x_k) \leq \lambda_1$, kar je protislovje. ■

Za lastne vrednosti hiperboličnega kvadratnega problema velja naslednji izrek o minimaksu, ki je posplošitev znanega Courant-Fischerjevega izreka o minimaksu za simetrične matrike.

Izrek 3.14 (Duffin, [5]) Za $i = 1, \dots, n$ velja

$$\lambda_i = \max_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=i}} \min_{\substack{x \in S \\ x \neq 0}} p_1(x) = \min_{\substack{R \subset \mathbb{R}^n \\ \dim(R)=n-i+1}} \max_{\substack{x \in R \\ x \neq 0}} p_1(x),$$

$$\lambda_{n+i} = \max_{\substack{S \subset \mathbb{R}^n \\ \dim(S)=i}} \min_{\substack{x \in S \\ x \neq 0}} p_2(x) = \min_{\substack{R \subset \mathbb{R}^n \\ \dim(R)=n-i+1}} \max_{\substack{x \in R \\ x \neq 0}} p_2(x).$$

3.7 Skaliranje in izbira optimalne linearizacije

PEP in QEP lahko numerično rešujemo na več načinov. Eden izmed standardnih načinov je linearizacija, kjer problem prevedemo na GEP. za reševanje GEP potem uporabimo npr. QZ algoritem za polne matrike oz. Arnoldijevo metodo za razpršene matrike.

Naj bo $L(\lambda) = \lambda X + Y$ linearizacija za PEP $P(\lambda) = \sum_{i=0}^m \lambda^i A_i$, kjer so A_i matrike $n \times n$ in $A_m \neq 0$. Torej obstajata matriki $E(\lambda)$ in $F(\lambda)$ s konstantno neničelno determinanto, da je

$$E(\lambda)L(\lambda)F(\lambda) = \begin{bmatrix} P(\lambda) \\ I_{(m-1)n} \end{bmatrix}.$$

Omenili smo že linearizacije iz razredov $\mathbb{L}_1$ in $\mathbb{L}_2$ in njun presek $\mathbb{DL}$. V primeru QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$ lahko pokažemo, da velja

$$\mathbb{DL}(Q) = \left\{ L(\lambda) = \begin{bmatrix} v_1 M & v_2 M \\ v_2 M & v_2 C - v_1 K \end{bmatrix} + \begin{bmatrix} v_1 C - v_2 M & v_1 K \\ v_1 K & v_2 K \end{bmatrix} : v \in \mathbb{C}^2 \right\}.$$

Radi bi, da se občutljivosti lastnih vrednosti z linearizacijo ne povečajo. Vprašanje je tudi, ali obstaja optimalni v in kako iz lastnega vektorja za L izločimo lastni vektor za Q .

Posebna primera, ki ju pogosto uporabljamo, sta linearizaciji, ki ju dobimo z izbiro $v = e_1$ oziroma $v = e_2$. Ustrezni linearizaciji, ki ju bomo označili z L_1 in L_2 , sta

$$\begin{aligned} L_1(\lambda) &= \lambda \begin{bmatrix} M & 0 \\ 0 & -K \end{bmatrix} + \begin{bmatrix} C & K \\ K & 0 \end{bmatrix} \\ L_2(\lambda) &= \lambda \begin{bmatrix} 0 & M \\ M & C \end{bmatrix} + \begin{bmatrix} -M & 0 \\ 0 & K \end{bmatrix}. \end{aligned}$$

Linearizaciji L_1 in L_2 ohranjata simetričnost, če so matrike M, C, K simetrične. To pa ne velja za standardno linearizacijo C_1 in razreda $\mathbb{L}_1$, ki je podana z

$$C_1(\lambda) = \lambda \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} C & K \\ -I & 0 \end{bmatrix}.$$

Zanimalo nas bo, kdaj se občutljivost lastnih vrednosti in vektorjev ne poveča bistveno, če QEP lineariziramo z eno izmed linearizacij L_1, L_2 ali C_1 . Prav tako nas v primeru L_1 in L_2 zanima, koliko sta linearizaciji slabši od optimalne linearizacije iz množice $\mathbb{DL}(Q)$.

Naj bo λ enostavna lastna vrednost Q z desnim lastnim vektorjem x in levim lastnim vektorjem y . Občutljivost lastne vrednosti je definirana kot

$$K_Q(\lambda) = \limsup_{\epsilon \rightarrow 0} \left\{ \frac{|\Delta\lambda|}{\epsilon\lambda} : (Q(\lambda + \Delta\lambda) + \Delta Q(\lambda + \Delta\lambda))(x + \Delta x) = 0, \right. \\ \left. \|\Delta M\|_2 \leq \epsilon m, \|\Delta K\|_2 \leq \epsilon k, \|\Delta C\|_2 \leq \epsilon c \right\}, \quad (3.6)$$

kjer je $m = \|M\|_2, c = \|C\|_2$ in $k = \|K\|_2$.

Izkaže se, da je

$$K_Q(\lambda) = \frac{(|\lambda|^2 m + |\lambda| d + k) \|y\|_2 \|x\|_2}{|\lambda| \|y^H (2\lambda M + C)x\|}.$$

Podobno bi lahko zapisali občutljivost lastne vrednosti lineariziranega problema. Za $L(\lambda) = \lambda X + Y$ dobimo

$$K_L(\lambda) = \frac{(|\lambda| \|X\|_2 + \|Y\|_2) \|w\|_2 \|z\|_2}{|\lambda| |w^H X z|}.$$

Sedaj lahko definiramo faktor rasti $\Phi_L(\lambda) = K_L(\lambda)/K_Q(\lambda)$.

Če definiramo

$$\rho = \frac{\max(m, c, k)}{\min(m, k)},$$

kjer je $m = \|M\|_2$, $c = \|C\|_2$ in $k = \|K\|_2$, potem se izkaže, da je problem dobro skaliran kadar je $\rho \approx 1$.

Izrek 3.15 (Higham, Mackey, Tisseur [7]) Če je λ enostavna lastna vrednost kvadratnega problema lastnih vrednosti Q , potem velja:

- a) $K_L(\lambda, e_1) \leq 2\sqrt{2}\rho \inf_v (K_L(\lambda, v))$, če je M nesingularna in $|\lambda| \geq 1$,
- b) $K_L(\lambda, e_2) \leq 2\sqrt{2}\rho \inf_v (K_L(\lambda, v))$, če je K nesingularna in $|\lambda| \leq 1$.

Iz zgornjega izreka sledi, da se v primeru, ko je $\rho \approx 1$, občutljivosti lastne vrednosti pri linearizaciji L_1 ali L_2 ne razlikuje veliko od občutljivosti pri optimalni linearizaciji. Če pa gledamo faktor rasti, potem se izkaže, da če uporabimo eno izmed linearizacij L_1, L_2, C_1 , potem je $\Phi_L(\lambda) \approx 1$, če velja:

- a) C_1 : $m \approx k \approx c$,
- b) L_1 : $|\lambda| \geq 1$ in $\rho \approx 1$,
- c) L_2 : $|\lambda| \leq 1$ in $\rho \approx 1$.

Kadar je $\rho \not\approx 1$, je priporočljivo uporabiti predhodno skaliranje QEP, kot so ga predlagali Fan, Lin in van Dooren v [6]. Če v $Q(\lambda)$ vstavimo $\lambda = \mu\gamma$, dobimo

$$\begin{aligned} Q(\lambda) &= \lambda^2 M + \lambda C + K \\ &= \mu^2 (\gamma^2 M) + \mu (\gamma C) + K \\ &= \mu^2 \tilde{M} + \mu \tilde{C} + K = \tilde{Q}(\mu). \end{aligned}$$

Za γ izberemo tisto vrednost, ki minimizira

$$\max \left(\|\tilde{M}\| / \|\tilde{C}\|, \|\tilde{K}\| / \|\tilde{C}\| \right) = \max (\gamma m / c, k / (\gamma c)),$$

kar je $\gamma = \sqrt{k/m}$. Za skalirani QEP $\tilde{Q}$ potem velja, da je $\tilde{\rho} = \max(1, c/\sqrt{mk})$.

Izrek 3.16 (Higham, Mackey, Tisseur [7]) Naj bo λ enostavna lastna vrednost kvadratnega problema lastnih vrednosti $Q(\lambda) = \lambda^2 M + \lambda C + K$, kjer je $\gamma = \sqrt{k/m}$. Naj bo ali

- a) $c \leq \max(k, m)$, $m \approx k$, $P = Q$ in $L \in \mathbb{DL}(Q)$, ali

$$b) \ c \leq \sqrt{(k, m)}, P = \tilde{Q} \text{ in } L \in \mathbb{DL}(\tilde{Q}).$$

Potem velja:

$$a) \text{ če je } K \text{ nesingularna in } |\lambda| \geq 1, \text{ je } K_L(\lambda, e_1) \approx K_P(\lambda),$$

$$b) \text{ če je } M \text{ nesingularna in } |\lambda| \leq 1, \text{ je } K_L(\lambda, e_2) \approx K_P(\lambda).$$

Kaj pa lastni vektor? Vemo, da je v primeru, ko je x lastni vektor za Q za lastno vrednost λ , $z = \begin{bmatrix} \lambda x \\ x \end{bmatrix}$ lastni vektor za L . Pri numeričnem računanju dobimo približek za lastni vektor za L , ki ima bločno obliko $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$. Ali je bolje izračunati približek za x iz z_1 ali iz z_2 ? Izkaže se, da je v primeru, ko je $|\lambda| \geq 1$ bolje izbrati z_1 , če je $|\lambda| \leq 1$ pa z_2 .

3.8 Numerične metode za polne QEP

Poleg vseh metod, ki smo jih obravnavali za splošne NEP, imamo za QEP (in PEP) še nekaj posebnih metod. Prva možnost je linearizacija, ki jo potem ponavadi kombiniramo s QZ algoritmom. Na kratko je postopek predstavljen v algoritmu 3.1.

Če je $A - \lambda B$ linearizacija QEP Q , potem s QZ algoritmom dobimo unitarni matriki Q, Z in zgornji trikotni matriki S, T , da je

$$Q^H A Z = S, \quad Q^H B Z = T.$$

Potem so lastne vrednosti QEP enake $\lambda_i = s_{ii}/t_{ii}$ za $i = 1, \dots, 2n$, kjer v primeru $t_{ii} = 0$ vzamemo $\lambda_i = \infty$.

Algoritem 3.1 Linearizacija in QZ algoritem

zgeneriraj matriki A in B , da je $A - \lambda B$ linearizacija $Q(\lambda) = \lambda^2 M + \lambda C + K$

izračunaj unitarni matriki Q, Z in zgornji trikotni S, T , da je $Q^H A Z = S$ in $Q^H B Z = T$

$k = 1, \dots, 2n$

$$\lambda_k = s_{kk}/t_{kk}$$

reši $(S - \lambda_k T)\phi = 0$ za $\phi \neq 0$

$$\xi = Z\phi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

$$r_1 = Q(\lambda_k)\xi_1/\|\xi_1\|, \quad r_2 = Q(\lambda_k)\xi_2/\|\xi_2\|$$

če je $\|r_1\| \leq \|r_2\|$ vzemi $x_k = \xi_1$, sicer $x_k = \xi_2$.

Čeprav vemo, da je QZ algoritem obratno stabilen za GEP, pa to še ne pomeni, da je algoritem 3.1 obratno stabilen postopek za reševanje QEP. Težava je v tem, da se pri motnji ne upošteva struktura problema. Obratna napaka za izračunani lastni par (λ, ξ) za GEP je

$$\eta_{\text{GEP}}(\lambda, \xi) = \frac{\|(A - \lambda B)\xi\|}{((\|A\| + |\lambda|\|B\|)\|\xi\|)}.$$

Ker je QZ algoritem obratno stabilen, lahko pričakujemo, da bo $\eta_{\text{GEP}}(\lambda, \xi) = \mathcal{O}(u)$, kjer je u osnovna zaokrožitvena napaka. Tega pa ne moremo pokazati za obratno napako para (λ, x) za QEP, kjer je

$$\eta_{\text{QEP}}(\lambda, x) = \frac{\|Q(\lambda)x\|}{((|\lambda|^2\|M\| + |\lambda|\|C\| + \|K\|)\|x\|)}.$$

Tako se lahko zgodi, da sta tako $\eta_{\text{QEP}}(\lambda, \xi_1)$ kot $\eta_{\text{QEP}}(\lambda, \xi_2)$ veliko večja od $\eta_{\text{GEP}}(\lambda, \xi)$.

3.8.1 Simetrične in definitne linearizacije

Dan je QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$. Če so matrike M, C, K simetrične, potem lahko QEP lineariziramo kot GEP $A - \lambda B$, kjer sta matriki A in B simetrični.

Sama simetrija nam pri QZ algoritmu, ki ga uporabimo v algoritmu 3.1, ne pomaga, saj je QZ algoritem ne zna upoštevati. Če pa je matrika B dodatno še definitna, potem lahko GEP $A - \lambda B$ preko razcepa Choleskega $B = LL^T$ prevedemo na simetrični problem lastnih vrednosti $L^{-1}AL^{-T}y = \lambda y$ za $y = L^T\xi$. V nadaljevanju lahko uporabimo QR algoritem za simetrične matrike ali poljuben drug algoritem za simetrični problem lastnih vrednosti.

Lahko se zgodi, da je par (A, B) simetričen in definiten, kar pomeni da je $\gamma(A, B) > 0$, kjer je γ t.i. Crawfordovo število

$$\gamma(A, B) = \min_{\|z\|=1} \sqrt{(z^H A z)^2 + (z^H B z)^2},$$

a nobena izmed matrik A, B ni definitna. To se npr. zgodi pri linearizaciji nadkritično dušene QEP. Na simetričnem definitnem paru (A, B) lahko potem uporabimo posplošitev Jacobi-jeve metode.

Za nedefinitne simetrične pare (A, B) sicer obstajajo algoritmi, ki znajo izkoristiti simetrijo, a zaenkrat noben izmed njih ni numerično stabilen.

3.8.2 Tridiagonalne matrike

Če so matrike M, C, K v QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$ tridiagonalne, potem lahko vrednost $p(\lambda) = \det(Q(\lambda))$ izračunamo v $\mathcal{O}(n)$, prav tako $p'(\lambda)$ in $p''(\lambda)$. Uporabimo lahko tričlensko rekurzivno formulo, Bohtetov LU+MV algoritem ali QR razcep (podobno kot pri algoritmu Kublanovskaje). To lahko uporabimo za učinkovito računanje lastnih vrednosti, več podrobnosti lahko najdete v [15].

Če je problem dodatno še hiperboličen, potem lahko z metodo deli in vladaj, ki jo kombiniramo z Laguerrovo metodo, učinkovito izračunamo vse lastne vrednosti.

Za splošne tridiagonalne (oz. pasovne) QEP lahko še vedno uporabimo metodo deli in vladaj, ki jo kombiniramo z Ehrlich-Aberthovo ali Durand-Kernerjevo metodo za iskanje vseh ničel polinoma p .

Na žalost se splošnega QEP ne da pretvoriti v tridiagonalnega, saj za splošne matrike M, C, K ne obstajata nesingularni matriki U in V , da bi bile matrike UMV, UCV, UKV vse tridiagonalne.

3.8.3 Žiroskopski problemi

Žiroskopski QEP ima obliko $G(\lambda) = \lambda^2 M + \lambda C + K$, kjer sta matriki M in K hermitski, M je pozitivno definitna, matrika C pa je poševno hermitska, kar pomeni $C = -C^H$.

Iz $G(\lambda)^H = G(-\bar{\lambda})$ sledi, da so lastne vrednosti G razporejene simetrično glede na imaginarno os. Lastne vrednosti, ki niso strogo imaginarne, tako nastopajo v parih $(\lambda, -\bar{\lambda})$.

Če so matrike M, C, K realne, potem velja tudi $G(\lambda)^T = G(-\lambda)$ in v tem primeru imajo lastne vrednosti t.i. hamiltonsko strukturo. Lastne vrednosti, ki niso strogo realne ali imaginarne, tako nastopajo v četvorkah $(\lambda, -\lambda, \bar{\lambda}, -\bar{\lambda})$, če pa je λ strogo realna ali strogo imaginarna, potem je lastna vrednost tudi $-\lambda$.

Od tod sledi, da je potreben pogoj, da je žiroskopski sistem

$$M\ddot{q}(t) + C\dot{q}(t) + Kq(t) = 0 \quad (3.7)$$

(šibko) stabilen ta, da vse lastne vrednosti QEP G ležijo na imaginarni osi.

Če definiramo QEP $Q(\lambda) = -G(-i\lambda) = \lambda^2 M + \lambda(iC) - K$, potem lahko hitro preverimo, da je Q hermitski QEP, saj je M hermitska pozitivno definitna, $K^H = K$ in $(iC)^H = iC$.

- Če je matrika K pozitivno definitna, potem je Q hiperboličen QEP. Od tod sledi, da so lastne vrednosti G čisto imaginarne in polenostavne, kar pomeni, da je žiroskopski sistem (3.7) šibko stabilen.
- Če je K negativno definitna in C realna, potem zaradi $C^T = -C$ obstaja tak neničelni vektor x_0 , da je $x_0^H C x_0 = 0$. Potem je

$$d(x_0) = i^2(x_0^H C x_0)^2 + 4(x_0^H M x_0)(x_0^H K x_0) < 0,$$

to pa pomeni, da Q ni hiperboličen problem. Torej lastne vrednosti G niso nujno čisto imaginarne in polenostavne in stabilnost G ni zagotovljena.

Ker se pri žiroskopskem problemu lastne vrednosti pojavljajo v četvorkah oziroma parih, bi bilo dobro, da bi se ta lastnost prenesla v strukturi tudi na linearizacijo. S tem bi dobili možnost razvoja natančnejših in učinkovitejših algoritmov za žiroskopske probleme. Linearizacija, ki ustreza tem zahtevam je taka, da je ena izmed matrik hamiltonska, druga pa poševno hamiltonska.

Definicija 3.17 Za $2n \times 2n$ matriko A pravimo, da je hamiltonska, če je $(AJ)^T = AJ$ za $J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$, oziroma poševno hamiltonska, če je $(AJ)^T = -AJ$.

Iz definicije sledi, da je matrika A v bločni 2×2 obliki hamiltonska, če ima obliko

$$A = \begin{bmatrix} C & D \\ E & -C^H \end{bmatrix},$$

kjer sta matriki D in E hermitski. Podobno je matrika B v bločni 2×2 obliki poševno hamiltonska, če ima obliko

$$A = \begin{bmatrix} C & D \\ E & C^H \end{bmatrix},$$

kjer sta matriki D in E poševno hermitski.

Ustrezni linearizaciji žiroskopskega QEP G sta

$$\begin{bmatrix} K & 0 \\ C & K \end{bmatrix} - \lambda \begin{bmatrix} 0 & K \\ -M & 0 \end{bmatrix},$$

ki je primerna, če je M singularna, in

$$\begin{bmatrix} 0 & -K \\ M & 0 \end{bmatrix} - \lambda \begin{bmatrix} M & C \\ 0 & M \end{bmatrix},$$

ki je primerna, če je K singularna. Kadar sta matriki M in K obe nesusingularni, sta dobri obe linearizaciji.

Ko enkrat lineariziramo žiroskopski QEP z GEP hamiltonsko-poševno hamiltonske oblike, lahko v nadaljevanju uporabimo posebne algoritme, ki upoštevajo strukturo problema in učinkovito in natančno izračunajo lastne vrednosti QEP.

Tu pridejo v poštev t.i. simplektične metode, ki uporabljajo simplektične unitarne transformacije.

Definicija 3.18 Matrika S je simplektična, če je $SJ = S^H J$.

Lema 3.19 Naj bo H hamiltonska matrika, S pa nesusingularna simplektična matrika. Potem je $S^{-1}HS$ tudi hamiltonska matrika.

Simplektične unitarne matrike so:

a) Householderjeva simplektična zrcaljenja

$$H(k, u) = \begin{bmatrix} P & 0 \\ O & P \end{bmatrix},$$

kjer je

$$P = I - \frac{2}{u^H u} u u^H, \quad u = [0 \quad \cdots \quad 0 \quad u_k \quad \cdots \quad u_n]^T.$$

b) Givensove simplektične rotacije

$$G(k, c, s) = \begin{bmatrix} C & S \\ -S & C \end{bmatrix},$$

kjer je $C = \text{diag}(1, \dots, 1, c, 1, \dots, 1)$, $S = \text{diag}(0, \dots, 0, s, 0, \dots, 0)$, $|c|^2 + |s|^2 = 1$ in $\bar{c}s \in \mathbb{R}$.

Obstajajo posplošitve QZ algoritma za hamiltonski GEP, ki uporabljajo simplektične transformacije, prav tako obstajajo posplošitve Jacobijeve oz. Eberleinove metode.

3.8.4 Palindromski polinomski problemi lastnih vrednosti

Pri palindromskem polinomskem problemu sta polinom $P(\lambda) = A_0 + \lambda A_1 + \cdots + \lambda^m A_m$ in pripadajoči vzvratni polinom $P_R(\lambda) = \lambda^m A_0 + \lambda^{m-1} A_1 + \cdots + A_0$ povezana na poseben način. Za polinom P pravimo, da je

- *palindromski*, če velja $P(\lambda) = P_R(\lambda)$,
- **-palindromski*, če velja $P(\lambda) = P_R(\lambda)^*$, kjer je $*$ = T ali $*$ = H ,
- *anti-palindromski*, če velja $P(\lambda) = -P_R(\lambda)$,
- **-anti-palindromski*, če velja $P(\lambda) = -P_R(\lambda)^*$, kjer je $*$ = T ali $*$ = H .

Poleg zgoraj navedenih so še druge lastnosti, ki so pomembne za računanje lastnih vrednosti. Tako za polinom P pravimo, da je:

- *sod*, če velja $P(-\lambda) = P(\lambda)$,
- *lih*, če velja $P(-\lambda) = -P(\lambda)$,
- **-sod*, če velja $P(-\lambda)^* = P(\lambda)$, kjer je $*$ = T ali $*$ = H ,
- **-lih*, če velja $P(-\lambda)^* = -P(\lambda)$, kjer je $*$ = T ali $*$ = H .

Če ima polinom P katero izmed zgornjih lastnosti, potem lastne vrednosti nastopajo v parih:

- $(\lambda, 1/\lambda)$, če je P palindromski, anti-palindromski, T-palindromski ali T-anti-palindromski,
- $(\lambda, 1/\bar{\lambda})$, če je P *-palindromski ali *-anti-palindromski,
- $(\lambda, -\lambda)$, če je P sod, lih, T-sod ali T-lih,
- $(\lambda, -\bar{\lambda})$, če je P *-sod ali *-lih.

Za polinome, ki imajo zgornje lastnosti, potrebujemo linearizacijo, ki bi ohranila strukturo, nato pa v nadaljevanju še prilagojene algoritme, ki znajo to izkoristiti za ekonomičnejše ali natančnejše računanje.

Zgled 3.6 Polinom $P(\lambda) = \lambda^2 A_0^T + \lambda A_1 + A_0$, kjer je $A_1 = A_1^T$ in $A_0 \neq 0$, je T-palindromski. Lineariziramo ga lahko z naslednjo linearizacijo iz $\mathbb{L}_1$:

$$L(\lambda) = \lambda Z + Z^T = v_1 \left(\lambda \begin{bmatrix} A_0^T & A_1 - A_0 \\ A_0^T A_0^T \end{bmatrix} + \begin{bmatrix} A_0 & A_0 \\ A_1 - A_0^T & A_0 \end{bmatrix} \right).$$

To je linearizacija, če P nima lastne vrednosti -1 , saj je -1 ničla $p(x; v) = v_1 x + v_1$. Vidimo, da je linearizacija prav tako T-palindromska.

Zgled 3.7 Polinom $P(\lambda) = \lambda^3 A^T + \lambda^2 B^T + \lambda B + A$ je prav tako T -palindromski. V $\mathbb{L}_1$ ga lahko lineariziramo z $v = [1 \quad -1 \quad 1]^T$, kar nam da linearizacijo $L(\lambda) = \lambda Z + Z^T$, kjer je

$$\begin{bmatrix} A^T & B^T - A & B + A \\ -A^T & B^T - B + A^T & B^T - A \\ A^T & -A^T & A^T \end{bmatrix}.$$

To je linearizacija v primeru, ko nobena lastna vrednost P ni ničla polinoma $p(x; v) = x^2 - x + 1$, ničli pa sta $\frac{1}{2}(1 \pm i\sqrt{3})$.

Poglavje 4

Metode podprostorov za kvadratni problem lastnih vrednosti

Dan je kvadratni problem lastnih vrednosti $Q(\lambda) = \lambda^2 M + \lambda C + K$, kjer so matrike M, C in K velike in razpršene. V tem primeru ne moremo izračunati vseh lastnih vrednosti z direktnimi metodami, lahko pa uporabimo metode podprostorov in izračunamo nekaj lastnih vrednosti, ki so najbližje izbrani tarči. Za začetek bomo najprej na kratko pogledali, kako uporabljamo metode podprostorov na navadnem in posplošenem problemu lastnih vrednosti, potem pa bomo to posplošili še na kvadratni (polinomski) problem lastnih vrednosti.

4.1 Metode podprostorov

Radi bi izračunali nekaj lastnih vrednosti matrike A , ki je velika in razpršena. Ne moremo si privoščiti, da bi izračunali vse lastne vrednosti preko npr. QR razcepa, saj bi na ta način porabili preveč časa in potrebovali preveč prostora.

Tu pridejo v poštev iterativne metode podprostorov, kjer namesto matrike A potrebujemo le podprogram, ki zna za poljubni vektor x izračunati produkt Ax (in po potrebi podoben podprogram še za množenje z A^T).

Najpreprostejša metoda tega tipa je znana potenčna metoda, s katero lahko izračunamo dominantni lastni par, vse kar potrebujemo pa je, da znamo v vsakem koraku vektor množiti z matriko A . Če na začetku izberemo vektor u_1 , potem pri potenčni metodi generiramo zaporedje vektorjev $Au_1, A^2u_1, \dots$, za katero se da pokazati, da po smeri skonvergira proti dominantnemu lastnemu vektorju matrike A .

4.1.1 Podprostor Krilova

Za matriko A in vektor u_1 definiramo *podprostor Krilova* kot

$$\mathcal{K}_k(A, u_1) = \text{Lin}(u_1, Au_1, \dots, A^{k-1}u_1).$$

Če začetni vektor u_1 ni vsebovan v invariantnem podprostoru dimenzije manjše od k , potem je

$\dim(\mathcal{K}_k(A, u_1)) = k$. Zapišemo lahko tudi

$$\mathcal{K}_k(A, u_1) = \{p(A)u_1 : p \in \mathcal{P}_{k-1}\},$$

kjer je $\mathcal{P}_{k-1}$ prostor vseh polinomov stopnje manjše ali enake $k - 1$.

Iz potenčne metode vemo, da vektorji $A^j u_1$ konvergirajo proti dominantnemu lastnemu vektorju, a v podprostoru $\mathcal{K}_k$, ki vsebuje še prejšnje vektorje iz potenčne metode, lahko vedno dobimo še boljšo aproksimacijo za dominanten lastni vektor.

4.1.2 Arnoldijev algoritem

Baza $\{u_1, Au_1, \dots, A^{k-1}u_1\}$ za podprostor Krilova $\mathcal{K}_k(A, u_1)$ je zelo slabo pogojena in zato neprimerna za praktično uporabo. Stabilnejšo ortonormirano bazo dobimo s pomočjo Gram-Schmidtove ortogonalizacije. Seveda moramo uporabiti tako imenovano modificirano Gram-Schmidtovo metodo (MGS), ki je stabilnejša od klasične. Denimo, da stolpci matrike $U_j = [u_1 \cdots u_j]$ sestavljajo ortonormirano bazo za $\mathcal{K}_j$. Potem je

$$\mathcal{K}_{j+1} = \text{Lin}(u_1, \dots, u_j, Au_j)$$

in naslednji bazni vektor dobimo iz

$$\tilde{u}_{j+1} = (I - U_j U_j^H) Au_j, \quad u_{j+1} = \frac{1}{\|\tilde{u}_{j+1}\|_2} \tilde{u}_{j+1}.$$

Ustrezni algoritem se imenuje *Arnoldijev algoritem* in je prikazan v algoritmu 4.1. Arnoldijev algoritem se konča pri izbranem k ali pa, ko je $h_{j+1,j} = 0$.

Algoritem 4.1 Arnoldijev algoritem. Začetni podatki so matrika A , normiran začetni vektor u_1 in dimenzija k . Algoritem vrne ortonormirane stolpce $u_1, \dots, u_k$, ki tvorijo ortonormirano bazo za podprostor Krilova $\mathcal{K}_k(A, u_1)$.

```

j = 1, 2, ..., k
  z = Au_j
  h_j = U_j^H z_j
   $\tilde{u}_{j+1} = z_j - U_j h_j$ 
   $h_{j+1,j} = \|\tilde{u}_{j+1}\|_2$ 
  če je  $h_{j+1,j} = 0$ , potem prekini računanje
   $u_{j+1} = \tilde{u}_{j+1} / h_{j+1,j}$ 

```

Trditev 4.1 Arnoldijev algoritem se lahko izvaja do $j = k$, kjer je $k = \dim \mathcal{K}_n(A, u_1)$. Za $j = 1, \dots, k$ velja

$$AU_j = U_j H_j + h_{j+1,j} u_{j+1} e_j^T,$$

kjer je H_j zgornja Hessenbergova matrika velikosti $j \times j$, stolpci matrike $U_j = [u_1 \cdots u_j]$ pa so ortonormirana baza za $\mathcal{K}_j(A, u_1)$.

Pišemo lahko tudi

$$AU_k = U_{k+1} \tilde{H}_k,$$

kjer $\tilde{H}_k$ dobimo tako, da H_k dodamo še vrstico $[0 \cdots 0 \ h_{k+1,k}]$.

Izkaže se, da moramo za numerično stabilno ortogonalizacijo modificirano Gram-Schmidtovo metodo pogosti dvakrat zapored. Dovolj pa je že, če ortogonalizacijo ponovimo samo na določenih vektorjih. V poštev pridejo vsi vektorji, katerih norma se močno zmanjša, ko od njih odštejemo komponente že ortonormiranih vektorjev. To je namreč znak, da je mogoče prišlo do odštevanja približno enako velikih količin. Dobljeni postopek se imenuje RGS (repeated Gram-Schmidt), celotna različica Arnoldijeve metode z RGS pa je predstavljena v algoritmu 4.2. Kriterij za reortogonalizacijo je, da se norma vektorja zmanjša za več kot 30 procentov.

Algoritem 4.2 Arnoldijev algoritem s ponovljeno ortogonalizacijo. Začetni podatki so matrika A , normiran začetni vektor u_1 in dimenzija k . Algoritem vrne ortonormirane stolpce $u_1, \dots, u_k$, ki tvorijo ortonormirano bazo za podprostor Krilova $\mathcal{K}_k(A, u_1)$.

```

j = 1, 2, ..., k
  z = Au_j
  h_j = U_j^H z_j
   $\tilde{u}_{j+1} = z_j - U_j h_j$ 
  če je  $\|\tilde{u}_{j+1}\|_2 < 0.7\|z_j\|_2$ , potem
     $\hat{h}_j = U_j^H \tilde{u}_{j+1}$ 
     $h_j = h_j + \hat{h}_j$ 
     $\hat{u}_{j+1} = \tilde{u}_{j+1} - U_j \hat{h}_j$ 
   $h_{j+1,j} = \|\tilde{u}_{j+1}\|_2$ 
  če je  $h_{j+1,j} = 0$ , potem prekini računanje
   $u_{j+1} = \hat{u}_{j+1} / h_{j+1,j}$ 

```

4.1.3 Galerkinov pogoj in Ritzeve vrednosti

Približka za lastno vrednost μ in lastni vektor v iz podprostora Krilova $\mathcal{K}_k$ dobimo iz *Galerkinovega pogoja*

$$Av - \mu v \perp \mathcal{K}_k, \quad v \in \mathcal{K}_k.$$

Stolpci matrike U_k tvorijo ortonormirano bazo za podprostor $\mathcal{K}_k$, kar pri Arnoldijevem algoritmu pomeni:

- μ je lastna vrednost $k \times k$ matrike $H_k = U_k^H A U_k$,
- $v = U_k w$, kjer je w normiran lastni vektor matrike H_k , ki pripada μ .

Potem imenujemo μ *Ritzeva vrednost*, v *Ritzev vektor*, skupaj pa je (μ, v) t.i. *Ritzev par*. Za ostanek velja

$$r = Av - \mu v = h_{k+1,k} u_{k+1} e_k^T w,$$

torej

$$\|r\|_2 = |h_{k+1,k}| \cdot |e_k^T w|.$$

Ta formula nam omogoča, da velikost ostanka, kar je običajni kriterij za konvergenco, izračunamo le iz k -tega elementa vektorja w . Prav tako iz zgornje ocene sledi, da so v primeru, ko

je $h_{k+1,k}$ dovolj blizu 0, vse lastne vrednosti matrike H_k zelo dobri približki za lastne vrednosti matrike A .

Za podprostor Krilova lahko hitro preverimo, da velja

$$\mathcal{K}_k(\alpha A + \beta I, u_1) = \mathcal{K}_k(A, u_1)$$

za poljuben β in $\alpha \neq 0$. Torej dobimo isti podprostor Krilova ne glede na to, če matriko pomnožimo s skalarjem oziroma uporabimo pomik.

To se mora odražati tudi na Ritzevih vrednostih. Ker so metode podprostorov Krilova invariantne na skaliranje in pomike, je položaj lastnih vrednosti glede na koordinatno izhodišče nepomemben. Namesto tega je pomembno, katere lastne vrednosti so zunanje in katere notranje.

Če vzamemo najmanjši krog (katerega središče ni nujno koordinatno izhodišče), ki vsebuje vse lastne vrednosti, potem lahko pričakujemo, da bo za naključno izbrane začetne vektorje metoda Krilova najprej skonvergirala k tistim lastnim vrednostim, ki so blizu roba tega kroga. Obstajajo pa pristopi, kako lahko metode podprostorov Krilova pripravimo do tega, da konvergirajo k notranjim lastnim vrednostim.

Denimo, da iščemo lastne vrednosti, ki so blizu izbranega cilja $\tau \in \mathbb{C}$. Konvergenco k tem t.i. notranjim lastnim vrednostim lahko izboljšamo, če namesto za matriko A vzamemo podprostor Krilova, ki ga generira matrika $(A - \tau I)^{-1}$. Tej metodi pravimo *shift-and-invert Arnoldi*. Težava pri uporabi te metode je, da moramo v vsakem koraku, ko razširimo podprostor Krilova, rešiti sistem z matriko $A - \tau I$. Ker je to sistem z veliko in razpršeno matriko, ga moramo prav tako rešiti s katero izmed iterativnih metod. Če sistem z matriko $A - \tau I$ rešimo le približno, potem je to metoda *inexact shift-and-invert Arnoldi*.

Uporabimo lahko tudi *polinomske predpogojevanje*. V tem primeru skonstruiramo fiksen polinom p nizke stopnje, ki naj bi imel pri neželenih lastnih vrednostih majhne absolutne vrednosti in potem namesto z matriko A delamo z matriko $p(A)$.

4.1.4 Harmonične Ritzeve vrednosti

Pri Galerkinovem pogoju smo zahtevali, da je ostanek pravokoten na podprostor, v katerem iščemo približke za lastne vektorje. Lahko pa uporabimo različna podprostor. Denimo, da za matriko A iščemo približke za lastne vektorje v iskalnem podprostoru $\mathcal{V}_k$, zahtevamo pa, da je ostanek pravokoten na testni podprostor $\mathcal{W}_k$.

Približke za lastne vrednosti in vektorje potem dobimo iz t.i. *Petrov–Galerkinovega pogoja*

$$Av - \mu v \perp \mathcal{W}_k, \quad v \in \mathcal{V}_k.$$

Če sta V_k in W_k matriki z ortonormiranimi stolpci, ki razpenjajo iskalni podprostor $\mathcal{V}_k$ in testni podprostor $\mathcal{W}_k$, potem je μ lastna vrednost $k \times k$ posplošenega problema lastnih vrednosti

$$W_k^H A V_k s = \mu W_k V_k s.$$

Sedaj μ imenujemo *vrednost Petrova*, $v = V_k s$ pa *vektor Petrova*.

Dostikrat, ni pa nujno, podprostor izberemo tako, da sta V_k in W_k biortogonalni bazi, kar pomeni $W_k^H V_k = I$. Potem so vrednosti Petrova kar lastne vrednosti matrike $W_k^H A V_k$.

Pri harmoničnih Ritzevih vrednostih za podprostor $\mathcal{W}_k$ izberemo $A\mathcal{V}_k$. Iz pogoja

$$Au - \theta u \perp A\mathcal{V}_k, \quad v \in \mathcal{V}_k,$$

dobimo posplošeni problem lastnih vrednosti

$$V_k^H A^H A V_k s - \theta V_k^H A^H V_k s.$$

Naj stolpci V_k sestavljajo bazo za $\mathcal{V}_k$, ki jo izberemo tako, da stolpci $W_k = AV_k$ sestavljajo ortonormirano bazo za $A\mathcal{V}_k$. Potem se posplošeni problem spremeni v

$$V_k^H A^H V_k s = \frac{1}{\theta} s = W_k^H A^{-1} W_k s.$$

To pomeni, da je θ^{-1} Ritzeva vrednost za matriko A^{-1} glede na podprostor $A\mathcal{V}_k$. Imenujemo jo *harmonična Ritzeva vrednost* in pričakujemo, da bo θ dober približek za notranje lastne vrednosti. Izračun matrike A^{-1} ni nikoli potreben, saj je $W_k^H V_k s = \theta^{-1} s$.

Če iščemo lastne vrednosti v bližini cilja τ , dobimo

$$V^H (A - \tau I)^H (A - \tau I) V_k s = (\theta - \tau)^{-1} V_k^H (A - \tau I) V_k s.$$

4.1.5 Lanczosev algoritem

Če je matrika A simetrična, se simetričnost pri Arnoldijevem algoritmu prenese na zgornjo Hessenbergovo matriko H , ki je tako simetrična in tridiagonalna. Konstrukcija ortogonalne baze za podprostor Krilova je močno poenostavljena. Bazo dobimo z Lanczosevim algoritmom, katerega osnovna različica je predstavljena v algoritmu 4.3.

Algoritem 4.3 Lanczosev algoritem. Začetni podatki so simetrična matrika A , normiran začetni vektor u_1 in dimenzija k . Algoritem vrne ortonormirane stolpce $u_1, \dots, u_k$, ki tvorijo ortonormirano bazo za podprostor Krilova $\mathcal{K}_k(A, u_1)$.

$\beta_0 = 0, u_0 = 0$
 $j = 1, 2, \dots, k$
 $z = Au_j$
 $\alpha_j = u_j^T z$
 $z = z - \alpha_j u_j - \beta_{j-1} u_{j-1}$
 $\beta_j = \|z\|_2$
 če je $\beta_j = 0$, potem prekini računanje
 $u_{j+1} = z / \beta_j$

Z Lanczosevim algoritmom dobimo $AU_k = U_k T_k + \beta_k u_{k+1} e_k^T$, kjer je

$$T_k = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{k-1} \\ & & & \beta_{k-1} & \alpha_k \end{bmatrix}.$$

Časovna zahtevnost Lanczosevega algoritma in prostorske zahtevnosti so za red manjše kot pri Arnoldijevi metodi. Poleg tega se da zaradi simetrije matrike A več povedati o sami konvergenki Ritzevih vrednosti. Ritzeve vrednosti pri k se prepletajo z Ritzevim vrednostmi pri $k + 1$, saj se lastne vrednosti T_k prepletajo z lastnimi vrednostmi T_{k+1} . Zaradi tega Ritzeve vrednosti monotonno konvergirajo proti lastnim vrednostim matrike A .

Izrek 4.2 Za Lanczosev algoritem velja:

- a) Če so $\theta_1, \dots, \theta_k$ Ritzeve vrednosti, ki jih dobimo v k -tem koraku Lanczosevega algoritma, potem obstaja k lastnih vrednosti $\lambda_1, \dots, \lambda_k$ matrike A , da je $|\theta_i - \lambda_i| \leq \beta_k$ za $i = 1, \dots, k$.
- b) Če je $v = U_k s$ Ritzev vektor za Ritzevo vrednost θ , potem velja

$$\|Av - \theta v\|_2 = \beta_k |e_k^T s|.$$

Težava pri Lanczosevem algoritmu pa je, da hitreje kot pri Arnoldijevem algoritmu pride do izgube ortogonalnosti. Ker vsako novo smer ortogonaliziramo le na zadnja dva vektorja (kar teoretično seveda zadošča), to pri numeričnem računanju ni vedno dovolj. Kot stranski pojav se zato pojavijo tudi prividi lastnih vrednosti, ko se zunanje lastne vrednosti začnejo pojavljati kot večkratne lastne vrednosti, čeprav so enostavne.

Opazimo, da se ortogonalnost baze za podprostor Krilova poslabša takrat, ko kak izmed Ritzevih vektorjev skonvergira do lastnega vektorja. Zaradi tega, kadar želimo izračunati več lastnih vrednosti, ni priporočljivo uporabljati Lanczoseve metode brez dodatne reortogonalizacije.

Algoritem 4.4 Lanczosev algoritem s polno reortogonalizacijo. Začetni podatki so simetrična matrika A , normiran začetni vektor u_1 in dimenzija k . Algoritem vrne ortonormirane stolpce $u_1, \dots, u_k$, ki tvorijo ortonormirano bazo za podprostor Krilova $\mathcal{K}_k(A, u_1)$.

$j = 1, 2, \dots, k$
 $z = Au_j$
 $\alpha_j = u_j^T z$
 $z = z - U_k U_k^T z$
 $z = z - U_k U_k^T z$
 $\beta_j = \|z\|_2$
 če je $\beta_j = 0$, potem prekini računanje
 $u_{j+1} = z/\beta_j$

Izrek 4.3 (Paige) Če izvajamo Lanczosev algoritem v aritmetiki z osnovno zaokrožitveno napako ϵ , potem v k -tem koraku za Ritzeve vektorje $v_1, \dots, v_k$ velja

$$v_i^T q_{k+1} = \frac{\mathcal{O}(\epsilon \|A\|)}{\|r_i\|_2},$$

kjer je $r_i = Av_i - \theta_i v_i$, oziroma

$$v_i^T q_{k+1} = \frac{\mathcal{O}(\epsilon \|A\|)}{\beta_k |e_k^T s_i|},$$

kjer je $v_i = U_k s_i$.

Ena možnost je *Lanczoseva metoda s polno reortogonalizacijo*, ki je prikazana v algoritmu 4.4. Da je algoritem res stabilen, vsak vektor reortogonaliziramo dvakrat.

Namesto polne obstaja tudi selektivna reortogonalizacija, kjer novo smer reortogonaliziramo le na nekatere izmed prejšnjih smeri. Metoda je predstavljena v algoritmu 4.5. V vsakem koraku moramo izračunati vse Ritzeve vrednosti $\theta_1, \dots, \theta_k$ in pripadajoče lastne vektorje $s_1, \dots, s_k$ matrike T_k . To lahko zapišemo kot $T_k = S_k \Theta S_k^T$, kjer je $S_k = [s_1 \cdots s_k]$ in $\Theta_k = \text{diag}(\theta_1, \dots, \theta_k)$. Potem novo smer z ortogonaliziramo še na vse Ritzeve vektorje $v_i = U_k s_i$, za katere velja $\beta_k |e_k^T s_i| \leq \sqrt{\epsilon} \|T_k\|$.

Algoritem 4.5 Lanczosev algoritem s selektivno reortogonalizacijo. Začetni podatki so simetrična matrika A , normiran začetni vektor u_1 in dimenzija k . Algoritem vrne ortonormirane stolpce $u_1, \dots, u_k$, ki tvorijo ortonormirano bazo za podprostor Krilova $\mathcal{K}_k(A, u_1)$.

```

 $\beta_0 = 0, u_0 = 0$ 
 $j = 1, 2, \dots, k$ 
   $z = Au_j$ 
   $\alpha_j = u_j^T z$ 
   $z = z - \alpha_j u_j - \beta_{j-1} u_{j-1}$ 
  izračunaj  $T_k = S_k \Theta S_k^T$ 
  za vse  $i = 1, \dots, k$ , kjer je  $\beta_k |e_k^T s_i| \leq \sqrt{\epsilon} \|T_k\|$ 
     $z = z - (v_i^T z) v_i$ 
   $\beta_j = \|z\|_2$ 
  če je  $\beta_j = 0$ , potem prekini računanje
   $u_{j+1} = z / \beta_j$ 

```

4.1.6 Jacobi–Davidsonova metoda

Pri metodah, ki temeljijo na podprostorih Krilova (Arnoldi, Lanczos, ...) je podprostor, v katerem iščemo aproksimacije (Ritzeve, Petrove, harmonične Ritzeve vrednosti, ...) v bistvu odvisen le od začetnega vektorja v_1 . Lahko pa bi iskalni podprostor širili tudi drugače. Če vektorji $v_1, \dots, v_k$ tvorijo ortonormirano bazo podprostora $\mathcal{V}_k$ dimenzije k , potem je potrebno v vsakem koraku splošne metode podprostorov izbrati novo smer v_{k+1} s katero razširimo podprostor $\mathcal{V}_k$ na $\mathcal{V}_{k+1}$, da dobimo še boljše približke za iskano lastno vrednost.

Primer takšne metode je Jacobi–Davidsonova metoda, ki sta jo razvila van der Vorst in Sleijpen leta 1996. Temelji pa na Jacobijevi metodi (1840) in Davidsonovi metodi (1975).

Naj bo $\mathcal{V}_k$ podprostor dimenzije k in naj bo θ_k Ritzeva vrednost za A in $\mathcal{V}_k$ z ustreznim normiranim Ritzevim vektorjem u_k . Sedaj popravek za u_k iščemo v podprostoru $u_k^\perp$. Ortogonalna projekcija matrike A na ta podprostor je $B = (I - u_k u_k^H) A (I - u_k u_k^H)$. Zapišemo (upoštevamo $\theta_k = u_k^H A u_k$)

$$A = B + A u_k u_k^H + u_k u_k^H A - \theta_k u_k u_k^H.$$

Nastavek je $A(u_k + v) = \lambda(u_k + v)$, od tod pa iz $Bu_k = 0$ sledi

$$(B - \lambda I)v = -r + (\lambda - \theta_k - u_k^H A v)u_k.$$

Ker sta leva stran in ostanek r ortogonalna na u_k , mora biti faktor pred u_k enak 0 in dobimo

$$(B - \lambda I)v = (I - u_k u_k^H)(A - \lambda I)(I - u_k u_k^H)v = -r.$$

Namesto λ vzamemo približek θ_k in dobimo t.i. *korekcijsko enačbo*

$$(I - u_k u_k^H)(A - \theta_k I)(I - u_k u_k^H)v = -r, \quad (4.1)$$

ki jo rešimo le približno (npr. z metodo GMRES) in z novim vektorjem razširimo bazo $\mathcal{V}_k$.

Osnovna različica Jacobi–Davidsonove metode je v grobem predstavljena v algoritmu 4.6. Ko dimenzija iskalnega podprostora preveč naraste, približki pa še niso skonvergirali do lastnega para, izvedemo ponovni zagon. Iz iskalnega podprostora izberemo nekaj najboljših približkov in jih uporabimo za nov začetni podprostor.

Algoritem 4.6 Jacobi–Davidsonova metoda za iskanje dominantne lastne vrednosti matrike A . Začetni podatki so matrika A in normiran začetni vektor v_1 .

$V_1 = [v_1], u = v_1, \theta = u^H A u, r = A u - \theta u$
 $k = 1, 2, \dots$
 približno reši $(I - u u^H)(A - \theta I)(I - u u^H)t = -r, \quad t \perp u$
 ortogonaliziraj t glede na V_k (RGS) in razširi V_k v V_{k+1}
 poišči dominantni lastni par (θ, s) matrike $V_{k+1}^H A V_{k+1}$, kjer je $\|s\| = 1$
 izračunaj Ritzev vektor $u = V_{k+1}s$ in ostanek $r = A u - \theta u$
 testiraj konvergenco in po potrebi končaj

V primeru simetrične matrike Ritzeve vrednosti monotono konvergirajo proti dominantni lastni vrednosti. V bližini rešitve imamo kubično konvergenco (če bi korekcijsko enačbo reševali točno). Ker namesto točnega reševanja uporabljamo GMRES (v primeru simetrične matrike pa lahko tudi simetrično varianto MINRES), moramo poiskati kompromis med številom zunanjih iteracij in številom računanj produkta Ax .

V primeru nesimetrične matrike ali pri računanju notranjih lastnih vrednosti simetrične matrike imamo težave, ki pa se dajo odpraviti z uporabo harmoničnih Ritzevih vrednosti. Z njimi lahko poiščemo lastne vrednosti, ki so po absolutni vrednosti najmanjše in tako s premiki dobimo iskane lastne vrednosti.

Pri reševanju korekcijske enačbe (4.1) je dobro uporabiti predpogojevanje. Denimo, da imamo na voljo tako matriko K , da je $K^{-1}(A - \theta_k I) \approx I$. Pri uporabi moramo tudi K zožiti na isti podprostor, torej uporabimo

$$\tilde{K} = (I - u_k u_k^H)K(I - u_k u_k^H).$$

Če za reševanje korekcijske enačbe uporabljamo metode podprostorov Krilova, moramo v vsakem koraku izračunati produkt

$$z = \tilde{K}^{-1} \tilde{A} w,$$

kjer je $\tilde{A} = (I - u_k u_k^H)(A - \theta_k I)(I - u_k u_k^H)$. To izvedemo tako, da najprej izračunamo vektor $\tilde{A} w$. Velja

$$\tilde{A} w = (I - u_k u_k^H)g,$$

kjer je $g = (A - \theta_k I)w$, saj je $u_k^H w = 0$. Potem rešimo sistem $\tilde{K}z = (1 - u_k u_k^H)g$. Ker je $u_k^H g = 0$, vektor z zadošča sistemu $Kz = g - \beta u_k$, od koder sledi $z = K^{-1}g - \beta K^{-1}u_k$, kjer je

$$\beta = \frac{u_k^H K^{-1}g}{u_k^H K^{-1}u_k}.$$

4.1.7 Jacobi–Davidsonova metoda s harmoničnimi Ritzevimi vrednostmi

Z uporabo harmoničnih Ritzevih vrednosti lahko računamo notranje lastne vrednosti. Če je $\mathcal{V}_k$ podprostor dimenzije k , potem je θ Ritzeva vrednost matrike A , če za nek $u \in \mathcal{V}_k$, $u \neq 0$, velja

$$Au - \theta u \perp \mathcal{V}_k.$$

Pravimo, da je θ harmonična Ritzeva vrednost matrike A glede na podprostor $\mathcal{W}_k$, če je θ^{-1} Ritzeva vrednost matrike A^{-1} glede na $\mathcal{W}_k$. Naj bo $\mathcal{V}_k$ podprostor dimenzije k . Potem je θ harmonična Ritzeva vrednost A glede na $A\mathcal{V}_k$ natanko tedaj, ko za nek $u \in \mathcal{V}_k$, $u \neq 0$, velja

$$Au - \theta u \perp A\mathcal{V}_k.$$

Naj bodo stolpci matrike $V_k = [v_1 \cdots v_k]$ baza za $\mathcal{V}_k$ in stolpci matrike $W_k = [w_1 \cdots w_k]$ baza za $\mathcal{W}_k := A\mathcal{V}_k$. Potem je θ harmonična Ritzeva vrednost A glede na $A\mathcal{V}_k$ natanko tedaj, ko je

$$(W_k^H V_k)^{-1} W_k^H A V_k s = \theta s$$

za neničelni vektor $s \in \mathbb{C}^k$.

Če sta V_k in W_k taki ortogonalni bazi, da je $W_k = AV_k$, potem je $W_k^H AV_k = I$ in izračunati moramo lastne vrednosti matrike $(W_k^H V_k)^{-1}$. To lahko naredimo brez inverza, saj lastne vrednosti in vektorje dobimo iz lastnih parov $W_k^H V_k$.

Naj bo (θ_k, u_k) harmonični Ritzev par iz k -tega koraka Jacobi–Davidsonove metode. Ostanek $r = Au_k - \theta u_k$ je ortogonalen na $z_k := Au_k$. Pri korekcijski enačbi sedaj iščemo popravek, ki bo ortogonalen na z_k (za razliko od standardne metode, kjer mora biti popravek ortogonalen na u_k). Namesto ortogonalne projekcije na $z_k^\perp$ uporabimo poševno projekcijo

$$I - \frac{u_k z_k^H}{z_k^H u_k},$$

saj tako u_k pošljemo v 0. Nova korekcijska enačba je

$$\left(I - \frac{u_k z_k^H}{z_k^H u_k} \right) (A - \theta I) \left(I - \frac{u_k z_k^H}{z_k^H u_k} \right) t = -r.$$

Jacobi–Davidsonova metoda s harmoničnimi Ritzevimi vrednostmi je zapisana v algoritmu 4.7.

Algoritem 4.7 Jacobi–Davidsonova metoda s harmoničnimi vrednostmi za iskanje notranje lastne vrednosti matrike A . Začetni podatki so matrika A in normiran začetni vektor v_1 .

$$V_1 = [v_1], u = v_1, \theta = u^H A u, r = Au - \theta u, z = Au$$

$$k = 1, 2, \dots$$

$$\text{približno reši } \left(I - \frac{uz^H}{z^H u} \right) (A - \theta I) \left(I - \frac{uz^H}{z^H u} \right) t = -r, \quad t \perp z$$

ortogonaliziraj t glede na V_k (RGS) in razširi V_k v V_{k+1}

poišči minimalni lastni par (θ, s) matrike $(W_k^H V_k)^{-1} W_k^H A V_k$, kjer je $\|s\| = 1$

izračunaj Ritzev vektor $u = V_{k+1} s$, ostanek $r = Au - \theta u$ in $z = Au$

testiraj konvergenco in po potrebi končaj

4.2 Iterativne metode za QEP preko linearizacije

Dan je QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$, kjer so M, C, K matrice velikosti $n \times n$. Z eno izmed linearizacij to prevedemo na GEP $L(\lambda) = A - \lambda B$. Če so matrice M, C, K razpršene, potem to velja tudi za $2n \times 2n$ matriki A in B in za reševanje GEP lahko uporabimo iterativne metode podprostorov.

4.2.1 Arnoldijeva metoda na $A - \lambda B$

Denimo, da je matrika B nesingularna. Potem pri Arnoldijevi metodi delamo s podprostorom Krilova $\mathcal{K}_k(S, v)$, kjer je $S = B^{-1}A$. To pomeni, da moramo v vsakem koraku algoritma rešiti sistem z matriko B .

Zgled 4.1 Denimo, da za linearizacijo uporabimo standardno prvo pridruženo formo, torej je

$$A = \begin{bmatrix} 0 & N \\ -K & -C \end{bmatrix}, \quad B = \begin{bmatrix} N & 0 \\ 0 & M \end{bmatrix},$$

kjer je N poljubna nesingularna matrika. Od tod dobimo

$$S = B^{-1}A = \begin{bmatrix} 0 & I \\ -M^{-1}K & -M^{-1}C \end{bmatrix}.$$

Če poznamo faktorizacijo matrice M (ali če je npr. M diagonalna), lahko potem učinkovito generiramo podprostor Krilova za matriko S . V nasprotnem primeru za reševanje sistema z matriko B (oziroma M) spet uporabimo eno izmed iterativnih metod. $\square$

Če iščemo lastne vrednosti blizu izbrane tarče σ , potem lahko uporabimo t.i. *shift-and-invert* spektralno transformacijo. Naj bo izbrana tarča σ taka, da je matrika $A - \sigma B$ nesingularna. Iz $Ax = \lambda Bx$ sledi

$$(A - \sigma B)x = (\lambda - \sigma)Bx,$$

od tod pa

$$\frac{1}{\lambda - \sigma}x = (A - \sigma B)^{-1}Bx.$$

Za novi S vzamemo $(A - \sigma B)^{-1}B$. Če izvajamo sedaj Arnoldijevo metodo za S , dobimo lastne vrednosti QEP, ki so blizu σ .

Zgled 4.2 Če za linearizacijo uporabimo standardno prvo pridruženo formo, torej

$$A = \begin{bmatrix} 0 & N \\ -K & -C \end{bmatrix}, \quad B = \begin{bmatrix} N & 0 \\ 0 & M \end{bmatrix},$$

potem velja

$$S = (A - \sigma B)^{-1}B = \begin{bmatrix} I & 0 \\ \sigma I & I \end{bmatrix} \begin{bmatrix} Q(\sigma)^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} -(C + \lambda M) & -M \\ I & 0 \end{bmatrix}.$$

Od tod sledi, da za izračun produkta z matriko S zadošča poznati npr. LU faktorizacijo $n \times n$ matrice $Q(\sigma)$. $\square$

Namesto spektralne transformacije na matrikah A, B iz linearizacije $L(\lambda)$ lahko to naredimo tudi direktno na QEP $Q(\lambda)$. Če delamo shift-and-invert direktno na QEP, dobimo

$$\widehat{Q}(\mu) = \mu^2 \widehat{M} + \mu \widehat{C} + \widehat{K},$$

kjer je $\mu = 1/(\lambda - \sigma)$ in

$$\begin{aligned}\widehat{M} &= \sigma^2 M + \sigma C + K = Q(\sigma) \\ \widehat{K} &= M \\ \widehat{C} &= C + 2\sigma M = Q'(\sigma).\end{aligned}$$

Če na linearizaciji uporabljamo Arnoldijevo metodo (ali pa katerokoli drugo metodo podprostorov, kot je npr. dvostranska Lanczoseva metoda), potem delamo z matrikami $2n \times 2n$ in vektorji dolžine $2n$, ki so oblike $\begin{bmatrix} x \\ \lambda x \end{bmatrix}$. Obstaja pa tudi nekaj pristopov, kako rešiti QEP preko metod podprostorov z vektorji in matrikami reda n .

4.2.2 SOAR (Second-order Arnoldi)

Metodo SOAR sta razvila Bai in Su leta 2005. Podrobno je predstavljena v članku [2].

Definicija 4.4 *Dani sta matriki A, B in vektor $u_0 \neq 0$. Vektorji iz zaporedja*

$$\begin{aligned}r_0 &= u_0, \\ r_1 &= Ar_0, \\ r_2 &= Ar_{j-1} + Br_{j-2} \quad \text{za } j \geq 2\end{aligned}$$

tvorijo zaporedje Krilova 2. reda. Podprostor $\mathcal{G}_n(A, B, u) = \text{Lin}(r_0, \dots, r_{n-1})$ imenujemo podprostor Krilova 2. reda.

Podprostor Krilova 2. reda je posplošitev podprostora Krilova. V primeru $B = 0$ tako velja $\mathcal{G}_n(A, B, u) = \mathcal{K}_n(A, u)$.

Denimo, da QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$ lineariziramo v GEP $L(\lambda) = F - \lambda G$, kjer je

$$F = \begin{bmatrix} -C & -K \\ I & 0 \end{bmatrix}, \quad B = \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix}.$$

Potem je

$$S = G^{-1}F = \begin{bmatrix} -M^{-1}C & -M^{-1}K \\ I & 0 \end{bmatrix} =: \begin{bmatrix} A & B \\ I & 0 \end{bmatrix}.$$

Če izvajamo standardno Arnoldijevo metodo na $F - \lambda G$ z začetnim vektorjem $v = \begin{bmatrix} u \\ 0 \end{bmatrix}$, potem za $j \geq 1$ velja

$$\begin{bmatrix} r_j \\ r_{j-1} \end{bmatrix} = S^j v.$$

Algoritem 4.8 Metoda SOAR za izračun ortogonalne baze za $\mathcal{G}_j(A, B, u)$. Začetni podatki so matriki A, B in neničelni začetni vektor u .

```

 $q_1 = u / \|u\|_2, \quad p_1 = 0$ 
 $j = 1, \dots, n$ 
   $r = Aq_j + Bp_j, \quad s = q_j$ 
   $i = 1, \dots, j$ 
     $t_{ij} = q_i^T r$ 
     $r = r - t_{ij}q_i$ 
     $s = s - t_{ij}p_i$ 
   $t_{j+1,j} = \|r\|_2$ 
  če je  $t_{j+1,j} = 0$ , potem končaj
   $q_{j+1,j} = (1/t_{j+1,j})r$ 
   $p_{j+1,j} = (1/t_{j+1,j})s$ 

```

Ta zveza nakazuje, da bi namesto $K_n(S, v)$ morale zadoščati, če poznamo podprostor $\mathcal{G}_j(A, B, u)$. Z algoritmom SOAR, ki je predstavljen v algoritmu 4.8, skonstruiramo ortogonalno bazo za $\mathcal{G}_j(A, B, u)$. Izkaže se (glej [2]), da lahko računanje baze izvedemo brez računanja vektorjev p_i s približno pol manj dela.

Algoritem 4.8 se konča, če je $t_{j+1,j} = 0$. To se lahko zgodi zaradi dveh vzrokov:

- a) Deflacija. Vektorji $r_0, \dots, r_j$ so linearno odvisni, vektorji

$$\begin{bmatrix} r_0 \\ 0 \end{bmatrix}, \begin{bmatrix} r_1 \\ r_0 \end{bmatrix}, \dots, \begin{bmatrix} r_j \\ r_{j-1} \end{bmatrix} \quad (4.2)$$

pa ne. V tem primeru lahko nadaljujemo tako, da vzamemo $q_{j+1} = 0$, $t_{j+1,j} = 1$ in $p_{j+1} = s$. Vektorji $q_0, \dots, q_{n-1}$ še vedno tvorijo bazo za $\mathcal{G}_n(A, B, u)$, katerega dimenzija je manjša od n .

- b) Zaustavitev. V tem primeru so tudi vektorji (4.2) linearno odvisni. Do zaustavitve pride natanko pri istem j , pri katerem se zaustavi tudi navadni Arnoldijev algoritem za matriko

$$S \text{ z začetnim vektorjem } v = \begin{bmatrix} u \\ 0 \end{bmatrix}.$$

Grob algoritem, kako uporabimo SOAR za računanje lastnih vrednosti QEP, je predstavljen v algoritmu 4.9. Prednost SOAR pred uporabo Arnoldijeve metode na linearizaciji je, da pri SOAR metodi projicirani manjši QEP ohrani lastnosti originalnega problema, kot so npr. simetričnost, hiperboličnost in žiroskopskost. Vse te lastnosti izgubimo, če uporabimo Arnoldijevo metodo na linearizaciji.

4.2.3 Jacobi–Davidsonova metoda

Tudi Jacobi–Davidsonovo metodo lahko posplošimo na QEP. Pri tej metodi tako kot pri SOAR delamo z originalnim problemom, zato se lastnosti problema lahko prenesejo tudi na manjši projicirani problem.

Algoritem 4.9 Uporaba SOAR za računanje lastnih vrednosti QEP.

1. Zgeneriraj ortonormirano bazo $Q_m = [q_1 \ \cdots \ q_m]$ z začetnim vektorjem u za $\mathcal{G}_n(A, B, u)$, kjer je $A = -M^{-1}C$, $B = -M^{-1}K$.
2. $M_m = Q_m^T M Q_m$, $C_m = Q_m^T C Q_m$, $K_m = Q_m^T K Q_m$
3. Za reduciran QEP $(\theta^2 M_m + \theta C_m + K_m)g = 0$ izračunaj Ritzev par (θ, z) , kjer je $z = Q_m g$.
4. Oцени natančnost Ritzevega para, kriterij je

$$\frac{\|(\theta^2 M + \theta C + K)z\|}{|\theta|^2 \|M\| + |\theta| \|C\| + \|K\|} < \epsilon.$$

Denimo, da rešujemo QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$. Naj bo $\mathcal{V}_k$ podprostor dimenzije k , v katerem iščemo približke za lastne vektorje. Galerkinov pogoj za Ritzev par (θ, u) , kjer je $u \in \mathcal{V}_k$, $u \neq 0$, je

$$Q(\theta)u \perp \mathcal{V}_k.$$

Če so stolpci matrike $V_k = [v_1 \ \cdots \ v_k]$ ortonormirana baza za $\mathcal{V}_k$, potem dobimo projicirani QEP

$$(\theta^2 M_k + \theta C_k + K_k)s = 0,$$

kjer je $M_k = V_k^T M V_k$, $C_k = V_k^T C V_k$, $K_k = V_k^T K V_k$ in $u = Qs$.

Za razširitev iskalnega podprostora iščemo $s \perp u$, da je $Q(\lambda)(u + s) = 0$. Če to razvijemo okrog znanega približka za lastno vrednost θ , dobimo

$$Q(\theta)s = -Q(\theta)u + (Q(\theta) - Q(\lambda))u + (Q(\theta) - Q(\lambda))s. \quad (4.3)$$

Lema 4.5 Naj bo $v \in \mathcal{V}_k$ tak vektor, da je $v \perp r$, kjer je $r = Q(\theta)u$. Če je (λ, x) lastni par QEP in $v^H Q'(\lambda)x \neq 0$, potem je $|\lambda - \theta| = \mathcal{O}(\|s\|)$.

Pogoj $v^H Q'(\lambda)x \neq 0$ je izpolnjen, če je λ enostavna lastna vrednost in je vektor v dovolj blizu levega lastnega vektorja. Zgornjo lemo potem pokažemo s pomočjo razvoja (1.7). Od tod sledi, da za tretji člen na desni strani (4.3) lahko ocenimo, da je $\mathcal{O}(\|s\|^2)$ in ga zanemarimo.

Drugi člen $(Q(\theta) - Q(\lambda))u$ ne zanemarimo, temveč uporabimo projekcijo, da ga projiciramo stran. Če aproksimiramo

$$(Q(\theta) - Q(\lambda))u \approx (\theta - \lambda)Q'(\theta)u,$$

potem je prava projekcija

$$I - \frac{Q'(\theta)u u^H}{u^H Q'(\theta)u},$$

ki je dobro definirana, če je $v^H Q'(\theta)u \neq 0$. Najbolj naravna izbira za vektor v je, da vzamemo kar $v = u$. V tem primeru iskalni podprostor razširimo z vektorjem $s \perp u$, ki reši (približno) korekcijsko enačbo

$$\left(I - \frac{Q'(\theta)u u^H}{u^H Q'(\theta)u} \right) Q(\theta)(I - u u^H)s = -r.$$

Če iščemo lastne vrednosti v bližini tarče τ , potem lahko uporabimo varianto s harmoničnimi Ritzevimi vrednostmi. Če je iskalni podprostor $\mathcal{V}_k$, potem je θ harmonična Ritzeva vrednost natanko tedaj, ko je

$$Q(\theta)u \perp Q(\tau)\mathcal{V}_k$$

za $u \in \mathcal{V}_k$, $u \neq 0$. Več podrobnosti o tem lahko najdete v [8].

4.2.4 Alternative za Rayleighov kvocient

Hochstenbach in van der Vorst sta v [9] predstavila nekaj alternativ, ki jih lahko uporabimo namesto Rayleighovega kvocienta, da za dani vektor dobimo približek za lastno vrednost QEP.

Naj bo (λ, x) , kjer je $x \neq 0$, lastni par za QEP $Q(\lambda)x = (\lambda^2 M + \lambda C + K)x = 0$. Če je vektor u približek za lastni vektor x , potem pri Rayleighovem kvocientu približek za lastno vrednost θ dobimo iz Galerkinovega pogoja, kjer mora veljati

$$r(\theta, u) := (\theta^2 M + \theta C + K)u \perp u.$$

Od tod dobimo kvadratno enačbo $m\theta^2 + c\theta + k = 0$, kjer je $m = u^H M u$, $c = u^H C u$ in $k = u^H K u$. Enačba ima rešitvi θ_1 in θ_2 . Če je $u = x$, potem ena izmed rešitev v splošnem nima pomena, druga rešitev pa je enaka lastni vrednosti. Če je u blizu x , potem se lahko s primerjanjem norme ostankov $r(\theta_1, u)$ in $r(\theta_2, u)$ odločimo, katero izmed dveh rešitev vzeti za približek lastne vrednosti. Kadar pa u ni tako dobra aproksimacija za x , lahko s tem postopkom ne pridemo do dovolj dobrih približkov. V nadaljevanju bo predstavljenih nekaj drugih možnosti, ki jih lahko uporabimo kot približek za lastno vrednost.

Denimo, da je u približek za lastni vektor EP $Ax = \lambda x$. V tem primeru lahko primerjamo vektorja u in Au . Če je u lastni vektor, potem sta vektorja u in Au kolinearna in je naravno vzeti dolžino projekcije Au na u kot približek za lastno vrednost. Izkaže se, da je to kar Rayleighov kvocient $\rho(A, u)$.

Pri QEP imamo tri vektorje: Mu , Cu in Ku . Če je u lastni vektor, potem ti trije vektorji ležijo v isti ravnini in posplošitev prejšnjega principa z EP je, da vse tri vektorje projiciramo na isto ravnino. Pri tem predpostavimo, da sta pri lastnem vektorju x vektorja Mx in Cx linearno neodvisna, Kx je linearna kombinacija Mx in Cx , lastna vrednost λ pa je enolična (kar pomeni, da x ni lastni vektor še za kakšno drugo lastno vrednost).

Definiramo posplošeni residual

$$r(\mu, \nu, u) = (\mu M + \nu C + K)u.$$

Z dodatnimi pogoji bi radi dosegli, da bo μ dober približek za λ^2 , ν pa za λ . Potem lahko kot približek za λ vzamemo tudi μ/ν .

Če p in q tvorita bazo za ravnino, na katero bomo projicirali vektorje Mu , Cu in Ku , potem dobimo za μ in ν naslednji sistem

$$\begin{aligned} r(\mu, \nu, u) &\perp p \\ r(\mu, \nu, u) &\perp q, \end{aligned}$$

ki ga lahko zapišemo v obliki

$$W^H z \begin{bmatrix} \mu \\ \nu \end{bmatrix} = -W^H K u, \quad (4.4)$$

kjer je $Z = \begin{bmatrix} Mu & Cu \end{bmatrix}$ in $w = \begin{bmatrix} p & q \end{bmatrix}$. Za p in q izberemo vektorja iz $\text{Lin}(Mu, Cu, Ku)$.

Možnosti je več:

- a) Izberemo t.i. *ravnino najmanjših kvadratov*, kjer p in q izberemo tako, da minimizirata

$$\|(I - P)Mu\|_2^2 + \|(I - P)Cu\|_2^2 + \|(I - P)Ku\|_2^2,$$

kjer je P ortogonalna projekcija na ravnino $\text{Lin}(p, q)$. Rešitev je, da za p in q vzamemo dominantna leva singularna vektorja matrike $D = \begin{bmatrix} Mu & Cu & Ku \end{bmatrix}$. Iz enačbe (4.4) potem izračunamo μ in ν . Poleg ν in μ/ν lahko za približek za λ vzamemo tudi

$$\theta = \arg \min_{\theta \in \mathbb{C}} \left\| \begin{bmatrix} \theta^2 \\ \theta \end{bmatrix} - \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right\|_2.$$

Tej rešitvi pravimo *argmin rešitev*. Če pišemo $\theta = \theta_1 + i\theta_2$, potem dobimo sistem dveh polinomov stopnje 3 za θ_1 in θ_2 , ki imata rezultanto stopnje 5.

- b) Kot približek za lastno vrednost lahko vzamemo

$$\theta = \arg \min_{\theta \in \mathbb{C}} \|(\theta^2 M + \theta C + K)u\|_2,$$

saj pri EP za Rayleighov kvocient velja $\rho(A, U) = \arg \min_{\sigma \in \mathbb{C}} \|Ax - \sigma x\|_2$. Če tako kot pri prejšnji točki pišemo $\theta = \theta_1 + i\theta_2$, potem dobimo sistem dveh polinomov stopnje 3 za θ_1 in θ_2 . Rezultanta je tokrat stopnje 9.

Če pa že vnaprej vemo, da je $\theta \in \mathbb{R}$ (npr., če je problem hiperboličen), potem dobimo le kubično enačbo za θ .

- c) Minimiziramo normo residuala $\|r(\mu, \nu, u)\|_2$ po $(\mu, \nu) \in \mathbb{C}^2$. Rešitev je

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} = -Z^+ Ku,$$

kjer je $Z = \begin{bmatrix} Mu & Cu \end{bmatrix}$. Ko določimo μ in ν , nadaljujemo tako kot v točki a). Izkaže se, da se to ujema s projiciranjem na ravnino, ki jo tvorita vektorja Mu in Cu , torej z izbiro $p = Mu$ in $q = Cu$.

Ko po enem izmed zgoraj opisanih načinov izračunamo približek za lastno vrednost, lahko to izkoristimo za iterativno izboljšanje vektorja v metodi podprostorov, ki jo uporabljamo. Groba ideja postopka je prikazana v algoritmu 4.10.

Algoritem 4.10 Iterativno izboljšanje približka za lastni vektor.

Na standardni Ritz–Galerkinov način določi približek (θ, u) za lastni par in vzemi $u_1 = u$.
 $k = 1, 2, \dots$

iz u_k določi θ_k z eno izmed alternativ za Rayleighov kvocient

določi minimalni singularni vektor s_k za matriko $\theta_k^2 AU + \theta_k BU + CU$

$u_k = Us_k$

4.2.5 Računanje več lastnih vrednosti

Pri EP in GEP lahko izkoristimo Schurovo formo, da izračunamo več lastnih vrednosti. Pri Arnoldijevi (Lanczosevi) metodi dobimo več lastnih vrednosti naenkrat, pri Jacobi–Davidsonovi pa eno samo. Tudi pri Arnoldijevi metodi pa lahko tiste lastne vrednosti, ki so že skonvergirale, ločimo od ostalih.

Denimo, da smo pri Arnoldijevi metodi za EP z matriko A že našli lastne vektorje $v_1, \dots, v_{k-1}$. Sedaj nadaljujemo z v_k , vektorje $v_1, \dots, v_{k-1}$ pa pustimo zaklenjene v bazi. To pomeni, da v nadaljevanju uporabljamo podprostor $\text{Lin}(v_1, \dots, v_{k-1}, v_k, Av_k, \dots, A^{m-k}v_k)$. Prvih $k-1$ vektorjev je zaklenjenih, preostali vektorji pa so aktivni.

Druga možnost je, da nadaljujemo z matriko $\tilde{A} = (I - U_{k-1}U_{k-1}^H)A(I - U_{k-1}U_{k-1}^H)$, kjer stolpci matrike $U_{k-1} = [u_1 \ \dots \ u_{k-1}]$ sestavljajo ortonormirano bazo za invariantni podprostor, ki smo ga že našli.

Težje je delati deflacijo na QEP, saj nimamo na voljo Schurove forme. Meerbergen je v [12] predstavil metodo za zaklepanje lastnih vrednosti QEP. Ključnega pomena je naslednja lema, ki povezuje podprostore in Ritzeve vrednosti za QEP in pripadajoči GEP.

Naj bo QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$, uporabimo pa linearizacijo GEP $A - \lambda B$, kjer sta

$$A = \begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix}, \quad B = \begin{bmatrix} -C & -M \\ -M & 0 \end{bmatrix}.$$

Lema 4.6 *Naj bo V matrika z ortonormiranimi stolpci. Če projekcija QEP na $\text{im}(V)$ vrne Ritzev par (θ, u) , potem projekcija GEP na $\text{im}\left(\begin{bmatrix} V & 0 \\ 0 & V \end{bmatrix}\right)$ vrne Ritzev par $\left(\theta, \begin{bmatrix} u \\ \theta u \end{bmatrix}\right)$.*

Pri GEP $Ax = \lambda Bx$ je Schurova forma definirana preko $B^{-1}A$ kot $B^{-1}Ax = XS$, kar je ekvivalentno $AX = BXS$, kjer je S zgornja trikotna, X pa unitarna matrika. Delna Schurova forma je $AX_p = BX_p S_p$, kjer je S_p zgornja trikotna matrika velikosti $p \times p$, matrika X_p pa je velikosti $n \times p$ z ortonormiranimi stolpci.

Če je V_k matrika z ortonormiranimi stolpci, ki tvorijo bazo za iskalni podprostor, potem dobimo približek za Schurovo formo $AU_k - BU_k S_k = F_k$, kjer je F_k ostanek, za katerega zaradi Galerkinovega pogoja velja $V_k^H F_k = 0$. To lahko zapišemo kot $V_k^H A V_k X_k - V_k^H B V_k X_k S_k = 0$, kjer je X_k taka matrika, da je $U_k = V_k X_k$.

Meerbergen podobno kot Bai in Su pri SOAR izkorišča povezave med Schurovimi vektorji in vrednostmi QEP in GEP, ki ga dobimo z linearizacijo. Tako npr. za linearizacijo vedno vemo, da obstaja delna Schurova forma

$$\begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix} \begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix} = \begin{bmatrix} -C & -M \\ -M & 0 \end{bmatrix} \begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix} S_{2k}, \quad (4.5)$$

kjer ima $2n \times 2k$ matrika $\begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix}$ ortonormirane stolpce, matrika S_{2k} velikosti $2k \times 2k$ pa je zgornja trikotna. Pri tem je pomembno opozoriti na to, da matrika U_{2k} ni nujno polnega ranga.

Definicija 4.7 Če je W_{2k} matrika $n \times 2k$ z ortonormiranimi stolpci, T_{2k} zgornja trikotna $2k \times 2k$ matrika z lastnimi vrednostmi na diagonalni in je

$$MW_{2k}T_{2k}^2 + CW_{2k}T_{2k} + KW_k = 0,$$

pravimo, da je to delna Schurova forma za QEP.

Delna Schurova forma za QEP ne obstaja vedno. Zadosten pogoj je podan v naslednji lemi.

Lema 4.8 Če ima matrika U_{2k} iz (4.5) poln rang, potem obstaja delna Schurova forma za QEP.

Dokaz. Iz (4.5) sledi $Y_{2k} = U_{2k}S_{2k}$, od tod pa

$$MU_{2k}S_{2k}^2 + CU_{2k}S_{2k} + KU_{2k} = 0.$$

Če je $U_{2k} = W_{2k}Z_{2k}$ QR razcep matrike U_{2k} , je matrika $T_{2k} = Z_{2k}S_{2k}Z_{2k}^{-1}$ zgornja trikotna in dobimo delno Schurovo formo, saj velja $MW_{2k}T_{2k}^2 + CW_{2k}T_{2k} + KW_k = 0$. ■

Ideja je, da podprostor, podobno kot pri SOAR, gradimo direktno za QEP, potem pa si pomagamo z lemo 4.6, da iz istega podprostora dobimo dvakrat večji podprostor za GEP. Na GEP lahko naredimo delno Schurovo formo in potem ustrezno zmanjšamo osnovni podprostor oziroma zaklenemo njegov del.

Na k -tem koraku imamo k -dimenzionalni podprostor za QEP, ki ga razpenjajo stolpci matrike V_k . Za GEP uporabimo $2k$ -dimenzionalni podprostor, ki ga razpenjajo stolpci matrike $\begin{bmatrix} V_k & 0 \\ 0 & V_k \end{bmatrix}$. Schurovi vektorji za GEP imajo potem obliko

$$\begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix} = \begin{bmatrix} V_k & 0 \\ 0 & V_k \end{bmatrix} \begin{bmatrix} X_{1k} \\ X_{2k} \end{bmatrix}.$$

Če vpeljemo $M_k = V_k^H M V_k$, $C_k = V_k^H C V_k$ in $K_k = V_k^H K V_k$, od tod dobimo

$$\begin{bmatrix} K_k & 0 \\ 0 & -M_k \end{bmatrix} \begin{bmatrix} X_{1k} \\ X_{2k} \end{bmatrix} = \begin{bmatrix} -C & -M \\ -M & 0 \end{bmatrix} \begin{bmatrix} X_{1k} \\ X_{2k} \end{bmatrix} S_{2k},$$

kar nam da $X_{2k} = X_{1k}S_{2k}$. Približek za delno Schurovo formo za GEP je torej

$$\begin{bmatrix} K & 0 \\ 0 & -M \end{bmatrix} \begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix} - \begin{bmatrix} -C & -M \\ -M & 0 \end{bmatrix} \begin{bmatrix} U_{2k} \\ Y_{2k} \end{bmatrix} S_{2k} = \begin{bmatrix} F_{2k} \\ 0 \end{bmatrix}$$

in velja

$$MU_{2k}S_{2k}^2 + CU_{2k}S_{2k} + KU_{2k} = F_{2k}.$$

Tudi če je $F_{2k} = 0$ to še ni nujno delna Schurova forma za QEP, saj matrika U_{2k} ni unitarna oziroma ni nujno polnega ranga.

Deflacija oz. zaklepanje lastnih vektorjev poteka pri GEP na naslednji način. Če razdelimo U_k na $U_k = [U_r \quad U_{k-r}]$, potem je

$$A[U_r \quad U_{k-r}] - B[U_r \quad U_{k-r}] \begin{bmatrix} S_r & S_{r,k-r} \\ 0 & S_{k-r} \end{bmatrix} = [F_r \quad F_{k-r}].$$

V primeru, ko je $\|F_r\| < \epsilon$, proglasimo $F_r = 0$ in zaklenemo U_r . To pomeni, da obravnavamo prvih r Ritzevih parov kot da so že skonvergirali do točnih lastnih parov.

Pri QEP zaklepanje izvedemo preko GEP. Zaradi zvez med QEP in GEP, je potrebno zakleniti prvih r stolpcev matrik V_k , X_{2k} in S_{2k} .

Podobno ravnamo pri ponovnem zagonu. Ponovni zagon uporabimo pri metodah podprostorov vsakič, ko postane iskalni podprostor prevelik. Tako pri GEP

$$A \begin{bmatrix} U_p & U_{k-p} \end{bmatrix} - B \begin{bmatrix} U_p & U_{k-p} \end{bmatrix} \begin{bmatrix} S_p & S_{p,k-p} \\ 0 & S_{k-p} \end{bmatrix} = \begin{bmatrix} F_p & F_{k-p} \end{bmatrix}$$

reduciramo na

$$AU_p - BU_p S_p = F_p$$

tako, da izločimo zadnjih $k - p$ Schurovih vektorjev. Pri QEP moramo zaradi zvez med GEP in QEP ponovni zagon narediti tako, da ga v bistvu izvedemo na GEP in iz podprostora dimenzije $2k$ izločimo $2k - 2p$ vektorjev. To moramo narediti na način, da je novi podprostor oblike

$$\begin{bmatrix} W_p & 0 \\ 0 & W_p \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix},$$

da lahko potem W_p uporabimo za začetni podprostor za QEP. Podrobnosti so opisane v [12].

Za razširitev iskalnega podprostora lahko uporabimo tudi inverzno iteracijo z residuali. Če za tarčo vzamemo σ_k in je tekoči približek za lastni par (ω_k, u_k) , potem pri inverzni iteraciji z residuali za QEP za nov vektor dobimo (pred normiranjem) $u_k - Q(\sigma_k)^{-1}Q(\omega_k)u_k$. Če delamo s podprostori lahko to uporabimo za razširitev podprostora tako, da za novo smer vzamemo $Q(\sigma_k)^{-1}Q(\omega_k)u_k$.

Če pa inverzno iteracijo z residuali izvajamo na linearizaciji $A - \lambda B$, potem za novo smer vzamemo $(A - \sigma_k B)^{-1}(A - \omega_k)B\tilde{u}_k$, kjer je $\tilde{u}_k$ vektor velikosti $2n$. Izkaže se, da sta iteraciji za QEP in GEP povezani z naslednjo lemo.

Lema 4.9 Če je

$$\begin{bmatrix} x \\ y \end{bmatrix} = (A - \sigma B)^{-1}(A - \omega) \begin{bmatrix} u \\ \omega u \end{bmatrix},$$

potem je $x = Q(\sigma)^{-1}Q(\omega)u$.

Poglavje 5

Metode podprostorov za splošni nelinearni problem

5.1 Uvod

Dan je problem $T(\lambda)x = 0$, kjer je $T : \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ gladka preslikava. Velja še, da je n velik in da so matrice $T(\lambda)$ razpršene. To pomeni, da za večino elementov velja, da so identično enaki 0 (torj za vsak $\lambda \in \mathbb{C}$), le $\mathcal{O}(n)$ je takih, ki niso identično enaki 0.

Podobno kot pri EP z razpršenimi matrikami, kjer imamo na voljo metode podprostorov, bi tudi pri NEP radi uporabili podobne metode.

Denimo, da je $V \in \mathbb{C}^{n \times k}$ matrika z ortonormiranimi stolpci, ki tvorijo bazo za iskalni podprostor $\mathcal{V}$. Tako kot pri ostalih problemih lastnih vrednosti (EP, GEP, QEP), tudi sedaj približke za lastne pare dobimo iz projiciranega problema

$$V^H T(\theta) V y = 0. \quad (5.1)$$

Pravimo, da je potem θ Ritzeva vrednost, vektor Vy , kjer je $y \neq 0$, pa Ritzev vektor. Projiciran problem je spet NEP, le da so matrice sedaj velikosti $k \times k$. Pri povsem splošnem NEP potem naletimo na problem, kako priti do rešitev projiciranega problema, saj je potrebno učinkovito izračunati $V^H T(\theta) V$ za različne θ . Večina NEP je takšne oblike, da jih sestavljajo fiksne matrice in parameter λ , kar pomeni, da lahko projiciramo matrice $n \times n$ v matrice $k \times k$ in potem rešimo manjši problem. To velja npr. za vse PEP.

Naj bo (θ, x) približek za lastni par. Sedaj potrebujemo novo smer za razširitev iskalnega podprostoru. Iz algoritmov, ki smo jih obravnavali v poglavju 2 za reševanje polnih NEP, dobimo naslednji ideji za izračun nove smeri:

a) inverzna iteracija: $v_a = T(\theta)^{-1} T'(\theta) x$,

b) inverzna iteracija z residuli: $v_b = x - T(\sigma)^{-1} T(\theta) x$, kjer je σ fiksna vrednost.

Izkaže se, da izbira a) ustreza posplošitvi Jacobi–Davidsonove metode, izbira b) pa posplošitvi Arnoldijeve metode.

5.2 Nelinearna Arnoldijeva metoda

Denimo, da za novo smer izberemo $x - T(\sigma)^{-1}T(\theta)x$. Ker bomo novo smer ortogonalizirali na prejšnje in ker je $x \in \mathcal{V}$, lahko za novo smer vzamemo kar $v = T(\sigma)^{-1}T(\theta)x$.

Zgled 5.1 V primeru GEP $T(\lambda) = A - \lambda B$ nas to pripelje do Arnoldijeve metode. velja namreč

$$\begin{aligned} T(\sigma)^{-1}T(\theta) &= (A - \sigma B)^{-1}(A - \theta B) \\ &= (A - \sigma B)^{-1}(A - \sigma B + (\sigma - \theta)B) \\ &= I + (\sigma - \theta)(A - \sigma B)^{-1}B. \end{aligned}$$

Vemo, da je podprostor Krilova invarianten na množenje s skalarjem in premike. To pomeni, da v primeru GEP dobimo podprostor Krilova za $(A - \sigma)^{-1}B$, kar ustreza metodi shift-and-invert Arnoldi.

Zaradi zgornje zveze imenujemo to metodo *nelinearna Arnoldijeva metoda*. Namesto reševanja sistema $T(\sigma)v = T(\theta)x$ ponavadi uporabimo predpogojevanje in vzamemo $v = MT(\theta)x$, kjer je $M \approx T(\sigma)^{-1}$.

Ker je konvergenca odvisna od tega, kako dober približek je σ , lahko premik σ in matriko M med algoritmom po potrebi posodobimo. V grobem je algoritem predstavljen v algoritmu 5.1.

Algoritem 5.1 Nelinearna Arnoldijeva metoda

izberi začetni premik σ in začetno bazo V za iskalni podprostor

določi matriko $M \approx T(\sigma)^{-1}$ za predpogojevanje

$m = 1, 2, \dots$

poišči lastni par (θ, y) projiciranega problema $V^H T(\theta) V y = 0$

$u = Vy, r = T(\theta)u$

če je $\|r\|/\|u\| \leq \epsilon$ smo našli lastno vrednost

$\lambda_m = \theta, x_m = u$

izberi nov premik θ in novo matriko $M \approx T(\sigma)^{-1}$

(po potrebi ponovni zagon)

pošči novi Ritzev par (θ, u) za naslednji lastni par

$r = T(\theta)u$.

$v = Mr$

$V = \text{rgs}([V \ v])$

(reortogonalizacija, če je potrebna)

Na začetku algoritma 5.1 moramo izbrati začetni iskalni podprostor $\mathcal{V}$. Tu nam pridejo zelo prav kakšne dodatne informacije o lastnih vrednostih in pripadajočih lastnih vektorjih, če jih seveda imamo na voljo. Sicer pa, če iščemo lastne vrednosti v bližini τ , lahko najprej z ključnim začetnim vektorjem naredimo nekaj korakov Arnoldijeve metode bodisi za problem lastnih vrednosti $T(\sigma)z = \mu z$ ali za posplošeni problem lastnih vrednosti $T(\sigma)z = \mu T'(\sigma)z$, potem pa vzamemo tiste vektorje, ki pripadajo po absolutni vrednosti najmanjšim lastnim vrednostim.

Če želimo izračunati več lastnih vrednosti, moramo poskrbeti, da metoda ne skonvergira k že izračunanim lastnim vrednostim. Zaradi tega, ker pri NEP nimamo na voljo Schurove forme, je to zelo težko.

Za konvergenco inverzne iteracije z residuali vemo, da je linearna in odvisna od razdalje $|\tau - \lambda|$. Zato je počasna konvergenca znak, da bi bilo dobro spremeniti premik τ .

5.3 Jacobi–Davidsonova metoda

Naj bo $\mathcal{V}$ iskalni podprostor, ki ga razpenjajo ortonormirani stolpci matrike V . Denimo, da smo iz tega podprostora dobili Ritzev par (θ, x) , kjer je $x = Vy$ za $y \neq 0$ in $V^H T(\theta) Vy = 0$. Če za izhodišče vzamemo Jacobi–Davidsonovo metodo za QEP, potem bi bila naravna izbira za korekcijsko enačbo

$$\left(I - \frac{px^H}{x^H p}\right) T(\sigma) \left(I - \frac{xx^H}{x^H x}\right) v = -r, \quad (5.2)$$

kjer je $p = T'(\theta)x$ in $r = T(\theta)x$.

Iz korekcijske enačbe sledi $T(\theta)v = -r + \alpha p$, kjer skalar α določimo tako, da je $v \perp x$. Od tod dobimo

$$v = -x + \alpha T(\theta)^{-1} T'(\theta)x.$$

Ker je $x \in \mathcal{V}$, je za novo smer ekvivalentno vzeti $T(\theta)^{-1} T'(\theta)x$, to pa ustreza inverzni iteraciji. Pričakujemo lahko kvadratično (v hermitskem primeru in za $\lambda \in \mathbb{R}$ pa celo kubično) konvergenco (če bi korekcijsko enačbo rešili eksaktno).

Dobljeni algoritem, prikazan v algoritmu 5.2, se zelo malo razlikuje od algoritma 5.1.

Algoritem 5.2 Nelinearna Jacobi–Davidsonova metoda

izberi začetni premik σ in začetno bazo V za iskalni podprostor

določi matriko $K \approx T(\sigma)$ za predpogojevanje

$m = 1, 2, \dots$

poišči lastni par (θ, y) projiciranega problema $V^H T(\theta) Vy = 0$

$u = Vy, \quad r = T(\theta)u$

če je $\|r\| / \|u\| \leq \epsilon$ smo našli lastno vrednost

$\lambda_m = \theta, \quad x_m = u$

izberi nov premik θ in novo matriko $K \approx T(\sigma)$

(po potrebi ponovni zagon)

poišči novi Ritzev par (θ, u) za naslednji lastni par

$r = T(\theta)u, \quad p = T'(\theta)u$

reši korekcijsko enačbo $\left(I - \frac{pu^H}{u^H p}\right) T(\sigma) \left(I - \frac{uu^H}{u^H u}\right) v = -r$

(reortogonalizacija, če je potrebna)

Skoraj nujno je uporabljati predpogojevanje. Če je $K \approx T(\sigma)$, potem za predpogojevanje korekcijske enačbe (5.2) uporabimo

$$\tilde{K} = \left(I - \frac{px^H}{x^H p}\right) K \left(I - \frac{xx^H}{x^H x}\right).$$

Pri popravljeni korekcijski enačbi s predpogojevanjem tako iščemo novo smer $v \perp x$, za katero

velja

$$\tilde{K}^{-1} \left(I - \frac{px^H}{x^H p} \right) T(\sigma) \left(I - \frac{xx^H}{x^H x} \right) v = -\tilde{K}^{-1} r.$$

Če želimo izračunati več kot le eno lastno vrednost, imamo pri nelinearnih problemih lastnih vrednosti težave, saj ne obstaja posplošitev Schurove forme. Vseeno pa lahko, če uporabljamo Jacobi–Davidsonovo metodo, uporabimo selekcijski kriterij, s katerim izločimo že izračunane lastne pare.

Naj bo λ_0 lastna vrednost, x_0 in y_0 pa njen desni in levi lastni vektor. V izreku 1.6 smo pokazali, da v primeru, ko je lastna vrednost enostavna, velja $y_0 T'(\lambda_0) y_0 \neq 0$. Sedaj definiramo deljeno diferenco

$$T[\lambda, \mu] := \begin{cases} \frac{T(\lambda) - T(\mu)}{\lambda - \mu} & \text{za } \lambda \neq \mu \\ T'(\lambda) & \text{za } \lambda = \mu. \end{cases},$$

ki je zvezno odvedljiva. Hitro se lahko prepričamo, da velja

$$y_i^H T[\lambda_i, \lambda_j] x_j = 0,$$

če je y_i levi lastni vektor za lastno vrednost λ_i in x_j desni lastni vektor za lastno vrednost $\lambda_j \neq \lambda_i$. Če pa je λ_i enostavna lastna vrednost, potem je $y_i^H T[\lambda_i, \lambda_i] x_i \neq 0$.

Na podlagi tega lahko razvijemo selekcijski kriterij, ki ga uporabimo pri izbiri Ritzevih vektorjev in vrednosti v Jacobi–Davidsonovi metodi. Denimo, da smo z metodo že izračunali d lastnih vrednosti s pripadajočimi levimi in desnimi lastnimi vektorji $(\lambda_1, x_1, y_1), \dots, (\lambda_d, x_d, y_d)$. Sedaj definiramo

$$\eta = \frac{1}{2} \min_{i=1, \dots, d} |y_i^H T'(\lambda_i) x_i|.$$

Pri izbiranju Ritzevega para (θ, u) sedaj zahtevamo, da mora za vse $i = 1, \dots, d$ veljati

$$|y_i^H T[\lambda_i, \theta] u| < \eta.$$

Ta pogoj zagotavlja, da par (θ, u) ne bo skonvergiral k enemu izmed že izračunanih lastnih parov. Res je, da za kriterij potrebujemo še leve lastne vektorje, a ko enkrat poznamo lastne vrednosti, jih lahko enostavno dobimo z inverzno iteracijo.

Zgled 5.2 Za poseben primer, ko imamo QEP $Q(\lambda) = \lambda^2 M + \lambda C + K$, dobimo

$$Q[\lambda, \mu] = (\lambda + \mu)M + C.$$

Literatura

- [1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst: *Templates for the solution of algebraic eigenvalue problems: A practical guide*, SIAM, Philadelphia, 2000.
- [2] Z. Bai, Y. Su, *SOAR: A second-order Arnoldi method for the solution of the quadratic eigenvalue problem*, SIAM J. Matrix Anal. Appl. **26** (2005), 640–659.
- [3] T. Betcke, *Efficient Methods For Nonlinear Eigenvalue Problems*, Diploma thesis, TU Hamburg-Harburg, 2002.
- [4] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, F. Tisseur: *NLEVP: A Collection of Nonlinear Eigenvalue Problems*, 2008.
- [5] R. J. Duffin: *A minimax theory for overdamped networks*, J. Rational Mech. Anal. **4** (1955), 221–233.
- [6] H. Fan, W. Lin, P. Van Dooren, *Normwise scaling of second order polynomial matrices*, SIAM J. Matrix Anal. Appl. **26** (2004), 252–256.
- [7] N. J. Higham, S. D. Mackey, F. Tisseur: *The Conditioning of Linearizations of Matrix Polynomials*, SIAM J. Matrix Anal. Appl. **28** (2006), 1005–1028.
- [8] M. E. Hochstenbach, G. L. Sleijpen, *Harmonic and refined Rayleigh–Ritz for the polynomial eigenvalue problem*, Numer. Linear Algebra Appl. **15** (2008), 35–54.
- [9] M. E. Hochstenbach, H. A. van der Vorst: *Alternatives to the Rayleigh quotient for the quadratic eigenvalue problem*, SIAM J. Sci. Comp. **25** (2003), 591–603.
- [10] P. Lancaster: *Lambda-matrices and vibrating systems*, Pergamon Press, Oxford, 1966.
- [11] A. S. Markus, *Introduction to the spectral theory of polynomial operator pencils*, Amer. Math. Soc., Providence, 1988.
- [12] K. Meerbergen, *Locking and restarting quadratic eigenvalue solvers*, SIAM. J. Sci. Comput. **22** (2001), 1814–1839.
- [13] V. Mehrmann, H. Voss: *Nonlinear Eigenvalue Problems: A Challenge for Modern Eigenvalue Methods*, GAMM Mitt. Ges. Angew. Math. Mech. **27** (2004) 121–152.
- [14] A. Neumaier: *Residual inverse iteration for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal. **22** (1985) 914–923.
- [15] B. Plestenjak: *Numerical methods for the tridiagonal hyperbolic quadratic eigenvalue problem*, SIAM J. Matrix Anal. Appl. **28** (2006) 1157–1172.

- [16] A. Ruhe: *Algorithms for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal. **10** (1973) 674–689.
- [17] K. Schreiber: *Nonlinear Eigenvalue Problems: Newton-type Methods and Nonlinear Rayleigh Functionals*, PhD thesis, TU Berlin, 2008.
- [18] F. Tisseur: *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl. **309** (2000) 339–361.
- [19] F. Tisseur, K. Meerbergen: *The quadratic eigenvalue problem*, SIAM Review **43** (2001), 235–286.