De numerieke benadering van de overdruk in de ondergrond met grote verschillen in permeabiliteit

L. El yaakoubi

Afstudeercommissie:
Prof.dr.ir. P. Wesseling (TU Delft),
Dr.ir. C. Vuik (TU Delft),
Ir. A.Segal (TU Delft),
Ir. E.Dufour (Shell RTS Rijswijk).

Inhoudsopgave

Hoofdstuk 1

Inleiding

Shell is een onderneming met een wereldwijde aanwezigheid in meer dan 120 landen. De activiteiten van Shell liggen in de eerste plaats op het gebied van olie en gas, maar Shell heeft ook belangen in chemie, steenkool, metalen, bosbouw en duurzame energie (biomassa, zon). Shell is in olie en gas actief van "put tot pomp", research, exploratie en produktie, trading en marketing, transport en verwerking.

Bij Research and Technical Services (RTS) in Rijswijk (onderdeel van Shell International Exploration and Production B.V.) houden ongeveer 900 medewerkers zich bezig met het ondersteunen van technologische ontwikkelingen in de werkmaatschappijen en nieuwe ondernemingen van Exploratie en Productie over de hele wereld. Het ontwikkelen van software vormt een van de belangrijkste activiteiten, maar ook op het gebied van services, bijvoorbeeld het verzamelen en verwerken van seismische gegevens, speelt RTS een vooraanstaande rol. De laatste tijd is de nieuwe manier van werken in een stroomversnelling geraakt. Hiermee wordt bedoeld onder andere het werken in teamverband in teamruimtes aan specifieke, vaak tijdelijke projecten en het opzetten van zogenaamde 'virtual global teams', waarbij een deel van de teamleden in Rijswijk werkt en een ander deel in een werkmaatschappij of nieuwe onderneming. De leden van dergelijke teams profiteren van de laatste ontwikkelingen op het gebied van kennis-uitwisseling via informatietechnologie en nieuwe media.

Een van de teams houdt zich bezig met het modelleren van geologische processen. Aan de hand van de geologische geschiedenis probeert men de kans op de aanwezigheid van winbare hoeveelheden gas en olie te voorspellen. Daarnaast is men sterk geïnteresseerd in de drukken van de in de bodem aanwezige vloeistoffen en gassen.

Een belangrijke reden hiervoor is dat in het geval van een hoge druk, boren met de nodige voorzichtigheid moet plaatsvinden. Door een te grote druk in de ondergrond kunnen zogenaamde spuiters ontstaan met alle nodige gevaren voor mens en materiaal. In de praktijk is het voorgekomen dat door het plotseling ontsnappen van een gasbel de bodem een drijfzandachtig karakter kreeg, waardoor een boortoren verzakte. Op zee zal het plotseling ontsnappen van een gas het drijfvermogen van een platform ernstig kunnen verminderen met alle gevolgen vandien. Een ander belangrijk aspect van de druk is dat olie kan wegsijpelen door de lagen heen. Het gevolg is dat er feitelijk geen olie meer te winnen is uit een gebied waar men olie had verwacht. Een ander belangrijk aspect van de druk is dat hij invloed heeft op de akoestische eigenschappen van de materie. Kennis van de overdruk is dus nodig om seismische gegevens beter te kunnen interpreteren. Het is voor een oliemaatschappij dus van zeer groot belang kennis te hebben van de drukken, die heersen in alle lagen van de ondergrond.

Hoofdstuk 2

Achtergrond van het probleem

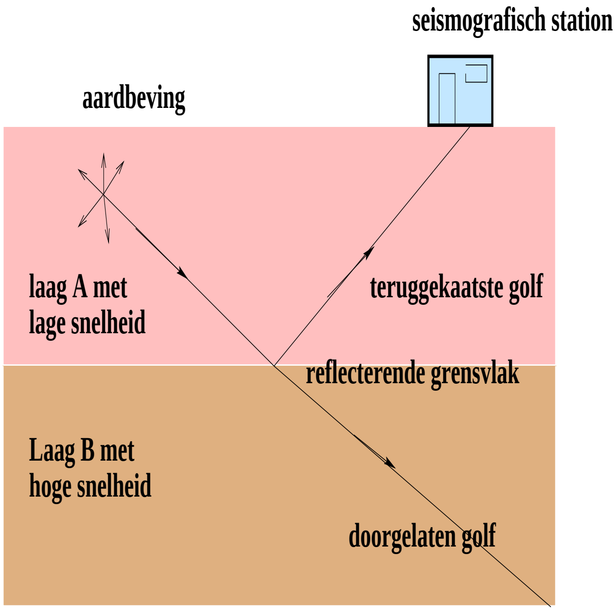
2.1 Geologiekennis

Het begrijpen van de veranderingen die op de aarde plaatsvinden en het begrijpen van de oorzaken van die veranderingen, is de doelstelling van de geologie. De fysische geologie richt zich op de samenstelling van de aarde en op de beschrijving en verklaringen van de dynamische processen en krachten, die van invloed zijn op het oppervlak en het inwendige van de aarde. Geologen zijn niet alleen geïnteresseerd in aardbevingen en vulkaanuitbarstingen, maar ook in praktische vraagstukken, zoals hoe gesteente verweert, hoe bodems ontstaan, wat het gedrag is van gesteente onder hoge temperatuur en druk, en dergelijke.

De aarde kan men beschouwen als een enorme machinerie, die door twee processen wordt beinvloed, een intern en een extern proces. Beide processen werken op warmte, ze zetten warmte-energie om in mechanische energie. Het interne proces wordt aangedreven door warmte, die vanuit het warme binnenste van de aarde naar het koele oppervlak getransporteerd wordt. Het externe proces wordt hoofdzakelijk door de zon aangedreven. De warmte van de zon voorziet de oceanen en de atmosfeer van bewegingsenergie. De beschrijving van de manier waarop de aardse machinerie werkt, wordt onderbouwd door de theorie van de plaattektoniek. Het idee achter plaattektoniek is dat het aardoppervlak verdeeld is in een paar grote, dikke platen, die langzaam ten opzichte van elkaar bewegen en in grootte veranderen.

De warmteprocessen van de aarde die al miljarden jaren aan de gang zijn, hebben bepaalde materialen in verschillende delen van de wereld geconcentreerd. Door kennis van de geologie, over het ontstaan van deze materialen en over het waarom van de verdeling ervan over het aardoppervlak, zijn wij in staat deze materialen met gericht onderzoek op te sporen. Evenals kennis van de geologie van belang is bij het opsporen van deze materialen, kan kennis van de geologie van belang zijn ten behoeve van het verminderen of voorkomen van schade aan het milieu.

2.2 Structuur van de aarde

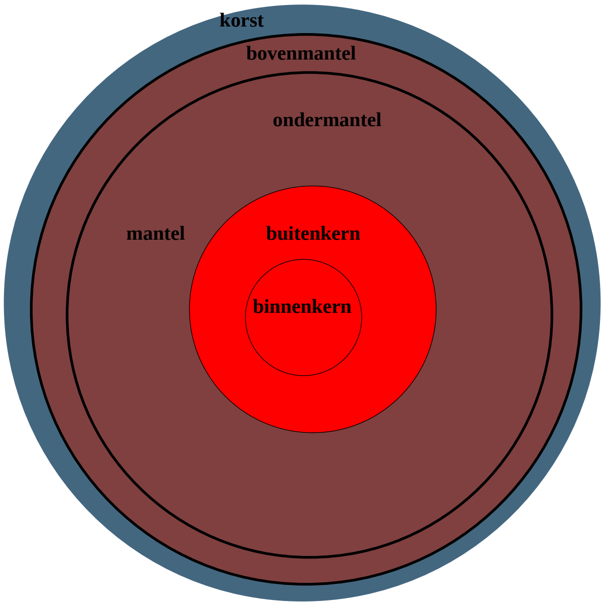


Figuur 2.1: De schematische weergave van de banen van seismische golven

De aarde heeft een straal van ongeveer 6370 kilometer, dus het is duidelijk dat geologen slechts net onder het oppervlak kunnen komen om het gesteente op een directe manier te bestuderen. Diepere delen van de aarde worden op een indirecte wijze bestudeerd. Dit gebeurt voor een groot deel met behulp van de geofysica [4], een tak van de geologie die natuurkundige wetten en theorieën gebruikt voor de bestudering van de aarde. De geofysica maakt gebruik van de voorplanting van de aardbevingstrillingen, het magneetveld, het zwaartekrachtsveld en van de warmte-uitstraling van de aarde. Door studie van trillingen ten gevolge van hevige aardbevingen en door de mens gemaakte explosies, wordt enige informatie verzameld over de structuur van de aarde. Eén van de manieren om van seismische trillingen gebruik te maken is het kijken naar reflecties en refracties van seismische golven. Aan een grensvlak tussen twee verschillende aardlagen wordt een binnenkomende golf ten dele naar de oppervlakte teruggekaatst (reflectie) en voor het overige deel onder een ander hoek doorgelaten (refractie)(zie figuur 2.1).

Studie van seismische reflecties en refracties heeft de wetenschap in staat gesteld de belangrijkste onderverdeling van de aarde vast te stellen (zie figuur 2.2). De buitenste laag is de korst, die slechts een dunne schil aan het aardoppervlak vormt. Onder de korst ligt de mantel. Deze is onderverdeeld in de bovenmantel en de ondermantel. De mantel bestaat uit vast gesteente. In het centrum van de aarde bevindt zich de kern, die uit een vaste binnenkern en een vloeibare buitenkern bestaat. De kern is van metaal en is de bron van het aardmagneetveld.

De korst en het bovenste deel van de mantel worden samen de lithosfeer genoemd (lithos betekent 'steen' in het Grieks). De mantel direct onder de



Figuur 2.2: De onderverdeling van de aarde

lithosfeer wordt asthenosfeer genoemd. Volgens de 'Plaattektoniek' theorie [4] is de lithosfeer opgebroken in platen, die ten opzichte van elkaar over de onderliggende asthenosfeer kunnen bewegen.

2.3 Geologische tijd

Het is zeker zo dat sommige geologische processen heel snel gaan, zoals landverschuivingen. Maar deze processen zijn een uiting van het plotseling vrijkomen van grote hoeveelheden opgeslagen energie. De meeste geologische processen gaan langzaam, maar gestaag. Zo zal een heuvel gedurende de periode van een mensenleven niet zichtbaar eroderen, maar wel over een periode van een miljoen jaar. De ouderdom van de aarde wordt geschat op 4.5 miljard (4500 miljoen) jaar.

Fossielen kunnen gebruikt worden om een standaard geologische tijdschaal, een wereldwijde relatieve tijdschaal, te construeren. Gebaseerd op fossielassociaties kunnen geologen voor elk tijdvak bepalen welke fossielen daarin kunnen voorkomen. De geologische tijdschaal kent, zoals te zien is in figuur 2.3, vier era's. Deze era's zijn onderverdeeld in perioden, die op hun beurt onderverdeeld zijn in tijdvakken.
Maar pas sinds de ontdekking van radioactiviteit kan men de absolute ouderdom van een gesteente bepalen. Radioactiviteit, het natuurlijke verval van de atoomkernen van bepaalde elementen waarbij straling uitgezonden wordt, levert een klok op die begint te tikken op het momemt dat bepaalde soorten gesteente gevormd worden. Dankzij deze gegevens was het mogelijk aan de geologische tijdschaal, die voorheen een relative schaal was, absolute ouder-

Geologische Tijdschaal

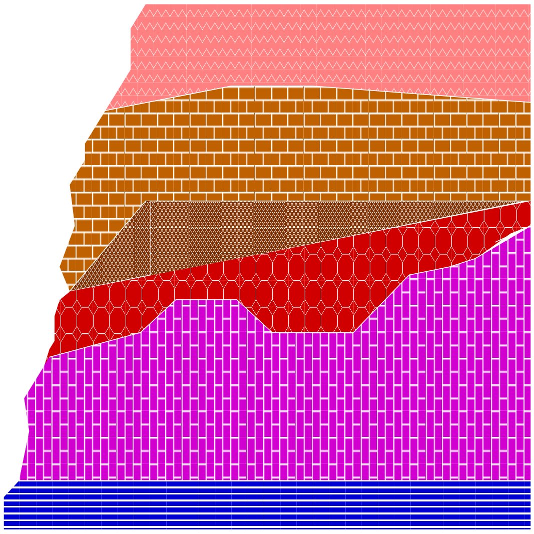
Era	Periode	Tijdvak	Ouderdom in miljoenen jaren voor heden	
	Kwartair	heden (Holeceen)	0.01	
		Pleistoceen	1.6	
Kenozoicum	Tertiair	Plioceen	5.3	
		Mioceen	23.7	
		Oligoceen	36.6	
		Palecoceen	57.8	
		Eoceen		66.4
Mesozoicum	Krijt		144	
	Jura		208	
	Trias			245
Paleozoicum	Perm		286	
	Carboon		360	
	Deoon		408	
	Siluur		438	
	Ordovicium		505	
	Cambrium			570
Precambrium	Proterzoisch			2500
	Archeisch Eon			
		ontstaan van de aarde		4500

Figuur 2.3: Geologische tijdschaal

dom te geven. Nu kan men met een hoge graad van betrouwbaarheid zeggen dat bijvoorbeeld het Kenozoicum zo'n 66 miljoen jaar geleden begon en het Paleozoicum ongeveer 570 miljoen jaar geleden. Het oudste gesteente dat werd gevonden op aarde is ongeveer 4 miljard jaar oud.

2.4 Sedimenten en sedimentaire gesteenten

Erosie is de belangrijkste proces die het gesteente dat door het interne warmteproces omhoog geduwd is, weer geheel afvlakken. Het stromen van water en lucht zorgt voor de erosie van het verweerde gesteente. Gesteente dat diep in de aarde onder hoge temperatuur en druk is ontstaan en door tektonische krachten omhoog is geduwd, is aan het oppervlak instabiel. Lucht en water breken de kristallen af en vorming van nieuw materiaal vindt plaats. Deze



Figuur 2.4: Een schets van een gelaagde struktuur

omzetting van instabiel materiaal heet verwering.
Sediment is de verzamelnaam voor al het losse materiaal dat afkomstig is van erosie van vast gesteente en chemische neerslag. Indien materiaal uit transport tot rust komt, treedt depositie of afzetting op. Sediment wordt afgezet, als stromend water of wind energie verliezen en hun vracht niet langer kunnen vervoeren. Meestal verzamelen dergelijke deeltjes zich in een laag aan het aardoppervlak. Sedimentaire structuren zijn verschijnselen in sedimentaire gesteente die meestal tijdens of vlak na afzetting zijn ontstaan. Een van de meest opvallende sedimentaire structuren is gelaagdheid. Meestal is de gelaagdheid horizontaal, omdat de meeste sedimenten in horizontale pakketten worden afgezet (zie figuur 2.4).
Sedimentdeeltjes worden naar korrelgrootte geclassificeerd. Alle deeltjes groter dan 2 millimeter worden samen wel als grind aangeduid, maar een verdere opdeling als in tabel 1.2 wordt ook gehanteerd. Bijvoorbeeld zand omvat alle korrels tussen 2 millimeter en 60 micron (0.06mm). Kleideeltjes zijn in grootte kleiner dan 2 micron.
Sedimentkorrels die in een rustig milieu tot afzetting komen, vormen een structuur waarbij veel open ruimte tussen de korrels aanwezig is. Dergelijke open ruimtes worden aangeduid met de term poriën of holtes. Het totaal volume aan holtes wordt poriënvolume genoemd. Bij doorgaande sedimentatie zullen, onder invloed van het toenemende gewicht, eerder afgezette sedimentkorrels worden samengedrukt, waarbij het poriënvolume afneemt. Dit proces wordt aangeduid met de term compactie.
Het percentage van een gesteente dat uit holle ruimtes bestaat, noemt men de porositeit. Het is een maat voor de hoeveelheid water die het gesteente kan bevatten. Bijna alle gesteenten kunnen water vasthouden. Sedimentaire gesteenten als zandsteen hebben in het algemeen een hoog poriëngehalte en kunnen dus veel water bevatten. Hoewel dus bijna alle gesteenten enig wa-

ter kunnen bevatten, zijn ze zeer variabel in de mate waarin het erdoorheen kan stromen. De permeabiliteit of doorlatendheid is de mate waarin vloeistoffen (water, maar ook aardolie) door een gesteente kunnen stromen. De orde van grootte van permeabiliteit en porositeit voor verschilende gesteentesoorten wordt gegeven in tabel (2.1).

soortgesteente	porositeit(in %)	permeabiliteit(in Darcy)
zand	37- 50	$0.2.10^2 - 1.8.10^2$
zandsteen	4 - 50	$5.10^{-4} - 30.10^{-1}$
Lei	0 - 80	$2.10^{-11} - 10^{-6}$
kalksteen	4 - 10	$10^{-11} - 5.10^{-10}$

Tabel 2.1: Porositeit en permeabiliteit van verschillende de soorten

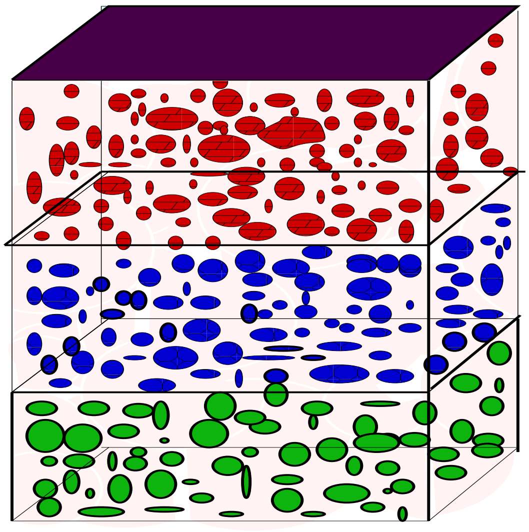
2.5 Ontstaan van olie

Van organische substanties blijven koolwaterstofverbindingen in het sediment over. Deze zijn bekend onder de verzamelnaam bitumen, dat gasvormig, vloeibaar of vast kan zijn.
Bitumen in sedimenten zijn het basismateriaal voor het ontstaan van aardolie. Het kan in bepaalde gevallen aan kleigesteenten gebonden zijn (olielei) en in zulke gevallen alleen met een ingewikkeld destillatieproces worden gewonnen. In de meeste gevallen heeft de natuur er al voor gezorgd. Als bitumineuze klei- of leisteen in de aardkorst aan sterke verhitting onderhevig is en bovendien onder zware druk komt te staan, wordt het bitumen van de klei gescheiden. Ze heeft een veel lager soortelijk gewicht dan de gesteenten en stijgt op, migreert, waar spleten en holten, of de porisiteit van gesteenten dit mogelijk maken. Het is mogelijk, dat de vloeistof op deze wijze tot het oppervlak doordringt. Er onstaan dan natuurlijke oliebronnen. In de meeste gevallen wordt de in de diepte ontstane aardolie op haar weg naar boven gestuit door afdichtende gesteenten, die geen druppel doorlaten. Dan wijkt de vloeistof uit naar poreuze of caverneuze gesteenten (vooral zandsteen of poreuze kalkgesteenten), waar ze zich verzamelt, accumuleert. Dergelijke gesteenten worden reservoirgesteenten genoemd.
In het volgende hoofdstuk zullen we een fysische beschrijving geven van het proces van stromen van olie en water en de daarbij behorende drukken.

Hoofdstuk 3

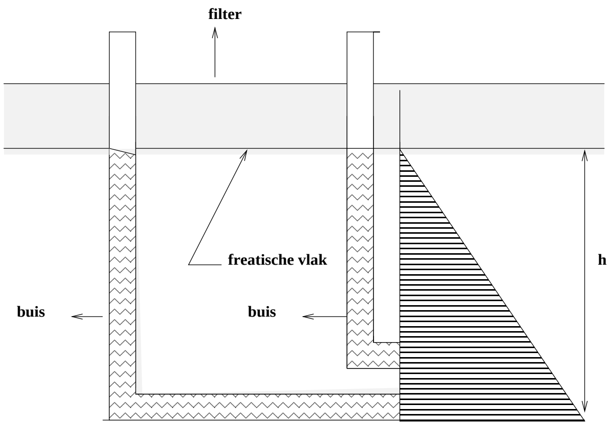
Beschrijving van de fysica

3.1 Beschrijving van het fysische probleem



Figuur 3.1: Een schets van structuur een poreus medium met verschillende gesteentes

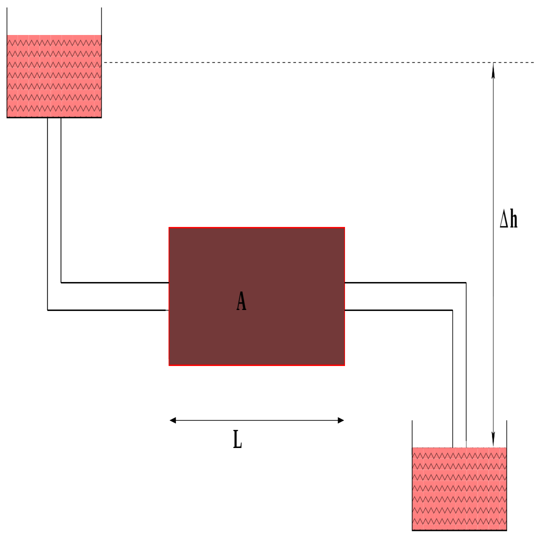
Er wordt uitgegaan van een gebied in een poreus medium waarin een vloeistof stroomt. Een poreus medium is een medium waarin de deformatiesnelheid van de rots en de vloeistof in principe verschillend zijn. Dit gebied bestaat uit sedimentlagen (zie figuur 3.1). Iedere sedimentlaag wordt beschouwd als een isotroop medium. Hij bevat een soort van gesteentedeeltje, gekenmerkt door een permeabiliteit k en porositeit ϕ.
Door deze gelaagdheid van de ondergrond is het medium anisotroop, dat houdt in dat het medium in verschillende richtingen verschillende eigenschappen bezit. De permeabiliteit hangt af van richting waarin gemeten wordt.
Wordt een buis A in een isotroop grondpakket(filter) geplaatst, dan blijkt na enige tijd het water dat in de buis gestegen is tot een hoogte h die evenredig is aan de gemeten waterdruk ter plaatse van het filter. Wanneer water in verticale richting geheel in rust is , zal de bovenkant van het water in de peilbuis



Figuur 3.2: Hydrostatisch verloop van de waterdruk in de grond

overeenkomen met het freatisch vlak: dit betekent dat het verloop van de waterdruk hydrostisch is (zie figuur (3.2). Deze hoogte heet h de *stijghoogte* en de druk in de freatisch vlak is gelijk aan de atmosferische druk.

Voor een eendimensionale stroming in een isotroop medium toonde Darcy



Figuur 3.3: Experiment van Darcy

experimenteel aan in 1856 aan dat de snelheid van een vloeistof v door een poreus medium (zie figuur 3.3) recht evenredig was met het verhang J(J=verlies van stijghoogte per lengte eenheid) [2].

$$v = CJ = C\frac{\Delta h}{L}.$$

C is een constante, afhankelijk van de eigenschappen van de vloeistof en het poreuze materiaal.
Om de vloeistofbeweging aan de wet van Darcy te kunnen voldoen wordt verondersteld dat de stroming viskeus en laminair is, dat wil zeggen dat een vloeistofdeeltje zich verplaatst langs de stroomlijnen die slechts geleidelijk van richting veranderen. In elk punt van de vloeistof is de snelheid voldoende klein om geen turbulentie te veroorzaken.
De porositeit wordt gedefinieerd als verhouding van het volume ingenomen door de vloeistof ten opzichte van het totaal volume:

$$\phi = \frac{V_{fluid}}{V_{totaal}}.$$

De spanningen in het poreuze gesteente worden voor een deel veroorzaakt door de geconcentreerde krachten die ze in hun contactpunten overbrengen, en gedeeltelijk door de druk in de vloeistof. Dit leidt er toe als een zinvolle spanningsgrootheid in te voeren het verschil tussen de gemiddelde verticale spanning σ_v en de vloeistof druk p_f [5]

$$\sigma_{eff} = \sigma_v - p_f. \tag{3.1}$$

Men noemt dit de effectieve spanning.
Bij modellering van ons probleem zullen wij met het verschijnsel compactie rekening houden. Dat wil zeggen dat de sedimentjes samengedrukt worden door het gewicht van de erop afgezette lagen. In dit geval wordt de porositeit afhankelijk van druk en de diepte.
Uit de praktijk is gebleken dat de maximale effectieve spanning afname van porositeit kan veroorzaken.
Bij eerste benadering geldt dat de porositeit ϕ van het medium een exponentiele functie van de effectieve druk σ_{eff} is [5]:

$$\phi = \phi_0 \exp^{-c\sigma_{eff}}.$$

Om de analyse eenvoudig te houden kijken we eerst naar een lineare uitdrukking voor ϕ:

$$\frac{d\phi}{d\sigma_{eff}} = \alpha. \tag{3.2}$$

Om ons mathematische model te vereenvoudigen kunnen een paar fysische aannames gemaakt worden:

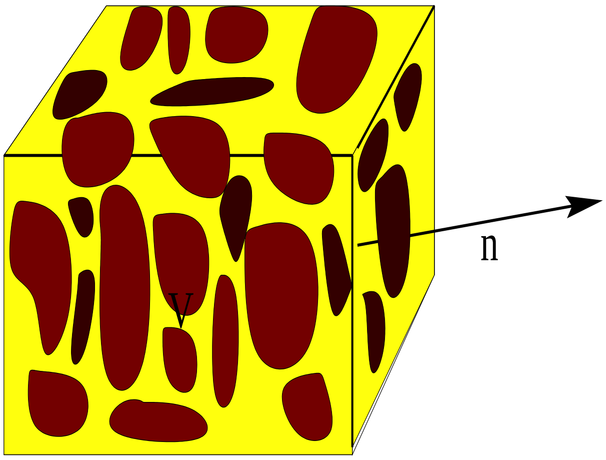
- Het medium is onsamendrukbaar, d.w.z de dichtheid zowel van de rots ρ_s als van de vloeistof ρ_v zijn onafhankelijk van de tijd t en de plaats $\bar{x}$: dus $\rho_s(t, \bar{x}) = \rho_s$ en $\rho_v(t, \bar{x}) = \rho_v$.

- In het medium vinden geen destructie en geen generatie van zowel de rotsdeeltjes als van de vloeistof plaats (bijvoorbeeld door chemische reacties)

Voor bepaling van de overdruk in de ondergrond zullen we van de volgende twee fysische wetten gebruik maken:

- de wet van massabehoud
- de wet van Darcy

3.2 Wet van massabehoud



Figuur 3.4: Volumeelement

Beschouw een volume V, omsloten door een oppervlak A , waarop **n** als uitwendige normaal gedefinieerd is (zie figuur 3.4).
Volgens de wet van massabehoud (als er geen bronnen in V aanwezig zijn) is de verandering van de massa in V gelijk aan de netto massaflux door A [5]:

$$\int_V \frac{\partial(\phi\rho)}{\partial t} dV = \int_A (\phi\rho\mathbf{u}).\mathbf{n} dA.$$

Hierin geldt : ρ is de dichtheid van het beschouwde medium en $\mathbf{u}$ is de absolute snelheid.
De divergentie stelling levert :

$$\int_V \frac{\partial(\phi\rho)}{\partial t} = - \int_V \nabla(\phi\rho\mathbf{u}) dV.$$

Omdat deze relatie voor een willekeurig volume V moet gelden, heeft de massabehoudswet de volgende vorm:

$$\frac{\partial(\phi\rho)}{\partial t} + \nabla.(\phi\rho\mathbf{u}) = 0. \tag{3.3}$$

Omdat er wordt aangenomen dat de dichtheden ρ voor beide stoffen onafhankelijk van de plaats en de tijd zijn, luidt de massabehoudwet (3.3) voor beide stoffen als volgt:

$$\frac{\partial(\phi)}{\partial t} + \nabla.(\phi \mathbf{u_f}) = 0.$$
$$\frac{\partial((1-\phi))}{\partial t} + \nabla.((1-\phi)\mathbf{u_s}) = 0.$$

Gebruikmakend van de volgende notatie voor de operator $\frac{D}{Dt}$:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{u_s}.\nabla,$$

wordt nu de continuiteitsvergelijking voor de vaste stof(rots) en voor de vloeistof op de volgende wijze geschreven:

$$\frac{1}{1-\phi}\frac{D\phi}{Dt} = \nabla.\mathbf{u_s}. \quad \text{voor rots} \tag{3.4}$$

$$\frac{1}{\phi}\frac{D\phi}{Dt} = -\nabla.\mathbf{u_f} + \frac{1}{\phi}(\mathbf{u_s} - \mathbf{u_f}).\nabla\phi. \quad \text{voor vloeistof} \tag{3.5}$$

Omdat zowel vloeistof als rots in beweging zijn, is er sprake van een gemiddelde vloeistofsnelheid q die gedefinieerd als volgt:

$$\mathbf{q} = \phi(\mathbf{u_f} - \mathbf{u_s}). \tag{3.6}$$

Door optellen van vergelijkingen (3.4) en(3.5) en verder door de substitutie van uitdrukking(3.6) in de verkregen vergelijking wordt na enige berekeningen een relatie tussen de porositeit en de gemiddelde relatieve vloeistofsnelheid afgeleid:

$$-\nabla.\mathbf{q} = \frac{1}{1-\phi}\frac{D\phi}{Dt}. \tag{3.7}$$

3.3 Wet van Darcy

De vergelijking van Darcy is geldig voor een stroming in een poreus medium waarin de weerstandkrachten tegen de stroming domineren. In ons fysische model zullen we aannemen dat de traagheidskrachten ten opzichte van de viskeuze krachten verwaarloosbaar zijn.
Voor een driedimensionale stroming in een poreus medium blijft de wet van Darcy ook geldig. Dit legt ook een lineair verband tussen de hydraulische gradient J en de gemiddelde vloeistofsnelheid q [2]:

$$\mathbf{q} = \frac{\mathbf{k}}{\mu}\mathbf{J}. \tag{3.8}$$

De permeabiliteit k is een eigenschap van het rotsdeeltje terwijl μ een eigenschap van de vloeistof is.
De hydraulische gradient $\mathbf{J}$ is gedefinieerd als de totale kracht per volumeëenheid

die nodig is om de vloeistof in beweging te brengen.
Op een vloeistof werken twee soorten krachten: de oppervlaktekracht $-\nabla p_f$ onder invloed van de drukgradient en de volume kracht $\rho_f \mathbf{g}$ onder invloed van de gravitatie .
De totale kracht per eenheid volume $\mathbf{J}$ wordt gegeven door:

$$\mathbf{J} = -\nabla p_f + \rho_f \mathbf{g}. \tag{3.9}$$

De overdruk p wordt gedefinieerd als het verschil tussen de vloeistof druk p_f en de hydrostatische druk p_h :

$$p = p_f - p_h. \tag{3.10}$$

De hydrostatische druk wordt gegeven door:

$$p_h = \rho_f g z. \tag{3.11}$$

Door combinatie van (3.9), (3.10) en (3.11) wordt de hydraulisch gradient $\mathbf{J}$ gelijk gesteld aan minus de gradient van p:

$$\mathbf{J} = -\nabla p.$$

Hiermee wordt de vergelijking van Darcy als volgt geformuleerd:

$$\mathbf{q} = -\frac{\mathbf{k}}{\mu} \nabla p. \tag{3.12}$$

3.4 Afleiding van de differentiaalvergelijking voor de overdruk

Substitutie van (3.12) in (3.7) en gebruikmakend van (3.2) leveren:

$$\frac{1}{\mu} \nabla.(\mathbf{k} \nabla p) = \frac{\alpha}{1-\phi} \frac{D\sigma_{eff}}{Dt}. \tag{3.13}$$

Uit vergelijkingen (3.1) en (3.10) wordt afgeleid dat de effectieve spanning een functie is van de overdruk:

$$\sigma_{eff} = \sigma - p.$$

Hierbij wordt σ gegeven door [5]:

$$\sigma(z) = \sigma_v - p_h = g \int_{z_0}^{z} (1-\phi)(\rho_s - \rho_f) dz.$$

Hierin geeft z de diepte aan.
Door deze uitdrukking van de effectieve spanning te gebruiken wordt de vergelijking (3.13) als volgt geschreven:

$$\frac{1}{\mu} \nabla.(\mathbf{k} \nabla p) = \frac{\alpha}{1-\phi} \left(\frac{D\sigma}{Dt} - \frac{Dp}{Dt} \right). \tag{3.14}$$

ofwel:

$$\frac{1}{\mu}\nabla.(\mathbf{k}\nabla p) = \frac{\alpha}{1-\phi}\left(\frac{\partial \sigma}{\partial t} - \frac{\partial p}{\partial t} - \mathbf{u_s}\nabla p - \mathbf{u_s}\nabla \sigma\right). \tag{3.15}$$

Omdat de snelheid $\mathbf{u_s}$ van de rots in de grond zeer laag is, kunnen we de termen met $\mathbf{u_s}$ in de vergelijking (3.15) verwaarlozen. Daarna wordt de volgende tijdsafhankelijke diffusievergelijking (PDE) verkregen:

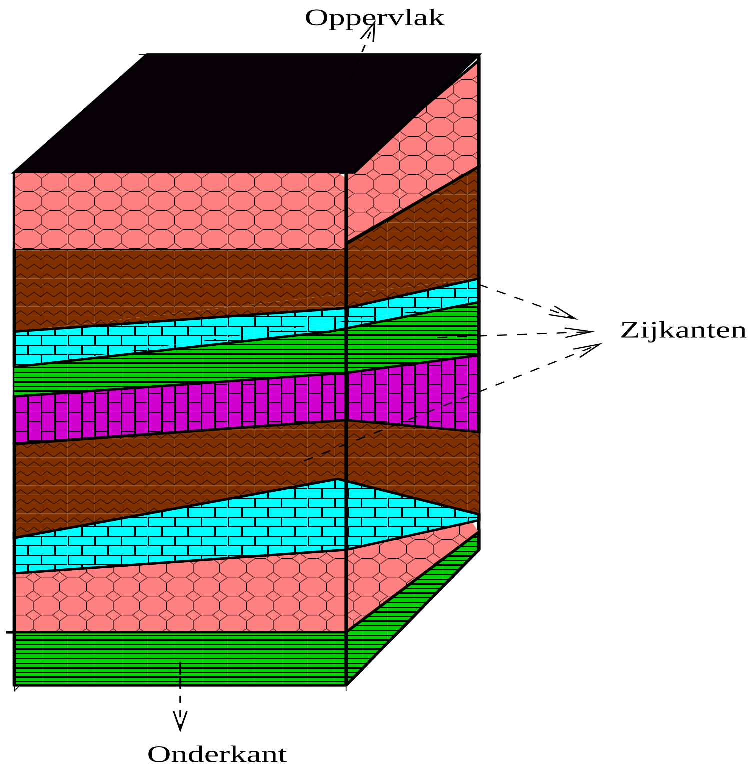
$$\frac{1}{\mu}\nabla.(\mathbf{k}\nabla p) = \frac{\alpha}{1-\phi}(\frac{\partial \sigma}{\partial t} - \frac{\partial p}{\partial t}). \tag{3.16}$$

Hierbij wordt aangenomen dat σ bekend is.

Hoofdstuk 4

Mathematische beschrijving

4.1 Randvoorwaaarden van het probleem



Figuur 4.1: Schets van een eindig gebied in de ondergrond

Om de diffusievergelijking (3.16) eenduidig op te kunnen lossen is het noodzakelijk zowel begin- als randvoorwaarden te geven. M.b.t. de beginvoorwaarden nemen we aan dat de druk bekend is op tijdstip $t = 0$.
Eerst beschouwen we een eindig gebied Ω (Zie figuur 4.1) in de ondergrond. De randen van het onderzochte gebied worden verdeeld in een natuurlijke rand (de bovenkant van de bovenste laag) en kunstmatige randen (de zijkanten en de onderkant van de onderste laag). Met natuurlijke rand wordt bedoeld de rand waarop de randvoorwaarden uit de fysica opgelegd kunnen worden.

We zullen nu de randvoorwaarden op de verschillende randen afleiden.

- Op de rand aan het oppervlak geldt dat de overdruk gelijk is aan de atmosferische druk, d.w.z

$$p = p_o \quad \text{op de bovenkant.}$$

  Dit is een inhomogene Dirichlet randvoorwaarde.

- De onderste rand wordt zo gekozen dat hij met een ondoordringbare laag samenvalt. Hierdoor is geen stroming mogelijk. Dit impliceert een homogene Neumann randvoorwaarde:

$$\frac{\partial p}{\partial n} = 0 \quad \text{op de onderkant.}$$

- De keuze van de randvoorwaarde op de rand aan zijkanten is niet meer uit de fysica te halen. Men lost dit op, door de rand zover te leggen van het fysisch interessante gebied, dat de invloed van de randvoorwaarde verwaarloosbaar is op de oplossing in het gebied van interesse. In dit geval kiezen we aan de rand een homogene Neumann randvoorwaarde:

$$\frac{\partial p}{\partial n} = 0.$$

4.2 Discretisatie van het probleem

Om de PDE numeriek op te lossen zullen we als discretisatiemethode de eindige elementen methode hanteren. Hiertoe zijn twee redenen. De eerste reden is dat in de werkelijkheid de struktuur van aardlagen zeer complex is. Als gevolg hiervan zijn we bij ruimtediscretisatie niet in staat met uniforme rekenroosters te werken. De tweede reden wordt gevormd door de discontinuiteiten die zich in ons diffusie probleem voordoen. Met name zijn de permeabiliteit en de porositeit sterk afhankelijk aan het type laag.

De zwakke formulering is een handig uitgangspunt voor de e.e.m. Om de zwakke formulering af te leiden, vermenigvuldigen we vergelijking (3.16) met een zogenaamde testfunctie ν in de oplossingruimte en integreren we het geheel over het gebied Ω met rand Γ. Dit heeft als resultaat dat voor iedere ν geldt na toepassing van stelling van Gauss:

$$\int_\Gamma \nu \frac{\mathbf{k}(x,p)}{\mu} \nabla p.\,\mathbf{n} d\Gamma - \int_\Omega \frac{\mathbf{k}(x,p)}{\mu} \nabla p \nabla \nu d\Omega = \int_\Omega \frac{\alpha \nu}{1-\phi(x,p)} (\frac{\partial \sigma}{\partial t} - \frac{\partial p}{\partial t}) d\Omega. \quad (4.1)$$

Door de gestelde randvoorwarden in de vorig paragraaf valt de randintegraal weg, d.w.z

$$\int_\Gamma \nu \frac{\mathbf{k}(x,p)}{\mu} \nabla p.\,\mathbf{n} d\Gamma = 0$$

Dat heeft als gevolg:

$$-\int_{\Omega} \frac{\mathbf{k}(x,p)}{\mu} \nabla p \nabla \nu d\Omega = \int_{\Omega} \frac{\alpha\nu}{1-\phi(x,p)} (\frac{\partial\sigma}{\partial t} - \frac{\partial p}{\partial t}) d\Omega. \qquad (4.2)$$

M.b.v de methode van Galerkin wordt de oplossing p benaderd door een lineaire combinatie van basisfuncties:

$$\bar{p} = \sum_{j=1}^{n} p_j(t)\varphi_j(\mathbf{x}) + p_D \text{ met } p_D = \sum_{j=n+1}^{N} p_j(t)\varphi_j(\mathbf{x}).$$

Hierbij vormen de functies φ_j een basis voor de oplossingsruimte. p_D dient aan de Dirichlet randvoorwaarden te voldoen. De parameters p_j $(j = 1, .., n)$ zijn onbekend en moeten nog bepaald worden.
De voor de hand liggende keuze voor ν in de oplossingruimte is :

$$\nu = \phi_i \text{ voor } i = 1, ..., n.$$

Door substitutie van de benadering $\bar{p}$ in vergelijking (4.2) komen we tot een stelsel van n tijdsafhankelijke vergelijkingen met n onbekenden:

$$\sum_{j=1}^{N} \Big(\frac{dp_j}{dt} \int_{\Omega} \frac{\alpha}{1-\phi} \varphi_i \varphi_j d\Omega - p_j \int_{\Omega} \frac{\mathbf{k}}{\mu} \nabla\varphi_j . \nabla\varphi_i d\Omega \Big) = \int_{\Omega} \frac{\alpha}{1-\phi} \frac{\partial\sigma}{\partial t} \varphi_i d\Omega \text{ voor } i = 1, ..., n \qquad (4.3)$$

Als we de matrices $\mathbf{M}$,$\mathbf{S}$ en respectievelijk de vektor $\mathbf{f}$ als volgt definieren:

$$m_{ij} = \int_{\Omega} \frac{\alpha}{1-\phi} \varphi_i \varphi_j d\Omega.$$

$$s_{ij} = -\int_{\Omega} \frac{\mathbf{k}}{\mu} \nabla\varphi_j . \nabla\varphi_i d\Omega,$$

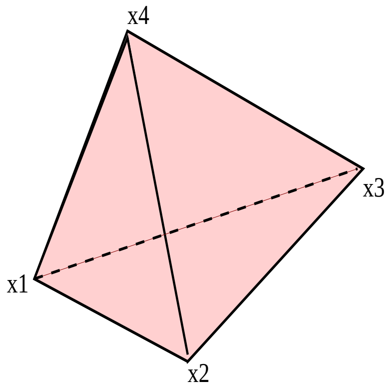
$$f_i = \int_{\Omega} \frac{\alpha}{1-\phi} \frac{\partial\sigma}{\partial t} \varphi_i d\Omega$$

dan wordt vergelijking (4.3) als volgt geschreven:

$$\sum_{j=1}^{n} m_{ij} \frac{dp_j}{dt} = \sum_{j=1}^{n} s_{ij} p_j + f_i \text{ voor } i = 1, ..., n$$

Het is gemakkelijk in te zien dat de matrixelementen van $\mathbf{M}, \mathbf{S}$ afhankelijk van de overdruk p zijn. Dat geldt ook voor de componenten van de vektor $\mathbf{f}$. Immers de porositeit ϕ en permeabiliteit $\mathbf{k}$ hangen van de overdruk p af.
In matrix-vektor notatie luidt het bovengenoemde stelsel van vergelijkingen als volgt:

$$\mathbf{M}(p) \frac{d\mathbf{p}}{dt} = \mathbf{S}(p)\mathbf{p} + \mathbf{f}(p). \qquad (4.4)$$

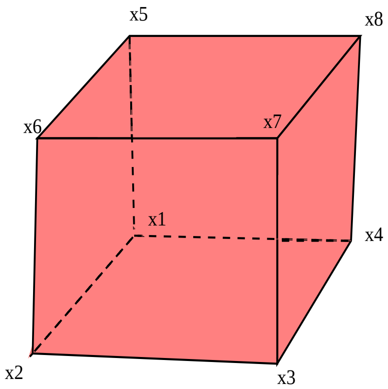


Figuur 4.2: tetraheder

Met $\mathbf{p}$ wordt de discrete benadering van p aangegeven.

Nu gaan we e.e.m. toepassen door het gebied in een aantal elementen te verdelen. Als voorbeeld beschouwen we twee soorten van lineaire elementen in $\mathcal{R}^3$. De ene met vier hoekpunten (tetraheders) (zie figuur 4.2) en de andere met acht hoekpunten (hexaheders)(zie figuur 4.3). Daarna kiezen we de bijbehorende basisfuncties φ_i die aan de volgende eisen voldoen:

- de basisfuncties φ_i zijn lineair in het geval van tetraheders en trilineair in het geval van hexaheders.
- $\varphi_i(x^j) = \delta_{ij}$



Figuur 4.3: Hexaheder

Het verkregen stelsel moet nog naar de tijd geintegreerd worden. Op het niet lineaire stelsel 4.4 gaan we het θ-schema toepassen [1]:

$$\mathbf{M}(\mathbf{p^a})\frac{(\mathbf{p}^{n+1} - \mathbf{p}^n)}{\Delta t} = \theta\mathbf{S}(\mathbf{p^a})\mathbf{p^{n+1}} + (1-\theta)\mathbf{S}(\mathbf{p^a})\mathbf{p^n} + ((1-\theta)\mathbf{f^n}(\mathbf{p^a}) + \theta\mathbf{f^{n+1}}(\mathbf{p^a})).$$

$$\text{met} \quad \mathbf{p^a} = \theta\mathbf{p^{n+1}} + (1-\theta)\mathbf{p^n} \quad \text{voor} \quad 0 \le \theta \le 1.$$

Voor verschillende waarden van θ krijgen we verschillende verschillende numerieke schema's.
De keuze $\theta = 0$ leidt tot de expliciete methode van Euler. Dit schema is echter alleen stabiel voor voldoende klein tijdstappen: $\Delta t \leq \frac{2}{\lambda_{max}}$.
In dit geval wordt het θ-schema gereduceerd tot het volgende schema:

$$\mathbf{M}(\mathbf{p^n})\mathbf{p^{n+1}} = \mathbf{M}(\mathbf{p^n})\mathbf{p^n} + \Delta t(\mathbf{S}(\mathbf{p^n})\,\mathbf{p^n} + \mathbf{f^n}(\mathbf{p^n})). \tag{4.5}$$

De keuze $\theta = 1/2$ of $\theta = 1$ leidt tot impliciete methodes. Deze zijn in de praktijk het meest aantrekkelijk. Hoewel het schema impliciet is en dus tijdens elke tijdstap een groot niet-lineair stelsel moet worden opgelost, kan de tijdstap veel groter gekozen worden want de bijbehorende schema's zijn onvoorwaardelijk stabiel.
Voor $\theta = 1$ wordt het achterwaartse Euler schema verkregen:

$$\Big(\mathbf{M}(\mathbf{p^{n+1}}) - \Delta t\mathbf{S}(\mathbf{p^{n+1}})\Big)\mathbf{p^{n+1}} = \mathbf{M}(\mathbf{p^{n+1}})\mathbf{p^n} + \Delta t\mathbf{f^{n+1}}(\mathbf{p^{n+1}})$$

Voor $\theta = \frac{1}{2}$ vinden we het Crank-Nicolson schema:

$$\Big(\mathbf{M}(\mathbf{p^a}) - \frac{\Delta t}{2}\mathbf{S}(\mathbf{p^a})\Big)\mathbf{p^{n+1}} = \Big(\mathbf{M}(\mathbf{p^a}) + \frac{\Delta t}{2}\mathbf{S}(\mathbf{p^a})\Big)\mathbf{p^n} + \frac{\Delta t}{2}(\mathbf{f^n}(\mathbf{p^a}) + \mathbf{f^{n+1}}(\mathbf{p^a})),$$

$$\text{met}\ \ \mathbf{p^a} = \frac{\mathbf{p^{n+1}} + \mathbf{p^n}}{2}.$$

Omdat de bovengenoemde schema's niet-linear zijn, gaan we de Picard iteratiemethode toepassen. Voor de eenvoud zullen we ons beperken tot één iteratieslag. Als schatting voor $\mathbf{p^{n+1}}$ kiezen we $\mathbf{p^n}$. Hieruit volgt dat we $\mathbf{p^n}$ als schatting voor $\mathbf{p^a}$ kunnen gebruiken.
Dit leidt tot de volgende lineaire schema's:
Voor $\theta = 1$:

$$\Big(\mathbf{M}(\mathbf{p^n}) - \Delta t\mathbf{S}(\mathbf{p^n})\Big)\mathbf{p^{n+1}} = \mathbf{M}(\mathbf{p^n})\mathbf{p^n} + \Delta t\,\mathbf{f^{n+1}}(\mathbf{p^n}). \tag{4.6}$$

Voor $\theta = \frac{1}{2}$:

$$\Big(\mathbf{M}(\mathbf{p^n}) - \frac{\Delta t}{2}\mathbf{S}(\mathbf{p^n})\Big)\mathbf{p^{n+1}} = \Big(\mathbf{M}(\mathbf{p^n}) + \frac{\Delta t}{2}\mathbf{S}(\mathbf{p^n})\Big)\mathbf{p^n} + \frac{\Delta t}{2}\,(\mathbf{f^n}(\mathbf{p^n}) + \mathbf{f^{n+1}}(\mathbf{p^n})). \tag{4.7}$$

De vergelijking (4.5), (4.6) of (4.7) kan wordt teruggebracht tot een lineair stelsel vergelijkingen van de vorm:

$$\mathbf{A}^n\mathbf{p}^{n+1} = \mathbf{b}^n. \tag{4.8}$$

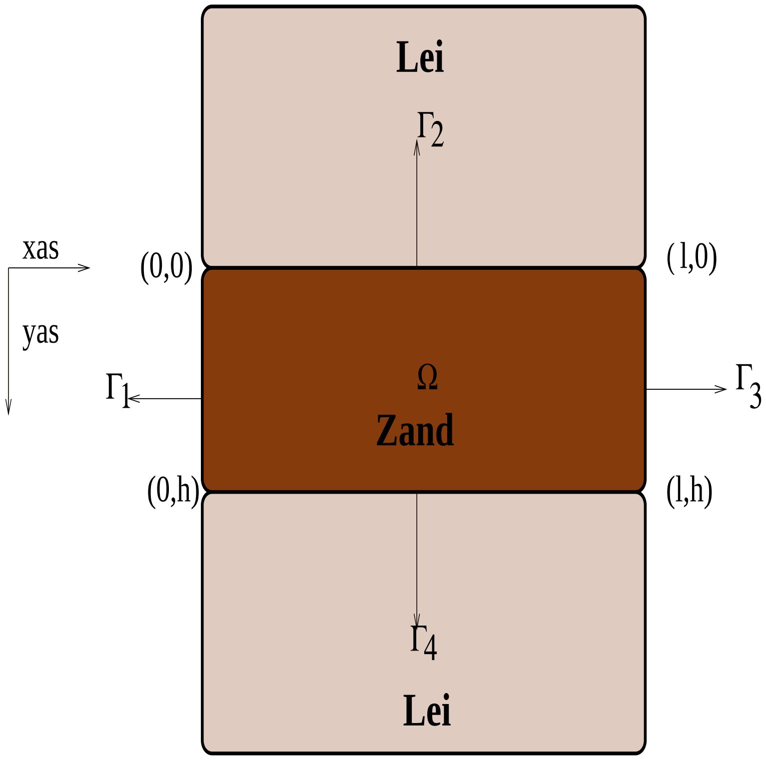
Hierin geldt dat de matrix $\mathbf{A^n}$ en de vektor $\mathbf{b^n}$ alleen van $\mathbf{p^n}$ afhankelijk zijn.
Om de numerieke analyse van de matrix $\mathbf{A^n}$ te vereenvoudigen vervangen we het tijdsafhankelijke probleem door het stationaire diffusieprobleem. Gemakshalve beperken we ons tot lineaire diffusievergelijkingen van de volgende vorm:

$$-\nabla.(\sigma\nabla p) = 0, \tag{4.9}$$

$$\text{met: } p = 0 \text{ op } \Gamma_D,$$
$$\text{en: } \frac{\partial p}{\partial n} = 0 \text{ op } \Gamma_N,$$

Hierin is σ een constante die van de eigenschap van het medium afhankelijk is $\sigma(=\frac{k}{\mu})$. Later zal σ ook per laag gaan varieren. Na discretisatie van (4.9) ontstaat het stelsel vergelijkingen:

$$\mathbf{Sp} = \mathbf{0},$$
$$\text{met: } s_{ij} = -\int_\Omega \sigma \nabla\varphi_j . \nabla\varphi_i d\Omega.$$



Figuur 4.4: Een schets van een Zand-laag

Kleine eigenwaarden kunnen zorgen voor een slechte conditie van het op te lossen stelsel vergelijkingen en dus voor een slecht convergentiegedrag voor iteratieve methoden. In dit geval wordt het lastig om het stelsel vergelijkingen op te lossen. Daarom is het handig alleerst om enig idee te krijgen over het bestaan van kleine eigenwaarden.
Omdat de permeabiliteit van ondergrondlagen in de praktijk onderling zeer sterk verschillend kan zijn, kan dit leiden tot een slechte verdeling van de eigenwaarden, d.w.z. de eigenwaarden liggen zeer ver van elkaar. Aan de hand van de fysica is het mogelijk het bestaan van de zeer kleine eigenwaarden te voorspellen.
Voor een eenvoudige analyse bekijken we de diffusie vergelijking (3.16), maar slechts de lokaal in een zandlaag die tussen twee lei-lagen ligt. Op basis van

de fysica voldoet de druk aan de volgende interface conditie op de onder-en bovenkant:

$$\frac{\partial p}{\partial y} = \epsilon \frac{\partial p_b}{\partial y} \text{ op } \Gamma_2.$$

$$\frac{\partial p}{\partial y} = \epsilon \frac{\partial p_o}{\partial y} \text{ op } \Gamma_4.$$

p_b geeft de overdruk in de bovendliggende leilaag aan . p_o geeft de overdruk in de onderliggende lei-laag aan. Op de zijkanten kiezen we de homogene Neumann randvoorwaarden:

$$\frac{\partial p}{\partial x} = 0 \text{ op } \Gamma_1, \Gamma_3.$$

De diffusievergelijking op een zandlaag kan als geformuleerd worden:

$$-\nabla.(\nabla p) = 0 \text{ in } \Omega. \tag{4.10}$$

$$\left.\frac{\partial p}{\partial x}\right|_{x=0} = 0, \; \left.\frac{\partial p}{\partial x}\right|_{x=l} = 0$$

$$\left.\frac{\partial p}{\partial y}\right|_{y=0} = \epsilon f_1(x), \; \left.\frac{\partial p}{\partial y}\right|_{y=h} = \epsilon f_2(x).$$

Vanwege het feit dat ϵ klein is, is dit nagenoeg Neumannprobleem met homogene randvoorwaarden. De oplossing van dit probleem is op een constante na bepaald. Daarom heeft een Neumann probleem één eigenwaarde 0.
In ons geval geldt dat de randvoorwaarden op Γ_2 en Γ_4 heel klein zijn ($O(\epsilon)$) en wij verwachten daarom ook een kleine eigenwaarde van die grootte.
Alvorens de invloed ϵ op de eigenwaarden van de matrix **S** duidelijk te maken bekijken eerst we het bijbehorende continue eigenwaardenprobleem:

$$\nabla.(\nabla p) + \lambda p = 0. \tag{4.11}$$

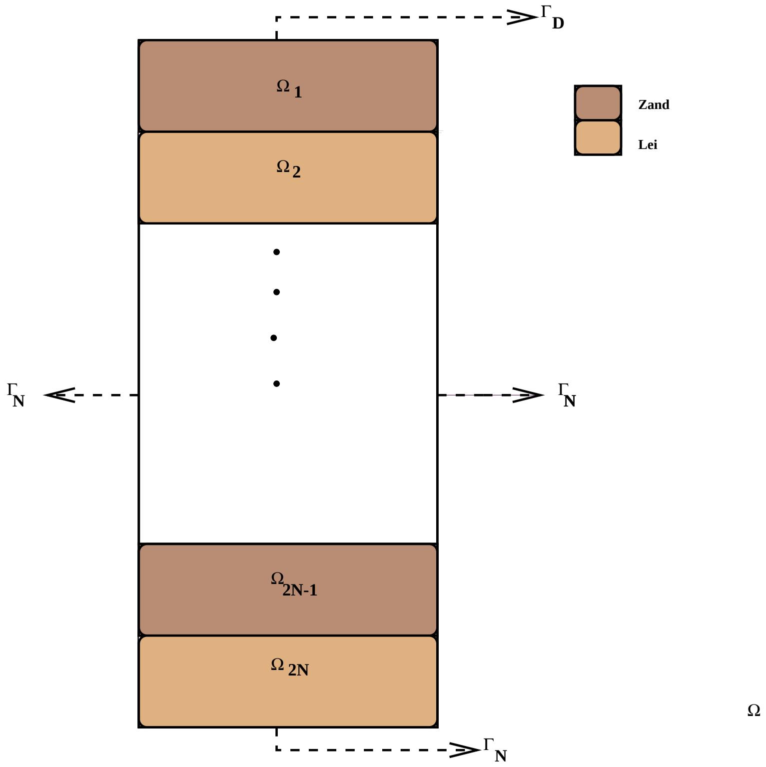
(4.11) bepalen (zie appendix A.1 in [10]), nl: In [10](appendix A.1) is gebleken dat het vergelijkbaar continue eigenwaardeprobleem (4.11) met speciale randvoorwaarden één nuleigenwaarde heeft.
Aan de hand een eenvoudig twee-dimensionaal voorbeeld zullen we de struktuur van de matrix **S** laten zien. Verder wordt gecontroleerd of er in het geval van niet homogene randvoorwaarden kleine eigenwaarden van de matrix zijn en hoeveel.

4.3 Voorbeeld

Beschouw een gebied in de ondergrond met $2N$ afwisselene lagen (lei resp.zand) (zie figuur 4.5).
In dit gebied wordt de bovengenoemde diffusievergelijking :

$$-\nabla.(\sigma_k \nabla p) = 0 \text{ op } \Omega_k (k = 1, ..., 2N)$$

$$\text{met: } p = 0 \text{ op } \Gamma_D,$$



Figuur 4.5: Eenvoudige structuur van ondergrond

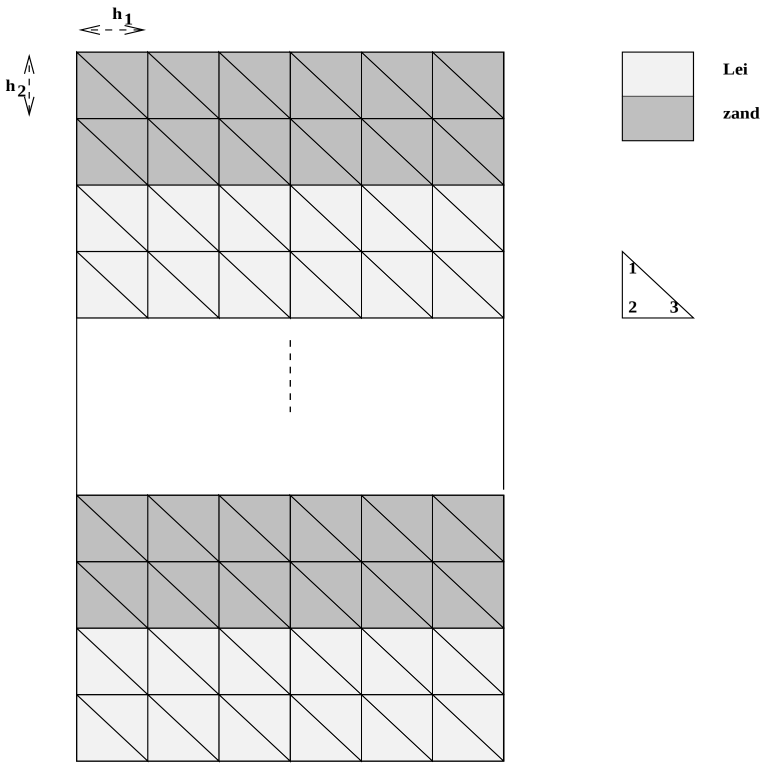
$$\text{en: } \frac{\partial p}{\partial n} = 0 \text{ op } \Gamma_N.$$

Voor zandlagen nemen we $\sigma = 1$ en voor leilagen $\sigma = \epsilon$, m.a.w:

$$\sigma_k = \begin{cases} 1 & \text{als k oneven is (hoog permeabel).} \\ \epsilon & \text{als k even is (laag permeabel).} \end{cases}$$

4.3.1 Structuur van de matrix S

Het gebied wordt verdeeld in een aantal eindige driehoekselementen e_i $(i = 1...M)$(zie figuur 4.6). Bij de Galerkin formulering zullen we de lineaire interpolatie functies voor een driehoek als basisfuncties gebruiken. Om eenvoudig de globale matrix $\mathbf{S}$ door de computer te laten construeren, berekenen we deelmatrices $\mathbf{S^i}$ voor een eindig element e_i. Door optellen van deze matrices op de juiste manier ontstaat $\mathbf{S}$. Dit proces wordt ook wel assemblage genoemd. Voor een eerste analyse beperken we ons tot het geval van rechte lagen waarin rechthoekige driehoeken gekozen worden.
Om de structuur van de matrix $\mathbf{S}$ te laten zien, hebben we de tridiagonale matrices D_α en diagonale matrices J_α nodig . Voor ieder reeel getal α zijn



Figuur 4.6: Een mogelijke verdeling van het ondergrondgebied in driehoekse-lementen

de matrices D_α en J_α als volgt gedefinieerd:

$$D_\alpha = \alpha \begin{pmatrix} a+b & -a & & & \\ -a & 2(a+b) & -a & & \\ & -a & 2(a+b) & -a & \\ & & \ddots & \ddots & \ddots \\ & & & -a & a+b \end{pmatrix},$$

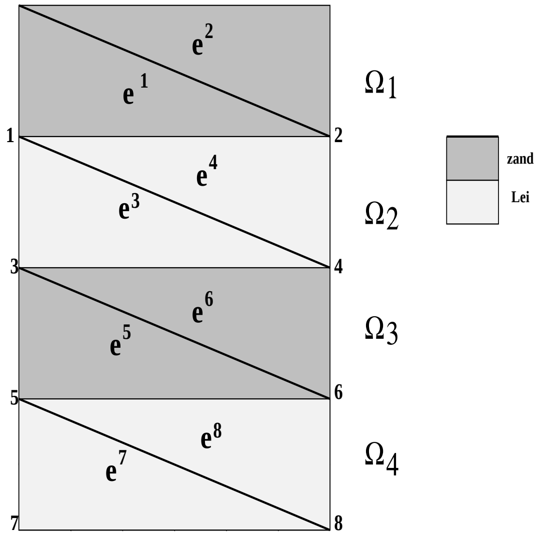
$$\text{en } \; J_\alpha = \alpha \begin{pmatrix} -b & & & & \\ & -2b & & & \\ & & -2b & & \\ & & & \ddots & \\ & & & & -b \end{pmatrix}.$$

Hier geldt: $a = \frac{h_1}{2h_2}$en $b = \frac{h_2}{2h_1}$.Voor een definitie van h_1 en h_2 zie figuur 4.6. In het geval van vier grondlagen (Z-L-Z-L) met een rij elementen per laag is gebleken uit appendix A.2 in [10] dat de matix $\mathbf{S}$ de volgende structuur heeft:

$$\mathbf{S_0} = \begin{pmatrix} D_{1+\epsilon} & J_\epsilon & 0 & 0 \\ J_\epsilon & D_{1+\epsilon} & J_1 & 0 \\ 0 & J_1 & D_{\epsilon+1} & J_\epsilon \\ 0 & 0 & J_\epsilon & D_\epsilon \end{pmatrix}.$$

Hierin worden de D_α en J_α gegeven door:

$$D_\alpha = \alpha\begin{pmatrix} a+b & -a \\ -a & a+b \end{pmatrix} \text{ en } J_\alpha = \alpha\begin{pmatrix} -b & 0 \\ 0 & -b \end{pmatrix}.$$



Figuur 4.7: Een verdeling van een gebied met 2N afwisselende lagen

Op analoge wijze kunnen we voor een gebied met 2N afwisselende zand-lei lagen met een rij elementen per laag (zie figuur 4.7) laten zien dat de matrix de volgende speciale structuur (bloktridiagonale structuur) heeft:

$$\mathbf{S_1} = \begin{pmatrix} D_{1+\epsilon} & J_\epsilon & & & & & & \\ J_\epsilon & D_{1+\epsilon} & J_1 & & & & & \\ & J_1 & D_{1+\epsilon} & J_\epsilon & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & J_\epsilon & D_{1+\epsilon} & J_\epsilon & \\ & & & & & J_1 & D_{\epsilon+1} & J_\epsilon \\ & & & & & & J_\epsilon & D_\epsilon \end{pmatrix}.$$

Indien we een horizontale nummering kiezen en lineaire driehoeken, heeft de matrix $\mathbf{S}$ in het algemene geval de volgende gedaante:

$$\mathbf{S} = \begin{pmatrix} D_2 & J_1 \\ J_1 & D_2 & J_1 \\ & \ddots & \ddots & \ddots \\ & & J_1 & D_2 & J_1 \\ & & & J_1 & D_{1+\epsilon} & J_\epsilon \\ & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon \\ & & & & & & & J_\epsilon & D_{\epsilon+1} & J_1 \\ & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & & & J_1 & D_{1+\epsilon} & J_\epsilon \\ & & & & & & & & & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon \\ & & & & & & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & & & & & & J_\epsilon & D_\epsilon \end{pmatrix}.$$

De matrix $\mathbf{S}$ kan anders worden geformuleerd:

$$\mathbf{S} = \begin{pmatrix} Z_0 \\ R_1 \\ L \\ R_2 \\ Z \\ \vdots \\ \vdots \\ R_2 \\ L \\ R_N \end{pmatrix}.$$

Hierbij is:

$$\mathbf{Z_0} = \begin{pmatrix} D_2 & J_1 & 0 & & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & J_1 & D_2 & J_1 & 0 & \cdots & 0 \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} J_1 & D_2 & J_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & J_1 & D_2 & J_1 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{L} = \begin{pmatrix} J_\epsilon & D_{2\epsilon} & J_\epsilon & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & J_\epsilon & D_{2\epsilon} & J_\epsilon & 0 & \cdots & 0 \end{pmatrix}, \mathbf{R_1} = \begin{pmatrix} 0 & \cdots & J_1 & D_{1+\epsilon} & J_\epsilon & 0 & \cdots & 0 \end{pmatrix}$$

$\mathbf{R_2} = \begin{pmatrix} 0 & \cdots & J_\epsilon & D_{1+\epsilon} & J_1 & 0 & \cdots & 0 \end{pmatrix}$ en $\mathbf{R_N} = \begin{pmatrix} 0 & \cdots & \cdots & 0 & J_\epsilon & D_\epsilon \end{pmatrix}$.

We kunnen gmakkelijk zien dat :

- de matrices **Z** en **L** corresponderen respectievelijk met de matrices die bij de interne knooppunten in een zandlaag en die bij een leilaag horen.

- de matrices $\mathbf{R_1}$ en $\mathbf{R_2}$ representeren de overgang van leilaag naar zandlaag respectievelijk van zandlaag naar leilaag.

- $\mathbf{R_0}$ en $\mathbf{R_N}$ zijn de matrices die bij de punten van de bovenrand en respectievelijk de onderrand behoren.

Omdat de matrix D_α symmetrisch is en J_α diagonaal is , kunnen we gemakkelijk controleren dat de matrix **S** ook een symmetrische matrix is.

4.3.2 Kleine eigenwaarden van de matrix S

Uit de fysica blijkt dat een zandlaag ingesloten door twee leilagen aanleiding geven tot één kleine eigenwaarde van de orde ϵ. Nu gaan we kijken of we dit resultaat terugvinden in het discrete probleem door de eigenwaarden van de matrix **S** te bepalen.
De eigenwaarden λ_i van de matrix $\mathbf{S_0}$ zijn zodanig dat:

$$|\mathbf{S_0} - \lambda_i I| = 0 \text{ met } I \text{ is de eenheidsmatrix}$$

Omdat we de eigenwaarden van S_0 niet eenvoudig te bepalen zijn voeren we een simpeler matrix $\mathbf{\bar{S}_0}$ in;

$$\mathbf{\bar{S}_0} = \begin{pmatrix} D_{1+\epsilon} & 0 & 0 & 0 \\ 0 & D_{1+\epsilon} & J_1 & 0 \\ 0 & J_1 & D_{\epsilon+1} & 0 \\ 0 & & & D_\epsilon \end{pmatrix}.$$

We kunnen laten zien (zie Appendix A.3 in [10]) dat een eigenwaarde λ_i van $\mathbf{S_0}$ benaderd kan worden door een eigenwaarde $\bar{\lambda}_i$ van $\mathbf{\bar{S}_0}$. Bovendien geldt $\lambda_i - \bar{\lambda}_i = O(\epsilon^p)$ met $1 \leq p$.
Een direkt gevolg hiervan is dat een kleine eigenwaarde van $\mathbf{\bar{S}_0}$ ook een kleine eigenwaarde van $\mathbf{S_0}$ is en omgekeerd.
De matrix $\bar{S}$ bestaat uit één blokmatrix $\mathbf{B}_\epsilon$ en één de blokmatrix D_ϵ , waarbij B_ϵ gegeven wordt door :

$$B_\epsilon = \begin{pmatrix} D_{1+\epsilon} & J_1 \\ J_1 & D_{\epsilon+1} \end{pmatrix}.$$

We kunnen makkelijk zien dat de matrix B_ϵ de zandlaag ingesloten door twee leilagen representeert. Bovendien levert B_ϵ één kleine eigenwaarde van de orde ϵ ,nl $\lambda_1 = \epsilon b$. Dit is in overeensteming met het fysische resultaat.
Voor een structuur met $2N$ afwisselende lagen gaan we nu het resultaat bevestigen dat iedere zandlaag ingesloten door twee leilagen kan leiden tot één

kleine eigenwaarde . Op analoge wijze geldt dat de eigenwaarden van de bijbehorende matrix $\mathbf{S}$

$$\mathbf{S_1} = \begin{pmatrix} D_{1+\epsilon} & J_\epsilon & & & & & & \\ J_\epsilon & D_{1+\epsilon} & J_1 & & & & & \\ & J_1 & D_{1+\epsilon} & J_\epsilon & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & J_\epsilon & D_{1+\epsilon} & J_\epsilon & \\ & & & & & J_1 & D_{\epsilon+1} & J_\epsilon \\ & & & & & & J_\epsilon & D_\epsilon \end{pmatrix}.$$

benaderd worden door de eigenwaarden van de volgende matrix $\bar{\mathbf{S}}_\mathbf{1}$:

$$\bar{\mathbf{S}}_\mathbf{1} = \begin{pmatrix} D_{1+\epsilon} & & & & \\ & B_\epsilon & & & \\ & & B_\epsilon & & \\ & & & \ddots & \\ & & & & D_\epsilon \end{pmatrix}.$$

Uit de structuur van de matrix $\bar{\mathbf{S}}_\mathbf{1}$ blijkt dat $\bar{\mathbf{S}}_\mathbf{1}$ een blokdiagonaal matrix is. Bovendien bestaat de matrix $\bar{\mathbf{S}}_\mathbf{1}$ o.a. uit $N-1$ blokmatrices $\mathbf{B}_\epsilon$. Iedere blokmatrix $\mathbf{B}_\epsilon$ komt overeen met een zandlaag die tussen leilagen ligt.

Nu gaan we van dit resultaat gebruikmaken om het aantal eigenwaarden te bepalen.
In het algemeen kunnen de eigenwaarden van de algemene matrix $\mathbf{S}$ benaderd worden door de eigenwaarden van de volgende blokdiagonaalmatrix :

$$\bar{\mathbf{S}} = \begin{pmatrix} B & & & & & & & & & & & \\ & B_\epsilon & & & & & & & & & & \\ & & D_{2\epsilon} & & & & & & & & & \\ & & & \ddots & & & & & & & & \\ & & & & D_{2\epsilon} & & & & & & & \\ & & & & & B_\epsilon & & & & & & \\ & & & & & & \ddots & & & & & \\ & & & & & & & \ddots & & & & \\ & & & & & & & & B_\epsilon & & & \\ & & & & & & & & & D_{2\epsilon} & & \\ & & & & & & & & & & \ddots & \\ & & & & & & & & & & & D_\epsilon \end{pmatrix}.$$

Hierin is de blokmatrix B_ϵ gegeven door:

$$B_\epsilon = \begin{pmatrix} D_{1+\epsilon} & J_1 & & & \\ J_1 & D_2 & J_1 & & \\ & \ddots & \ddots & \ddots & \\ & & J_1 & D_2 & J_1 \\ & & & J_1 & D_{1+\epsilon} \end{pmatrix}.$$

Analoog bestaat de matrix $\bar{S}$ uit $N-1$ blokmatrices B_ϵ. Uit fysische overwegingen is gebleken dat B_ϵ é én kleine eigenwaarde heeft. Controle met matlab voor een aantal gevallen heeft dit bevestigd, doch een algemeen bewijs hebben we niet kunnen geven.
We kunnen ook opmerken dat het totaal aantal kleine eigenwaarden die van de matrices $D_{2\epsilon}$ en van de matrix D_ϵ afkomstig zijn, precies gelijk is aan het totaal aantal knooppunten binnen leilagen die niet op de randen van de zandlagen liggen.
Uit de bovengenoemde resultaten kunnen we concluderen dat het aantal kleine eigenwaarden van de matrix $\mathbf{S}$ gelijk is aan het totaal aantal knooppunten in leilagen die niet op de randen van zandlagen liggen plus het aantal overgangen van zandlaag zonder Dirichlet randvoorwaarden naar leilaag.
Met behulp van matlab programma's (zie appendix A.3 in [10]) kunnen we nagaan voor een aantal verfijningen in de horizontale richting en in de vertikale richting dat het totaal aantal kleine eigenwaardeden van de matrix $\mathbf{S}$ gelijk is aan het totaal aantal knooppunten in leilagen die niet op de randen van zandlagen liggen plus het aantal overgangen van zandlaag zonder Dirichlet randvoorwaarden naar leilaag. Wij constateren dus dat behalve de fysisch verklaarbare kleine eigenwaarden, de discretizatie ook een nog groot aantal kleine eigenwaarden heeft ten gevolg van de leilagen. In het volgende hoofdstuk zullen we zien dat deze laatste eigenwaarden bij een geschikt gekozen oplossingtechniek geen probleem vormen, maar voor de eerste set van eigenwaarden een speciale aanpak vereist is.

Hoofdstuk 5

Oplosmethoden voor het gediscretiseerde stelsel

In de vorige paragraaf hebben we een stelsel lineaire vergelijkingen in de vorm $\mathbf{Ax} = \mathbf{b}$, dat ontstaat na discretisatie van de diffusie vergelijking (3.16) uitgeschreven. Voor de numerieke oplossing van $\mathbf{Ax} = \mathbf{b}$ zijn verschillende directe en iteratieve methoden beschikbaar.
Door de e.e.m als discretisatie te gebruiken kunnen we makkelijk zien dat de matrix $\mathbf{A}$ ijl is, hetgeen inhoudt dat er slechts enkele elementen per rij van A niet nul zijn. Bij de directe methoden is het uitbuiten van de ijlheid niet eenvoudig. Een andere eigenschap van de matrix A is dat hij in de praktijk groot is, dat wil zeggen dat een zeer groot aantal onbekenden in het gediscretiseerde stelsel voorkomen. Als gevolg daarvan nemen de directe methoden veel geheugenruimte in beslag en kosten veel rekenwerk. Rekening houdend met deze eigenschappen van de matrix A zijn iteratieve methoden te preferen bij een grote en ijle matrix. We zullen ons bij ons onderzoek beperken tot de gradiënten methoden. Bij deze methoden wordt de iterand in de (i+1)-de stap als volgt berekend uit die van de i- de stap:

$$\mathbf{x_{i+1}} = \mathbf{x_i} + \mathbf{P^{-1}}(\mathbf{b} - \mathbf{Ax_i}).$$

De grootheid $\mathbf{r_i} = \mathbf{b} - \mathbf{Ax_i}$ noemt men het residu. Met $\mathbf{r_0} = \mathbf{b} - \mathbf{Ax_0}$ gaat men met volledige inductie eenvoudig na dat na i slagen geldt: $\mathbf{x_i} = \mathbf{x_0} + \mathbf{y}$ met $\mathbf{y} \in span\{\mathbf{r_0}, \mathbf{Ar_0}, ..., \mathbf{A^{i-1}r_0}\} = K^i(\mathbf{A}, \mathbf{r_0})$.
De ruimte $K^i(\mathbf{A}, \mathbf{r_0})$ heet de Krylov-ruimte [6] met dimensie i, behorende bij $\mathbf{A}$ en $\mathbf{r_0}$.

5.1 De CG-methode

De geconjugeerde-gradiëntenmethode (meestal afgekort CG) [3] is een iteratieve methode voor het oplossen van $\mathbf{Ax} = \mathbf{b}$ waarbij A symmetrisch en positief definiet is. Men probeert de oplossing van het op te lossen stelsel te benaderen door $\mathbf{x_i}$ waarbij de afstand gemeten in de A-norm tussen de echte oplossing en een benadering hiervan zo klein mogelijk is. De algemene gedaante van de methode is dat bij iedere iteratie stap (i) een benadering van de oplossing $\mathbf{x_i}$, het residu $\mathbf{r_i}$ en de zoekrichting $\mathbf{p_i}$ bepaald worden door:

$$\mathbf{x_i} = \mathbf{x_{i-1}} + \alpha_i \mathbf{p_i}.$$

$$\mathbf{r_i} = \mathbf{r_{i-1}} - \alpha_i \mathbf{A p_i}.$$

$$\mathbf{p_{i+1}} = \mathbf{r_i} + \beta_{i+1} \mathbf{p_i}.$$

De coëfficiënt α_i wordt zodanig gekozen dat $\mathbf{r_i^T A^{-1} r_i}$ minimaal is.
Een geschikte keuze voor α_i is :

$$\alpha_i = \frac{(\mathbf{r_{i-1}}, \mathbf{r_{i-1}})}{(\mathbf{p_i}, \mathbf{A p_i})}.$$

De coëfficiënt β_i wordt zodanig gekozen dat $\mathbf{p_i}$ en $\mathbf{A p_{i-1}}$ orthogonaal zijn.
Een geschikte keuze voor β_i is :

$$\beta_i = \frac{(\mathbf{r_{i-1}}, \mathbf{r_{i-1}})}{(\mathbf{r_{i-2}}, \mathbf{r_{i-2}})}.$$

Een algoritmische beschrijving van de CG-methode [3] in *pseudocode* ziet er als volgt uit:

CG-algoritme
  $k = 0, \mathbf{x_0} = 0; \mathbf{r_0} = \mathbf{b}$
  while not $(r_k = 0)$ **do**
    $k = k + 1;$
    if $k = 1$ **then**
      $p_1 = r_0$
    else
      $\beta_k = \frac{(r_{k-1}, r_{k-1})}{(r_{k-2}, r_{k-2})}$
      $p_k = r_{k-1} + \beta_k p_{k-1}$
    endif
    $\alpha_k = \frac{(r_{k-1}, r_{k-1})}{(p_k, A p_k)}$
    $x_k = x_{k-1} + \alpha_k p_k$
    $r_k = r_{k-1} - \alpha_k A p_k$
  endwhile
endCG-algoritme

Als stopcriterium zullen we tevreden zijn met een vooraf vastgestelde precisie, m.a.w $||\mathbf{r_k}|| \leq \epsilon$, in plaats van $\mathbf{r_k} = 0$.
Na enige aanpassingen kan een variant van dit algoritme worden gegeven door:

CG-algoritme
$k = 0, x_0 = 0; r_0 = b; p_1 = r_0$
while $||r_0|| > \epsilon$ *do*
$k = k + 1;$
$\alpha_k = \frac{(r_{k-1}, r_{k-1})}{(p_k, Ap_k)}$
$x_k = x_{k-1} + \alpha_k p_k$
$r_k = r_{k-1} - \alpha_k Ap_k$
$\beta_k = \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})}$
$p_{k+1} = r_k + \beta_k p_k$
endwhile
endCG-algoritme

De convergentie van de geconjugeerde gradiënten methode (CG) wordt gegeven door een begrenzing van de relatieve fout. Stel dat $\mathbf{x}$ de exacte oplossing is van het lineaire stelsel $\mathbf{Ax} = \mathbf{b}$. Een bovengrens voor $||\mathbf{x} - \mathbf{x_k}||_A$ wordt gegeven door [6]

$$||\mathbf{x} - \mathbf{x_k}||_A \leq 2||\mathbf{x} - \mathbf{x_0}||_A \left(\frac{\sqrt{K_2(\mathbf{A})} - 1}{(\sqrt{K_2\mathbf{A}}) + 1} \right)^k .$$

Hierbij is $K_2(\mathbf{A})(= \frac{\lambda_{max}}{\lambda_{min}})$ het conditiegetal van de matrix $\mathbf{A}$. We zien uit deze ongelijkheid dat het CG-proces snel convergeren als $K_2(\mathbf{A})$ dichtbij 1 ligt. De bovengenoemde begrenzing van de relatieve fout kan in termen van de eigenwaarden uitgedrukt worden:

$$||\mathbf{x} - \mathbf{x_k}||_A \leq 2||\mathbf{x} - \mathbf{x_0}||_A \left(\frac{\sqrt{\lambda_{max}/\lambda_{min}} - 1}{\sqrt{\lambda_{max}/\lambda_{min}} + 1} \right)^k .$$

Hieruit kunnen we concluderen dat de CG methode toegepast op $\mathbf{Ax} = \mathbf{b}$ slecht zal convergeren als de kleinste eigenwaarde zeer klein is ten opzichte van de grootste eigenwaarde.
Uit resultaten van toepassing van CG methode op een testprobleem is gebleken dat het aantal benodigde iteraties om de convergentie te bereiken gridafhankelijk is, d.w.z afhankelijk van het aantal vertikale- en horizontale punten. Ook wordt opgemerkt dat het aantal benodigde iteraties veel groter is dan het aantal onbekenden. Dit slechte convergentiegedrag kan worden verklaard door het feit dat matrix A slecht geconditioneerd is, vanwege het grote aantal kleine eigenwaarden.

5.2 De ICCG-methode

We hebben in de vorige paragraaf (5.1) gezien dat het convergentiegedrag van de CG methode sterk is gerelateerd aan het conditiegetal van de matrix. Om de convergentie van het CG-proces te versnellen, maken we gebruik van preconditionering. Het idee van de preconditioneringsmethoden is dat het op te lossen stelsel $\mathbf{Ax} = \mathbf{b}$ tot een stelsel $\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{b}}$ getransformeerd wordt, waarbij

ervoor wordt gezorgd dat het conditiegetal van $\bar{\mathbf{A}}$ veel kleiner is dan het conditiegetal van $\mathbf{A}$. Hierna wordt op $\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{b}}$ de CG-methode toegepast. Het komt er eigenlijk op neer dat er wordt getracht het aanvankelijke conditiegetal $K_2(\mathbf{A})$ te verbeteren door de matrix $\mathbf{A}$ met een andere matrix $\mathbf{P^{-1}}$ te vermenigvuldigen zodanig dat: $K_2(\mathbf{P^{-1}A}) \leq K_2(\mathbf{A})$. $\mathbf{P}$ wordt de preconditioneringsmatrix genoemd. Uiteraard heeft dit alleen zin als $\mathbf{P^{-1}}$ goedkoop uit te voeren is en $\mathbf{P}$ eenvoudig kan worden berekend.
Voor CG is het noodzakelijk dat $\mathbf{P}$ symmetrisch en positief definiet is. Voor zo'n matrix bestaat een ontbinding $\mathbf{P} = \mathbf{CC}^T$, bijvoorbeeld de Choleski-ontbinding. We zoeken nu een oplossing van het stelsel

$$\mathbf{P^{-1}Ax} = \mathbf{P^{-1}b}$$

maar omdat $\mathbf{P^{-1}A}$ in het algemeen niet symmetrisch is vermenigvuldgen we links en rechts met $\mathbf{C^{-T}}$ en vinden

$$\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{b}}$$

Hierin geldt :$\bar{\mathbf{A}} = \mathbf{C^{-1}AC^{-T}}, \bar{\mathbf{x}} = \mathbf{C^{T}x}$ en $\bar{\mathbf{b}} = \mathbf{C^{-1}b}$.
Passen we hierop de CG-methode toe, dan geeft dit de gepreconditioneerde CG-methode (PCG).

PCG-algoritme
  $k = 0, x_0 = 0; r_0 = b;$
  Los op: $Pz_0 = r_0$
  $p_1 = z_0;$
  while$||r_k|| > \epsilon$ **do**
    $k = k + 1;$
    $\alpha_k = \frac{(r_{k-1}, z_{k-1})}{(p_k, Ap_k)}$
    $x_k = x_{k-1} + \alpha_k p_k$
    $r_k = r_{k-1} - \alpha_k Ap_k$
    Los op: $Pz_k = r_k$
    $\beta_k = \frac{(r_k, z_k)}{(r_{k-1}, z_{k-1})}$
    $p_{k+1} = z_k + \beta_k p_k$
  endwhile
endPCG-algoritme

De meest eenvoudige preconditionering is om de matrix $\mathbf{A}$ diagonaal te schalen. De diagonaalschaling houdt in dat we de matrix $\mathbf{A}$ door de matrix $\bar{\mathbf{A}} = \mathbf{C^{-1}AC^{-T}}$
met $\mathbf{C} = \mathbf{diag^{1/2}}(\mathbf{A})$ vervangen. Door deze keuze blijkt dat het conditiegetal van $\mathbf{A}$ verbeterd kan worden.
Uit resultaten van Matlab experimenten toegepast op ons testprobleem (zie appendix A.3 in [10]) is gebleken dat door de toepassing van de bovengenoemde preconditionering het aantal kleine eigenwaarden van de gepreconditioneerde matrix behoorlijk verminderd is. Het blijkt dat het aantal kleine eigenwaarden zich reduceert tot slechts het aantal lei-zand overgangen, dus is het aantal kleine eigenwaarden onafhankelijk van het grid geworden. In feite hebben we

slechts de al eerder fysisch gemotiveerde kleine eigenwaarden over. Voor een mathematisch bewijs van deze eigenschap verwijzen we naar [7].

Een veel gebruikte preconditionering is die van de incomplete Choleski decompositie. Hierbij wordt geprobeerd een matrix $\mathbf{P}$ te vinden zodat $K_2(\mathbf{P^{-1}A})$ in de buurt van 1 ligt, m.a.w $\mathbf{P^{-1}A} \approx \mathbf{I}$. Omdat A positief definiet symmetrisch is, kunnen we van A de Choleski decompositie $\mathbf{A} = \mathbf{LL^T}$ construeren. Ideaal is de keuze $\mathbf{P} = \mathbf{LL^T}$. Maar bij ons probleem is de matrix zeer ijl. Als gevolg hiervan kan $fill - in$ optreden, d.w.z het verschijnen van niet nul elementen op plaatsen waar $\mathbf{A}$ een nul element heeft. Voor het weglaten of minimaliseren van de fill-in gaan we incomplete Choleski decompositie toepassen. Men mag dus hopen dat via de incomplete decompositie een matrix $\mathbf{\bar{L}}$ zal worden geconstrueerd die op de echte $\mathbf{L}$ lijkt in de zin dat $\mathbf{\bar{L}\bar{L}^T} \approx \mathbf{A}$. Het bijbehorende PCG-algoritme wordt het ICCG-algoritme genoemd. In dit geval geldt $\mathbf{P} = \mathbf{\bar{L}\bar{L}^T}$. Het preconditioneerde stelsel wordt nu als volgt geschreven :

$$\mathbf{\bar{L}^{-1}A\bar{L}^{-T}y} = \mathbf{\bar{L}^{-1}b} \text{ met } \mathbf{x} = \mathbf{\bar{L}^{-T}y}$$

De beschrijving van dit algoritme [3] in *pseudocode* is de volgende:

ICCG-algoritme
$k = 0, x_0 = 0; r_0 = b; z_0 = \bar{L}^{-T}\bar{L}^{-1}r_0; p_1 = z_0$
while$||r_k|| > \epsilon$ **do**
$k = k + 1;$
$\alpha_k = \frac{(r_{k-1}, z_{k-1})}{(p_k, Ap_k)}$
$x_k = x_{k-1} + \alpha_k p_k$
$r_k = r_{k-1} - \alpha_k Ap_k$
$z_k = \bar{L}^{-T}\bar{L}^{-1}r_k$
$\beta_k = \frac{(r_k, z_k)}{(r_{k-1}, z_{k-1})}$
$p_{k+1} = z_k + \beta_k p_k$
endwhile
endICCG-algoritme

Door toepassing van ICCG methode blijkt dat het totaal benodigde iteraties om de convergentie te bereiken veel minder is dan bij die verkregen zonder preconditionering. Vergeleken met het aantal onbekenden is het echter nog steeds erg hoog. De reden is dat de kleine eigenwaarden die door het Neumann-achtige karakter van de diffusieprobleem in de zandlagen worden veroorzaakt, niet verdwijnen door de standaard preconditioneringstechnieken. Om toch de bijbehorende kleine eigenwaarden "kwijt" te raken gaan we over op een geheel nieuwe techniek [7]. Bij deze techniek (DICCG) wordt de oplossing geprojecteerd op een ruimte waarin de eigenvektoren behorend bij de kleine eigenwaarden niet behoren . Op deze manier zullen we de invloed van deze kleine eigenwaarden niet meer voelen en kunnen we een veel snellere convergentie verwachten. Die wordt in de volgende paragraaf uitgelegd.

5.3 De DICCG-methode

In de vorig paragraaf hadden we een gepreconditioneerd stelsel lineaire vergelijkingen in de vorm van $\mathbf{\bar{A}y} = \mathbf{\bar{b}}$. Deze laatste wordt verkregen na preconditionering van het oorspronkelijke stelsel $\mathbf{Ax} = \mathbf{b}$.
De eigenwaarden van $\mathbf{\bar{A}}$ worden gegeven door:

$$0 < \lambda_1 < ... < \lambda_m << \lambda_{m+1} < ... < \lambda_n.$$

Hierbij wordt aangegeven dat de eerste m eigenwaarden zeer klein zijn ten opzichte van de andere eigenwaarden.
Voor de definitie van DICCG methode nemen we aan dat de vektoren $\mathbf{v_1}, ..., \mathbf{v_m}$ gegeven zijn en onafhankelijk zijn. Voor deze vectoren definieren we de matrix: $\mathbf{V} = [\mathbf{v_1}...\mathbf{v_m}]$. In eerste instantie kunnen voor de vectoren $\mathbf{v_i}$ precies de eigenvectoren genomen worden, die behoren bij de kleine eigenwaarden. Later zullen we zien dat ze algemener gekozen mogen worden.
Nu definiëren we de deelruimte $\mathcal{V}$ opgespannen door deze eerste m vektoren $\mathbf{v_i}$:

$$\mathcal{V} = span\{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_m}\}.$$

Voor ieder iteratie kan het residu gedefinieerd worden : $\mathbf{r_i} = \mathbf{\bar{b}} - \mathbf{\bar{A}y_i}$. Eerder hebben we gezien dat het residu $\mathbf{r_i}$ in de Krylov-ruimte $K^i(\mathbf{\bar{A}}, \mathbf{r_0})$ ligt. Noteren we voortaan de Krylov-ruimte $K^i(\mathbf{\bar{A}}, \mathbf{r_0})$ met $\mathcal{U}$,

$$\mathcal{U} = span\{\mathbf{r_0}, \mathbf{\bar{A}r_0}, ..., \mathbf{\bar{A}^{i-1}r_0}\}.$$

Zonder verlies van algemeenheid zullen we aannemen dat de startvektor $\mathbf{\bar{x}_0}$ gelijk aan nul is, zodat $\mathbf{r_0} = \mathbf{\bar{b}}$. Dat heeft als resultaat:

$$\mathcal{U} = span\{\mathbf{\bar{b}}, \mathbf{\bar{A}\bar{b}}, ..., \mathbf{\bar{A}^{i-1}\bar{b}}\}.$$

Om het effect van de kleine eigenwaarden op het iteratie proces te annuleren zorgen we ervoor dat het residu $\mathbf{r_i}$ geen componenten in de richting van de kleinste eigenvektoren heeft.
Daarvoor hebben we een transformatie nodig die de ruimte $\mathcal{V}$ projecteert naar de deelruimte $\mathcal{W}$ van $\mathcal{R}^\mathbf{n}$, zodat $\mathcal{V} \cap \mathcal{W}$ een nulruimte is.
De operator $\mathbf{P}$ wordt gedefinieerd door:$\mathbf{P} = \mathbf{I} - \mathbf{VE^{-1}}(\mathbf{\bar{A}V})^\mathbf{T}$, met $\mathbf{E} = (\mathbf{\bar{A}V})^\mathbf{T}\mathbf{V}$.
De projectie $\mathbf{P}$ heeft de volgende eigenschappen:

(i) $\mathbf{P^T\bar{A}V} = \mathbf{0} \implies \mathbf{P^T\bar{A}z} = \mathbf{0}\ \forall \mathbf{z} \in \mathcal{U}$,

(ii) $\mathbf{PV} = \mathbf{0}$,

(iii) $\mathbf{P^2} = \mathbf{P}$,

(iv) $\mathbf{\bar{A}P} = \mathbf{P^T\bar{A}}$.

De oplossing $\mathbf{y}$ kan gesplitst worden in twee deelvektoren:

$$\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y} + \mathbf{Py}.$$

Hierbij ligt $\mathbf{Py}$ in de nieuwe oplossingruimte $\mathcal{W}$ en is $(\mathbf{I}-\mathbf{P})\mathbf{y}$ het gedeelte dat correspondeert met de 'kleine eigenvektoren'.
De eerste deelvektor $(\mathbf{I}-\mathbf{P})\mathbf{y}$ kan bepaald door:

$$(\mathbf{I}-\mathbf{P})\mathbf{y} = \mathbf{V}\mathbf{E}^{-1}(\bar{\mathbf{A}}\mathbf{V})^{\mathbf{T}}\mathbf{y} = \mathbf{V}\mathbf{E}^{-1}\mathbf{V}^{\mathbf{T}}\bar{\mathbf{b}}.$$

Voor de berekening van de tweede deelvector $\mathbf{Py}$ maken we gebruik van eigenschap (iv). Dan volgt $\bar{\mathbf{A}}\mathbf{Py} = \mathbf{P}^{\mathbf{T}}\bar{\mathbf{A}}\mathbf{y} = \mathbf{P}^{\mathbf{T}}\bar{\mathbf{b}}$. Vervolgens kan $\mathbf{y}$ berekend worden door toepassing van CG op het volgende stelsel :

$$\mathbf{P}^{\mathbf{T}}\bar{\mathbf{A}}\mathbf{y} = \mathbf{P}^{\mathbf{T}}\bar{\mathbf{b}}. \tag{5.1}$$

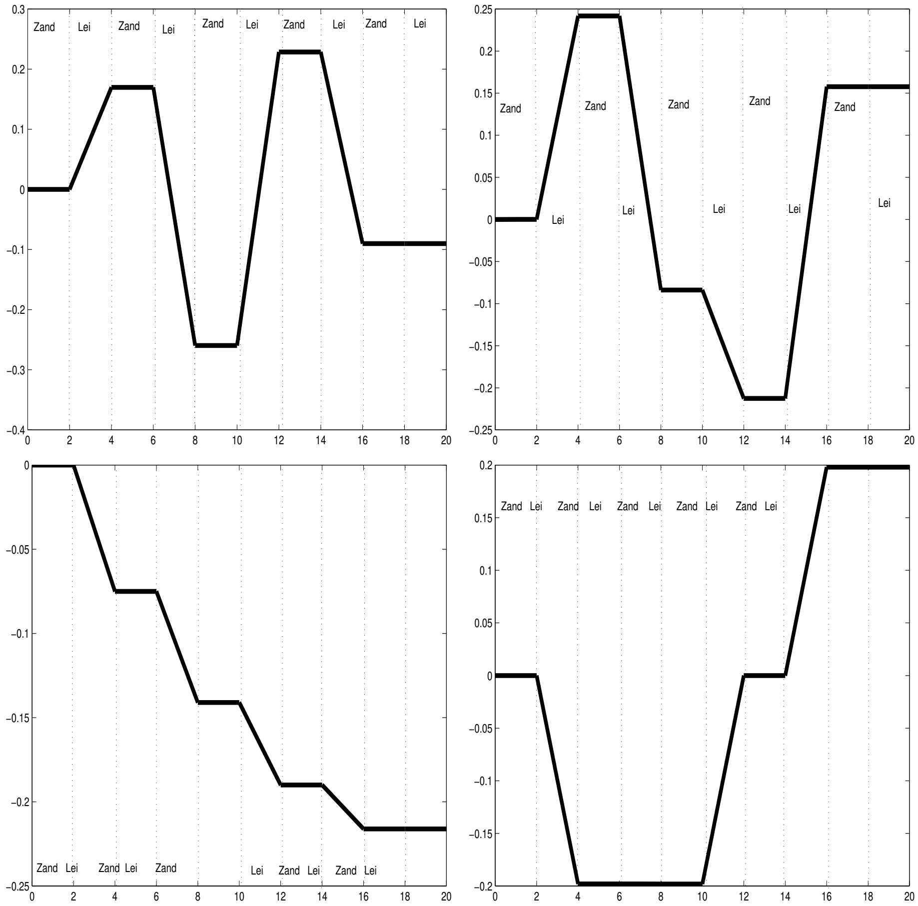
We kunnen makkelijk nagaan dat (5.1) een singulier stelsel is en $\mathbf{P}^{\mathbf{T}}\bar{\mathbf{A}}$ symmetrisch is, als de matrix $\mathbf{V}$ uit kleine eigenvectoren bestaat. Dit geef aanleiding tot de 'Deflated' ICCG methode. Opmerken dat de CG methode snel convergeert als de eigenwaarden precies 0 zijn.
Het bijbehorende algoritme [8] ziet er als volgt uit:

DICCG-algoritme
$k=0, x_0=0; = P^T r_0; p_1 = z_1 = L^{-T}L^{-1}r_0;$
while$||r_k|| > \epsilon$ **do**
$k = k+1;$
$\alpha_k = \frac{(r_{k-1}, z_{k-1})}{(p_k, P^T A p_k)};$
$x_k = x_{k-1} + \alpha_k p_k;$
$r_k = r_{k-1} - \alpha_k P^T A p_k;$
$z_k = L^{-T}L^{-1}r_k;$
$\beta_k = \frac{(r_k, z_k)}{(r_{k-1}, z_{k-1})};$
$p_{k+1} = z_k + \beta_k p_k;$
endwhile
endDICCG-algoritme

Het is duidelijk in te zien dat voor de toepassing van DICCG we eerst over de kleine eigenruimte moeten beschikken. De bepaling van de bijbehorende eigenwaarden en eigenvektoren is duur en bij eenmalig oplossen van het stelsel vergelijkingen zeker niet zinvol. Een goede keuze van projectie vektoren is heel belangrijk om een efficiënte DICCG methode te verkrijgen, d.w.z de bepaling dient goedkoper te zijn qua rekenwerk en geheugenruimte.
In [7] worden kleine eigenvektoren berekend en geanalyseerd voor een testprobleem. Uit de fysica wordt geprobeerd inzicht te krijgen over het gedrag van de eigenvektoren. Uit matlab berekeningen wordt een struktuur van de kleine eigenvectoren gevonden (zie figuur 5.1). Aan de hand van deze struktuur van eigenvectoren kunnen we de 'kleine eigenruimte' benaderen.
Deze analyse geeft ons het idee dat de benaderende eigenvektoren laags-gewijs geconstrueerd kunnen worden op de volgende manier:

- per zandlaag worden de eigenvektoren constant genomen.

- per leilaag voldoen de eigenvektoren aan de diffussievergelijking (4.9) met geschikte gekozen randvoorwaarden ($\frac{\partial p}{\partial n} = 0$ op de kunstmatige randen en p gegeven op de rand van de zandlaag(0 of 1)).



Figuur 5.1: Verticale doorsneden van de eigenvectoren in het geval van 10 lagen

Een mogelijke manier om vertikale doorsneden van projektievektoren v_i op te bouwen voor het gelaagde probleem uit paragraaf 4.3 is de volgende:

- de waarde van v_i is 1 in de $i+1^{er}$ hoog permeabele laag inclusief de interface punten en 0 in de andere hoog permeabele lagen,
- de waarde van v_i in laag permeabele lagen is een lineaire interpolatie van de aangenomen waarden in de interfaces.

Blijkbaar is deze wijze van constructie efficiënt, omdat de benaderende eigenvektoren uit veel nullen bestaan (ijl). Dit bespaart veel rekentijd en geheugen. In [8] wordt voor algemene geometrische configuraties van het gebied verwezen naar een algemene constructie van de projectievektoren. Bovendien wordt aangetoond dat de ruimte opgespannen door deze projektievektoren een goede schatting is van de kleine eigenruimte van de geprecondioneerde matrix. Daarnaast wordt aangetoond dat we de kleine eigenvectoren in de leilagen slechts zeer grof hoeven te benaderen om al een aanzienlijke versnelling van het iteratieproces te verkrijgen.
Voor de definitie van projectievektoren worden de hoog-permeable deelge-

bieden als volgt genummereerd: $\Omega_i, i \in \{1, ..., k^h\}$. Verder zijn k^s hoog-permeable deelgebieden zodanig dat $\bar{\Omega}_i \cap \partial\Omega^D = 0,\, i \in \{1, ..., k^s\}$.. Hierin geeft $\partial\Omega$ aan een rand van een gebied Ω. De rand $\partial\Omega^D$ voldoet aan Dirichlet voorwaarde. De rand $\partial\Omega^N$ voldoet aan Neumann voorwaarde.
En worden Dirichlet randvoorwaarden in de interfaces ($\partial\Omega_j \cap \Omega$) gebruikt en de homogene randvoorwaarden in de buiteranden($v_i = 0$ in $\bar{\Omega}_j \cap \partial\Omega^D$en $\frac{\partial v_i}{\partial n} = 0$ in $\bar{\Omega}_j \cap \partial\Omega^N$).
Bij deze gemelde constructie worden de projektievektoren als volgt gedefinieerd:

- $v_i = 1 \,\text{in}\, \bar{\Omega}_i \,\text{en}\, v_i = 0$ in $\bar{\Omega}_j,\ j \neq i,\ j \in \{1, ..., k^h\}$,
- v_i voldoet aan de eindige element discretisatie van de vergelijking:

$$-div(\sigma_j \nabla(v_i) = 0 \ \in \ \Omega_j,\ j \in \{k^h + 1, ..., k\},$$

Hoofdstuk 6

Gevoeligheidsanalyse

6.1 Inleiding

Uit het vorige hoofdstuk (5) is gebleken dat voor een efficiënte toepassing van de DICCG methode we de deflatievectoren dienen te benaderen. In de praktijk is het vaak niet mogelijk om de projectievectoren te kiezen zodanig dat deze precies de basis van de kleine eigenruimte vormen. Dit geeft aanleiding tot een benadering van de kleine eigenruimte. De DICCG methode kan gevoelig zijn voor zo'n benadering indien deze het optreden van nieuwe kleine eigenwaarden veroorzaakt.

In dit hoofdstuk gaan we het effect van de perturbatie van de kleine eigenruimte in de leilagen analyseren. Om deze gevoeligheidanalyse eenvoudig te houden zullen we eerst het geval beschouwen dat we één perturbatie van de kleine eigenvectoren in één richting van één van de eigenvectoren van $\mathbf{A}$ hebben. In dit geval kunnen we wat makkelijker de verstoorde eigenwaarden bepalen en de bijbehorende eigenvectoren construeren. Daarna zullen we onze analyse uitbreiden tot het geval dat meerdere kleine eigenvectoren in de richting van verschillende eigenvectoren geperturbeerd worden.

In beide gevallen gaan we de volgende uitspraak bewijzen: als de perturbatie zeer klein is, kunnen geen kleine eigenwaarden optreden, derhalve is de DICCG methode niet gevoelig voor een kleine perturbatie. Omgekeerd als de perturbatie zeer groot is, zullen er naast de nuleigenwaarden ook een aantal nieuwe kleine eigenwaarden tot stand komen en dus is de DICCG methode wel gevoelig voor grote verstoringen.

Uit een aantal numerieke testen(zie appendix A4.1) toegepast op een diagonaal matrix blijkt dat de perturbatie in meerdere eigenvectoren optreedt, dit tot meerdere kleine eigenwaarden kan leiden. In dit hoofdstuk gaan we dit resultaat analytisch bevestigen.

6.2 Perturbatie van de projectievectoren in de leilagen

Aangezien A een symmetrische matrix is, is A diagonaliseerbaar, m.a.w. er is een diagonale matrix Σ en een orthogonale matrix S zodat:

$$A = S\Sigma S^T \text{ met } S = [\mathbf{v_1}\mathbf{v_2}....\mathbf{v_n}].$$

Hierin zijn $\mathbf{v_i}$ voor $i = 1..n$ de orthonormale eigenvectoren van A.
De bijbehorende eigenwaaarden $\lambda_1, ..., \lambda_n$ zijn als volgt geordend:

$$0 < \lambda_1 \leq ... \leq \lambda_d << \lambda_{d+1} \leq ... \leq \lambda_n.$$

Hierin geeft d het totaal aantal kleine eigenwaarden van A aan of de dimensie van de kleine eigenruimte.
De projectiematrix P wordt gegeven door:

$$P = I - W((AW)^T W)^{-1}(AW)^T.$$

De matrix W is een matrix met als kolommen de projectievectoren $\hat{\mathbf{w}}_\mathbf{1}, ...\hat{\mathbf{w}}_\mathbf{i}$.
Om de bepaling van de geperturbeerde eigenwaarden van $P^T A$ en de bijbehorende eigenvectoren wat eenvoudiger te maken gaan we eerst laten zien (zie appendix A.3) dat de perturbatie van de kleine eigenruimte behorend bij A equivalent is met de perturbatie van de bijbehorende diagonaal matrix Σ.
D.w.z. er bestaan een projectie matrix $\hat{P}$ en een eigenruimte $\hat{W} = S^T W$ zodanig dat $P^T A$ en $\hat{P}^T \Sigma$ dezelfde eigenwaarden hebben, met

$$\hat{P} = I - \hat{W}((\Sigma\hat{W})^T\hat{W})^{-1}(\Sigma\hat{W})^T \text{ ofwel } \hat{P} = S^T P S.$$

We bekijken verder de diagonale matrix Σ en de projectiematrix $\hat{P}$. Het gevolg daarvan is dat de eigenvectoren eenheidsvectoren zijn.
In het geval van een willekeurige perturbatie in de leilagen kunnen de projectievectoren opgesplitst worden in twee deelvectoren (als voorbeeld zie appendix A4.2):

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}^k + \hat{\mathbf{w}}^g.$$

De eerste deelvector ligt in de kleine eigenruimte, derhalve is $\hat{\mathbf{w}}^k$ een lineaire combinatie van de kleine eigenvectoren. Die kan als volgt geschreven worden:

$$\hat{\mathbf{w}}^k = \sum_{j=1}^{d} \alpha_j \mathbf{e_j}$$

De tweede deelvector $\hat{\mathbf{w}}^g$ behoort tot het orthogonale complement van de kleine eigenruimte. Deze laatste correspondeert met de perturbatie als gevolg van de keuze van projectievectoren. Die kan als volgt uitgedrukt worden:

$$\hat{\mathbf{w}}^g = \sum_{j=d+1}^{p} \alpha_j \mathbf{e_j}$$

Bij onze analyse zullen we ervan uitgegaan dat als de projectievectoren geen basis van de kleine eigenruimte vormen, dat er sprake is van perturbatie van de kleine eigenruimte. De reden daarvan is dat in dit geval worden niet alle kleine eigenwaarden weggedaan.

6.3 Projectiematrix behorend bij projectievectoren die in de kleine eigenruimte liggen

In deze paragraaf nemen we aan dat de projectievectoren een basis van de kleine eigenruimte vormen. In dit geval worden de projectievectoren als lineare combinatie van de kleine eigenvectoren geformuleerd:

$$\begin{aligned}
\hat{\mathbf{w}}_\mathbf{1} &= r_{1,1}\mathbf{e_1} + r_{1,2}\mathbf{e_2} + ...r_{1,d}\mathbf{e_d} \\
\hat{\mathbf{w}}_\mathbf{2} &= r_{2,1}\mathbf{e_1} + r_{2,2}\mathbf{e_2} + ...r_{2,d}\mathbf{e_d} \\
&\vdots \\
\hat{\mathbf{w}}_\mathbf{d} &= r_{d,1}\mathbf{e_1} + r_{d,2}\mathbf{e_2} + ...r_{d,d}\mathbf{e_d}
\end{aligned}$$

ofwel

$$\hat{\mathbf{w}}_\mathbf{i} = \sum_{j=1}^{d} \mathbf{e_j} r_{i,j} \text{ voor } i = 1..d.$$

In compacte vorm:

$$\hat{W} = \hat{V}R^T \text{ met}$$

$$\hat{V} = [e_1.....e_d] \text{ en } R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1d} \\ r_{21} & r_{22} & & r_{2d} \\ \vdots & & & \vdots \\ & & & \\ r_{d1} & \cdots & & r_{dd} \end{pmatrix}.$$

Er geldt:

$$\Sigma\hat{W} = [\Sigma\hat{\mathbf{w}}_\mathbf{1}...\Sigma\hat{\mathbf{w}}_\mathbf{d}], \text{ met } \Sigma\hat{\mathbf{w}}_\mathbf{i} = \sum_{j=1}^{d} r_{ij}\lambda_j\mathbf{e_j} \text{voor } i = 1..d.$$

Anders geschreven:

$$\Sigma\hat{W} = \hat{V}DR^T, \text{ dus } (\Sigma\hat{W})^T = RD^T\hat{V}^T = RD\hat{V}^T,$$

hierbij is D een diagonaalmatrix die de volgende vorm heeft:

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix},$$

De projectiematrix is bepaald door:

$$\hat{P} = I - \hat{W}\hat{E}^{-1}(\Sigma\hat{W})^T.$$

Hier wordt de matrix $\hat{E}$ gegeven door:

$$\hat{E} = ((\Sigma\hat{W})^T\hat{W}) = RD\hat{V}^T\hat{V}R^T = RDR^T \text{ (want } \hat{V}^T\hat{V} = I).$$

Invullen daarvan in de formule voor $\hat{P}$ levert:

$$\hat{P} = I - \hat{V}R^T(R^{-T}D^{-1}R^{-1})RD\hat{V}^T$$

Na enig rekenwerk ziet er de projectiematrix $\hat{P}$ als er volgt uit:

$$\hat{P} = I - \hat{V}\hat{V}^T$$

Dit resultaat kunnen we ook krijgen voor de matrix A met $V = [\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_d}]$, waarbij V de matrix is met als kolommen de kleine eigenvectoren $\mathbf{v_i}$, $i = 1...d$ Immers:

$$P = S^T\hat{P}S = S^T(I - \hat{V}\hat{V}^T)S = S^TS - (S^T\hat{V})(\hat{V}^TS) = I - VV^T.$$

6.4 Perturbatie in de richting van één eigenvector

Met perturbatie in één richting van één eigenvector wordt bedoeld dat de geperturbeerde deelvectoren behorend bij de projectievectoren opgespannen zijn door één eigenvector :

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}^k + \hat{\mathbf{w}}^g, \text{ waarbij } \hat{\mathbf{w}}^g = \alpha_j\mathbf{e_j},\ j \in \{m+1, ...n\}.$$

Zonder verlies van algemeenheid zullen we aannemen dat de perturbatie in de richting van de eigenvector $\mathbf{e_{m+1}}$ optreedt .

In dit geval worden de projectievectoren als volgt geschreven:

$$\hat{\mathbf{w}}_\mathbf{1} = r_{1,1}\mathbf{e_1} + r_{1,2}\mathbf{e_2} + ... + r_{1,m}\hat{\mathbf{e}}_\mathbf{m} + r_{1,m+1}\mathbf{e_{m+1}}$$

$$\hat{\mathbf{w}}_\mathbf{2} = r_{2,1}\mathbf{e_1} + r_{2,2}\mathbf{e_2} + ... + r_{2,m}\hat{\mathbf{e}}_\mathbf{m} + r_{2,m+1}\mathbf{e_{m+1}}$$

$$\vdots$$

$$\hat{\mathbf{w}}_\mathbf{m} = r_{m,1}\mathbf{e_1} + r_{m,2}\mathbf{e_2} + ... + r_{m,m}\hat{\mathbf{e}}_\mathbf{m} + r_{m,m+1}\mathbf{e_{m+1}}$$

ofwel

$$\hat{\mathbf{w}}_\mathbf{i} = \sum_{j=1}^{m} r_{ij}\mathbf{e_j} + r_{im+1}\mathbf{e_{m+1}}. \tag{6.1}$$

Kortweg:

$$\hat{W} = \hat{V}R^T \text{ met}$$

$$\hat{V} = [\mathbf{e_1}.....\mathbf{e_{m+1}}] \,\text{en}\, R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m+1} \\ r_{2,1} & r_{2,2} & & r_{2,m+1} \\ \vdots & & & \vdots \\ & & & \\ r_{m,1} & r_{m,2} & & r_{m,m+1} \end{pmatrix},$$

Voor onze analyse zullen we de volgende twee matrices $\bar{R}$, D en de vector $\mathbf{r}$ invoeren:

$$\bar{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ & & & \\ r_{m,1} & r_{m,2} & \cdots & r_{m,m} \end{pmatrix}, \; D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{pmatrix} \text{en}$$

$$\mathbf{r^T} = \begin{pmatrix} -r_{1,m+1} & -r_{2,m+1} & \cdots & -r_{m,m+1} \end{pmatrix}.$$

De matrix $\bar{R}$ is een matrix waarvan de elementen de coefficienten zijn van de deelvectoren in de kleine eigenruimte met basis $\mathbf{e_1}...\mathbf{e_m}$.
De matrix D is een diagonaal matrix met als diagonaalelementen de kleine eigenwaarden van de matrix Σ.
Later zullen we zien dat $\bar{R}$, D en $\mathbf{r}$ nodig zijn voor de bepaling van de 'geperturbeerde' eigenwaarde van $\hat{P}^T\Sigma$.
We kunnen makkelijk controleren dat:

$$\Sigma = \begin{pmatrix} D & 0 & & \\ 0 & \lambda_{m+1} & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \text{en}\; R = \begin{pmatrix} \bar{R} & -\mathbf{r} \end{pmatrix}.$$

Nu kunnen we $\hat{W}$ en $\Sigma\hat{W}$ als volgt schrijven:

$$\hat{W} = \hat{V}\begin{pmatrix} \bar{R}^T \\ -\mathbf{r}^T \end{pmatrix} \text{en}\, \Sigma\hat{W} = \hat{V}\begin{pmatrix} D\bar{R}^T \\ -\lambda_{m+1}\mathbf{r}^T \end{pmatrix}.$$

In appendix A4.4 wordt aangetoond dat voor ieder vector $\hat{\mathbf{w}}$ van $span\{\mathbf{e_1}, \mathbf{e_2}, ...\mathbf{e_{m+1}}\}$ $\Sigma\hat{P}^T\hat{\mathbf{w}}$ ook in $span\{\mathbf{e_1}, \mathbf{e_2}, ...\mathbf{e_{m+1}}\}$ ligt.
Omdat per definitie geldt $\Sigma\hat{P}^T\mathbf{e_{per}} = \lambda_{per}\mathbf{e_{per}}$ gaan we een eigenvector $\mathbf{e_{per}}$ van $\Sigma\hat{P}^T$ zoeken die te schrijven is als lineare combinatie van deze eigenvectoren:

$$\mathbf{e_{per}} = \sum_{j=1}^{m} \beta_j\mathbf{e_j} + \beta_{m+1}\mathbf{e_{m+1}}.$$

Omdat de matrix $\Sigma\hat{P}^T$ symmetrisch is, dan moet $\mathbf{e_{per}}$ orthogonaal staan op de nuleigenvectoren $\hat{\mathbf{w}}_\mathbf{1}, ...\hat{\mathbf{w}}_\mathbf{m}$, van $\Sigma\hat{P}^T$ m.a.w:

$$\mathbf{e^T_{per}}\hat{\mathbf{w}}_\mathbf{i} = 0 \;\text{ voor } i = 1..m. \tag{6.2}$$

We noteren $\mathbf{e_{per}}$ nu als:

$$\mathbf{e_{per}} = \begin{pmatrix} \mathbf{x} \\ \beta_{m+1} \\ \mathbf{0} \end{pmatrix}$$

Uit (6.1) en (6.2) volgt:

$$\sum_{j=1}^{m} r_{i,j} x_i + \beta_{m+1} r_{i,m+1} = 0 \text{ ofwel } \sum_{j=1}^{m} r_{i,j} x_i = -\beta_{m+1} r_{i,m+1}.$$

De deelvector $\mathbf{x}$ moet nog bepaald worden door het volgende stelsel op te lossen:

$$\begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ & & & \\ r_{m,1} & r_{m,2} & & r_{m,m} \end{pmatrix} \mathbf{x} = \begin{pmatrix} -\beta_{m+1} r_{1,1+m} \\ -\beta_{m+1} r_{2,1+m} \\ \\ -\beta_{m+1} r_{m,1+m} \end{pmatrix}$$

of wel:

$$\bar{R}\mathbf{x} = \beta_{m+1}\mathbf{r}. \tag{6.3}$$

Om het bewijs iets eenvoudiger te maken, gaan we veronderstellen dat de eerste deelvectoren van de projectievectoren die een basis vormen van de kleine eigenruimte. Als gevolg daarvan is de matrix $\bar{R}$ inverteerbaar en wordt m dus de dimensie van de kleine eigenruimte ($m = d$).
Als gevolg daarvan bestaat er een unieke vector $\mathbf{x}$ waarvoor de vergelijking (6.3) geldig is.
De component β_{m+1} moet ongelijk aan nul zijn, anders is de oplossing $\mathbf{x}$ van (6.3) een nulvector, dan is ook $\mathbf{e_{per}}$ een nulvector. Als schaling van $\mathbf{e_{per}}$ kiezen we $\beta_{m+1} = 1$.
Uit appendix A4.5 volgt dat de projectie matrix $\hat{P}$ ziet er als volgt uit:

$$\hat{P} = I - \hat{V} \begin{pmatrix} \left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1} D & -\lambda_{m+1}\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1}\mathbf{x} \\ -\mathbf{x}^T\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1} D & \mathbf{x}^T\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1}\mathbf{x} \end{pmatrix} \hat{V}^T.$$

In appendix A4.6 wordt aangetoond dat voor $\mathbf{e_{per}}$ de volgende vergelijking geldt:

$$\Sigma\hat{P}\mathbf{e_{per}} = \begin{pmatrix} D\mathbf{x} - D\Big(\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x})\Big) \\ \lambda_{m+1} + \lambda_{m+1}\Big(\mathbf{x}^T\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x})\Big) \\ \mathbf{0} \end{pmatrix}. \tag{6.4}$$

Verder wordt bewezen dat $\mathbf{e_{per}}$ een eigenvector van $\Sigma\hat{P}$ is en de bijbehorende eigenwaarde λ_{per} als volgt gedefinieerd is:

$$\lambda_{per} = \lambda_{m+1}\Big((1 + \mathbf{x}^T\left(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T\right)^{-1}(D\underline{\mathbf{x}} - \lambda_{m+1}\mathbf{x}))\Big). \tag{6.5}$$

In appendix A4.3 wordt aangetoond dat λ_{per} ook de geperturbeerde eigenwaarde is van de bijbehorende volle matrix $A = S\Sigma S^T$.
Op basis van (6.5) zijn we nu in staat te bepalen hoe groot de eigenwaarden van $\Sigma\hat{P}$ als de coëfficiënten $r_{m+1,1}$, $r_{m+1,2} \cdots$en $r_{m+1,m}$ gegeven zijn.
Als de coëfficiënten $r_{m+1,1}$, $r_{m+1,2} \cdots$en$r_{m+1,m}$ naar nul gaan(zeer klein), dan gaat de vector $\mathbf{r}$ ook naar nul, want deze cöefficiënten zijn de min componenten van $\mathbf{r}$.
Omdat er geldt : $\bar{R}\mathbf{x} = \mathbf{r}$ ofwel $\mathbf{x} = \bar{R}^{-1}\mathbf{r}$, dan volgt :

$$||\mathbf{x}|| \leq ||\bar{R}^{-1}||\,||\mathbf{r}||$$

Hieruit kunnen we direct zien dat als $\mathbf{r}$ naar nul gaat, ook de vector $\mathbf{x}$ naar nul gaat .
Uit (6.5) volgt:

$$\lambda_{per} - \lambda_{m+1} = \lambda_{m+1}\left(\mathbf{x}^T(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x})\right).$$

Door toepassing van driehoeksongelijkheid kunnen we de volgende schatting verkrijgen:

$$|\lambda_{per} - \lambda_{m+1}| \leq \lambda_{m+1}||(\mathbf{x}^T(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x})||.$$

$$|\lambda_{per} - \lambda_{m+1}| \leq \lambda_{m+1}||\mathbf{x}^T||||(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}||\,||(D\underline{x} - \lambda_{m+1}\mathbf{x})||.$$

$$|\lambda_{per} - \lambda_{m+1}| \leq \lambda_{m+1}||\mathbf{x}||||(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}||\,(||D\mathbf{x}|| + ||\lambda_{m+1}\mathbf{x}||).$$

Tenslotte:

$$|\lambda_{per} - \lambda_{m+1}| \leq \lambda_{m+1}||(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}||\,(||D|| + \lambda_{m+1})\,||\mathbf{x}||\,||\mathbf{x^T}||). \quad (6.6)$$

In appendix A4.7 wordt bewezen dat de uitdrukking $||(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^T)^{-1}||$ naar $||D^{-1}||$gaat, indien $\mathbf{x}$ naar nul gaat.
Uit de schatting (6.6) kunnen we afleiden dat $|\lambda_{per} - \lambda_{m+1}|$ naar nul convergeert, als $\mathbf{x}$ naar nul convergeert.
Uit de volgende vergelijking:

$$\Sigma\hat{P}\mathbf{e_{per}} = \lambda_{per}\mathbf{e_{per}} = \lambda_{per}\begin{pmatrix} \mathbf{x} \\ 1 \\ \mathbf{0} \end{pmatrix},$$

volgt:

$$\lambda_{per}\mathbf{x} = D\mathbf{x} + D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}(-D\mathbf{x} + \lambda_{m+1}\mathbf{x})$$

Uitwerken van het rechterlid levert:

$$\lambda_{per}\mathbf{x} = D\mathbf{x} - D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}D\mathbf{x} + \lambda_{m+1}D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}\mathbf{x}.$$

Volgens appendix A4.7 is $\mathbf{x}$ een eigenvector van $D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}$ met $\frac{1}{1+\lambda_{m+1}\mathbf{x^T}D^{-1}\mathbf{x}}$ de bijbehorende eigenwaarde, waaruitvolgt:

$$\lambda_{per}\mathbf{x} = D\mathbf{x} + D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}D\mathbf{x} + \frac{\lambda_{m+1}}{1 + \lambda_{m+1}\mathbf{x}^T D^{-1}\mathbf{x}}\mathbf{x}.$$

Door toepassing van de driehoeksongelijkheid wordt de volgende schatting verkregen:

$$\lambda_{per}\,||\mathbf{x}|| \leq ||D||\,||\mathbf{x}|| - ||D(D+\lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}||\,||D||||\mathbf{x}|| + |\frac{\lambda_{m+1}}{1+\lambda_{m+1}\mathbf{x}^T D^{-1}\mathbf{x}}|\,||\mathbf{x}||.$$

Verder uitgewerkt:

$$\lambda_{per} \leq ||D|| + ||D(D+\lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}||\,||D|| + \frac{\lambda_{m+1}}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}} D^{-1}\mathbf{x}}.$$

Uit appendix A4.7 volgt: $||D(D+\lambda_{m+1}\mathbf{x}\,\mathbf{x}^T)^{-1}||$ is begrensd door een constante M en $\frac{\lambda_{m+1}}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}} D^{-1}\mathbf{x}}$ gaat naar nul, als $||\mathbf{x}||$ naar oneindig convergeert. Dit leidt tot de volgende schatting voor λ_{per} :

$$\lambda_{per} \leq (1+M)||\,||D||. \qquad (6.7)$$

Uit de schatting (6.7) kunnen we concluderen dat λ_{per} ook een kleine eigenwaarde van $\Sigma\hat{P}$ is, omdat de diagonaalelementen van D de kleine eigenwaarden van Σ zijn.
Ter illustratie gaan we dit bovengenoemde resultaat (6.7)toepassen op de volgende problemen:

Voorbeeld 1

Beschouwen we de volgende matrix:

$$A = S\Sigma S^T \text{met}$$

$S = [\mathbf{e_1}\,\mathbf{e_2}]$ en $\Sigma = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$.

Hierin is $\lambda_1(=\lambda)$ een kleine eigenwaarde behorend bij de eigenvector $\mathbf{e_1}$, en $\lambda_2(=1)$ een niet kleine eigenwaarde is behorend bij de eigenvector $\mathbf{e_2}$.
De gekozen projectie vector ziet er als volgt uit:

$$\mathbf{w_1} = \mathbf{e_1} + \alpha\mathbf{e_2} = \begin{pmatrix} 1 \\ \alpha \end{pmatrix}.$$

De geperturbeerde projectievector heeft de volgende vorm:

$$\mathbf{e_{per}} = x\mathbf{e_1} + \mathbf{e_2} = \begin{pmatrix} x \\ 1 \end{pmatrix}$$

In dit geval kunnen we makkelijk controleren dat :
$D = \lambda_1 = \lambda$, $\bar{R} = 1$ en $\mathbf{r} = -\alpha$.
Eenvoudig is na te gaan : $x = -\alpha$.
Uit formule (6.5) volgt dat de 'geperturbeerde' eigenwaarde λ_{per} de volgende is:

$$\lambda_{per} = \Big(1 + \alpha(\lambda+\alpha^2)^{-1}(-\alpha+\lambda\alpha)\Big).$$

Na enig rekenwerk krijgen we hetzelfde resultaat die in ([9]) hebben we gezien:

$$\lambda_{per} = \frac{\lambda(1+\alpha^2)}{\lambda+\alpha^2}.$$

Voorbeeld 2

Bekijken we de volgende matrix:

$$A = S\Sigma S^T \text{met}$$

$$S = [\mathbf{e_1}\,\mathbf{e_2}\mathbf{e_3}] \text{ en } \Sigma = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

Hierin zijn λ_1 en λ_1 de kleine eigenwaarden behorend bij de eigenvector $\mathbf{e_1}$ en $\mathbf{e_2}$. De gekozen projectievectoren hebben de volgende vorm:

$$\hat{\mathbf{w}}_\mathbf{1} = \alpha_1\mathbf{e_1} + \alpha_2\mathbf{e_2} + \alpha_3\mathbf{e_3} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

$$\hat{\mathbf{w}}_\mathbf{2} = \beta_1\mathbf{e_1} + \beta_2\mathbf{e_2} + \beta_3\mathbf{e_3} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

De eigenvector $\mathbf{e_{per}}$ van $\Sigma^T\hat{P}$ moet ook liggen in de ruimte opgespannen door eigenvectoren $e_1, e_2 \text{en} e_3$,dus is hij te schrijven als lineare combinatie daarvan:

$$\mathbf{e_{per}} = x_1\mathbf{e_1} + x_2\mathbf{e_2} + \mathbf{e_3} = \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \text{ met } \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

$\mathbf{e_{per}}$ moet orthogonaal zijn aan de projectievectoren $\hat{\mathbf{w}}_\mathbf{1}$ en $\hat{\mathbf{w}}_\mathbf{2}$. Dit zijn de eigenvectoren behorend bij nuleigenwaarden $\Sigma^T\hat{P}$. Deze orthogonaliteit leidt tot de volgende vergelijkingen:

$$x_1\alpha_1 + x_2\alpha_2 = -\alpha_3.$$

$$x_1\beta_1 + x_2\beta_2 = -\beta_3.$$

of wel:

$$\bar{R}\mathbf{x} = \mathbf{r},$$

Met:

$$\bar{R} = \begin{pmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{pmatrix} \text{ en } \mathbf{r} = \begin{pmatrix} -\alpha_3 \\ -\beta_3 \end{pmatrix}.$$

Als $\bar{R}$ inverteerbaar is, d.w.z $\alpha_2\beta_1 - \alpha_1\beta_2 \neq 0$, dan zijn de oplossingen van de bovengenoemde vergelijkingen de volgende:

$$x_1 = \frac{\alpha_2\beta_3 - \alpha_3\beta_2}{\alpha_1\beta_2 - \alpha_2\beta_1}.$$

$$x_2 = \frac{\alpha_1\beta_3 - \alpha_3\beta_1}{\alpha_2\beta_1 - \alpha_1\beta_2}.$$

Toepassing van (6.5) geeft dat de 'geperturbeerde' eigenwaarde λ_{per} de volgende is:

$$\lambda_{per} = \lambda_3\Big((1 + \mathbf{x}^T(D + \lambda_3\mathbf{x}\mathbf{x}^T)^{-1}(D\mathbf{x} - \lambda_3\mathbf{x}))\Big).$$

Hierin :

$$D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \text{en}\, \mathbf{x} = \begin{pmatrix} \frac{\alpha_2\beta_3-\alpha_3\beta_2}{\alpha_1\beta_2-\alpha_2\beta_1} \\ \frac{\alpha_1\beta_3-\alpha_3\beta_1}{\alpha_2\beta_1-\alpha_1\beta_2} \end{pmatrix}.$$

Nu gaan we het bovengenoemde resultaat gebruiken voor de berekeningen van de geperturbeerde λ_{per} in het voorbeeld van appendix A4.1 (tweede geval).
In dit geval hebben we:

$$D = 10^{-6}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

en $\alpha_1 = 4$, $\alpha_2 = 3$, $\beta_1 = 2$, $\beta_2 = 5$, $\lambda_3 = 1$.
De 'geperturbeerde' eigenwaarde λ_{per} is gelijk aan de waarde $1.0795e - 05$ die hebben we eerder in apendix A4.1 direct berekend met matlab.
Om de invloed van perturbatie duidelijk te maken zijn de geperturbeerde eigenwaarden λ_{per} berekend met matlab voor verschillende grootte van de perturbatie.
De numerieke resultaten zijn aangegeven aan de hand van de volgende tabel:

$\|\|\mathbf{r}\|\|\|$	geperturbeerde eigenwaarden	numerieke fout	geschatte verschill
10^3	9.8671e-06	1.5686552896763974e-07	1
10^2	1.0008e-05	3.055652278763287e-09	1
10^1	1.0146e-05	8.347576915725824e-13	1
10^0	2.4604e-05	3.732247428050051e-13	1
10^{-1}	0.0015	1.505089455844377e-15	0.9985
10^{-2}	0.1274	1.110223024625157e-16	0.8726
10^{-3}	0.9359	-1.110223024625157e-16	0.0641

6.5 Perturbatie in de richting van meerdere eigenvectoren

Onder perturbatie in de richting van meerdere eigenvectoren wordt verstaan dat de geperturbeerde deelvectoren behorend bij de projectievectoren opgespannen zijn door meerdere eigenvectoren :

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}^k + \hat{\mathbf{w}}^g, \text{ waarbij } \hat{\mathbf{w}}^g = \sum_{j=m+1}^{p} \alpha_j \mathbf{e_j}.$$

In dit geval worden de projectievectoren als volgt geformuleerd:

$$\hat{\mathbf{w}}_\mathbf{1} = r_{1,1}\mathbf{e_1} + r_{1,2}\mathbf{e_2} + ... + r_{1,m}\mathbf{e_m} + r_{1,m+1}\mathbf{e_{m+1}} + ... + r_{1,p}\mathbf{e_p}$$

$$\hat{\mathbf{w}}_\mathbf{2} = r_{2,1}\mathbf{e_1} + r_{2,2}\mathbf{e_2} + ... + r_{2,m}\mathbf{e_m} + r_{2,m+1}\mathbf{e_{m+1}} + ... + r_{2,p}\mathbf{e_p}$$

$$\vdots$$

$$\hat{\mathbf{w}}_\mathbf{m} = r_{m,1}\mathbf{e_1} + r_{m,2}\mathbf{e_2} + ... + r_{m,m}\mathbf{e_m} + r_{m,m+1}\mathbf{e_{m+1}} + ... + r_{m,p}\mathbf{e_p}$$

ofwel

$$\hat{\mathbf{w}}_{\mathbf{i}} = \sum_{j=1}^{m} r_{ij}\mathbf{e_j} + \sum_{j=m+1}^{p} r_{ij}\mathbf{e_j}. \tag{6.8}$$

In compacte vorm geschreven:

$$\hat{W} = \hat{V} R^T \text{ met}$$

$$\hat{V} = [\mathbf{e_1}.....\mathbf{e_p}] \,\text{en}\, R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,p} \\ r_{2,1} & r_{2,2} & & r_{2,p} \\ \vdots & & & \vdots \\ \\ r_{m,1} & r_{m,2} & & r_{m,p} \end{pmatrix},$$

Nu kunnen we $\hat{W}$ en $\Sigma\hat{W}$ als volgt schrijven:

$$\hat{W} = \hat{V} \begin{pmatrix} \bar{R}^T \\ -\underline{R}^T \end{pmatrix} \text{ en } \Sigma\hat{W} = \hat{V} \begin{pmatrix} \bar{D}\,\bar{R}^T \\ -\underline{D}\,\underline{R}^T \end{pmatrix},$$

waarbij zijn de matrices $\bar{R}$, $\bar{D}$ $\underline{R}$ en $\underline{D}$ gegeven door:

$$\bar{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \\ \\ r_{m,1} & r_{m,2} & \cdots & r_{m,m} \end{pmatrix}, \ \bar{D} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{pmatrix},$$

$$\underline{R} = \begin{pmatrix} -r_{1,m+1} & -r_{1,m+2} & \cdots & -r_{1,p} \\ -r_{2,m+1} & -r_{2,m+2} & \cdots & -r_{2,p} \\ \\ \\ -r_{m,m+1} & -r_{m,m+2} & \cdots & -r_{m,p} \end{pmatrix} \text{ en } \underline{D} = \begin{pmatrix} \lambda_{m+1} & & & \\ & \lambda_{m+2} & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}.$$

Als in het vorige geval kunnen we respectievelijk de matrices Σ en R splitsen als volgt:

$$\Sigma = \begin{pmatrix} \bar{D} & 0 & & & \\ 0 & \underline{D} & & & \\ & & \lambda_{p+1} & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix} \text{ en } R = \begin{pmatrix} \bar{R} & -\underline{R} \end{pmatrix}.$$

Voor de opbouw van de eigenvectoren gaan we op analoge wijze als in het vorige geval de vectoren $\mathbf{e_{per,k}}$ $(k = m+1...p)$ van $\Sigma\hat{P}^T$ zoeken die geschreven kan worden als lineaire combinatie van de eigenvectoren $\mathbf{e_j}$, $k = m+1...p$:

$$\mathbf{e_{per,k}} = \sum_{j=1}^{m} \beta_{k,j}\mathbf{e_j} + \sum_{j=m+1}^{p} \beta_{k,j}\mathbf{e_j}.$$

De vectoren $\mathbf{e_{per,k}}$, $k = m+1..p$ staan orthogonaal op de nuleigenvectoren $\mathbf{\hat{w}_1}, ...\mathbf{\hat{w}_m}$, van $\Sigma\hat{P}^T$ m.a.w:

$$\mathbf{e_{per,k}^T \hat{w}_i} = 0 \;\text{ voor } i = 1..m,\; k = m+1..p. \tag{6.9}$$

In eerste instantie zoeken we de vectoren $\mathbf{e_{per,k}}$ die de volgende vorm hebben:

$$\mathbf{e_{per,k}} = \sum_{j=1}^{m} x_{k,j}\mathbf{e_j} + \mathbf{e_{m+k}} = \begin{pmatrix} \mathbf{x_k} \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{6.10}$$

Uit (6.9)en(6.10) volgt voor ieder $k = 1+m..p$:

$$\begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ & & & \\ r_{m,1} & r_{m,2} & & r_{m,m} \end{pmatrix} \begin{pmatrix} x_{k,1} \\ x_{k,2} \\ \\ x_{k,m} \end{pmatrix} = \begin{pmatrix} -r_{1,k+m} \\ -r_{2,k+m} \\ \\ -r_{m,k+m} \end{pmatrix}$$

Kortweg:

$$\bar{R}\mathbf{x_k} = \mathbf{r_k}. \tag{6.11}$$

Hierin zijn voor ($k = m+1...p$)
$\mathbf{r_k} = (-r_{1,m+k}, ..., -r_{m,m+k})^T$.
$\mathbf{x_k} = (-r_{k,1}, ..., -r_{k,m})^T$.
In matrix-vorm:

$$\bar{R}X = \underline{R}. \tag{6.12}$$

Hierin is $X = [\mathbf{x_{1+m}}\ \mathbf{x_{2+m}}...\mathbf{x_p}]$, $\underline{R} = [\mathbf{r_{1+m}}\ \mathbf{r_{2+m}}...\mathbf{r_p}]$.
Voor de bepaling van de 'geperturbeerde' eigenvectoren gaan we de matrix E_{per} introduceren:

$$E_{per} = [\mathbf{e_{per,m+1}}, ..., \mathbf{e_{per,p}}].$$

ofwel:

$$E_{per} = \begin{pmatrix} X \\ I \\ \mathbf{0} \end{pmatrix}.$$

Uit appendix A5.1 volgt dat de projectie matrix $\hat{P}$ er als volgt er uitziet:

$$\hat{P} = I - \hat{V} \begin{pmatrix} (\bar{D} + X\underline{D}X^T)^{-1}\bar{D} & -(\bar{D} + X\underline{D}X^T)^{-1}X\underline{D} \\ -X^T(\bar{D} + X\underline{D}X^T)^{-1}\bar{D} & X^T(\bar{D} + \lambda_{m+1}X\underline{D}X^T)^{-1}X\underline{D} \end{pmatrix} \hat{V}^T.$$

In appendix A5.2 wordt aangetoond dat voor E_{per} de volgende vergelijking geldt:

$$\Sigma\hat{P}E_{per} = \begin{pmatrix} \bar{D}X - \bar{D}(\bar{D}+X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{D} + \underline{D}X^T(\bar{D}+X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{0} \end{pmatrix}. \qquad (6.13)$$

Nu gaan we de matrix D_{per} als volgt definiëren:

$$D_{per} = \underline{D} + \underline{D}X^T(\bar{D}+X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}). \qquad (6.14)$$

In appendix A5.2 hebben we bewezen dat er geldt bovendien:

$$\Sigma\hat{P}E_{per} = E_{per}D_{per} = \begin{pmatrix} X \\ I \\ \mathbf{0} \end{pmatrix} D_{per}. \qquad (6.15)$$

Laat $\mathbf{g}$ een eigenvector van D_{per} zijn en λ_{per} de bijbehorende eigenwaarde. Er geldt dus: $D_{per}\mathbf{g} = \lambda_{per}\mathbf{g}$.
Nu hebben we :

$$\Sigma\hat{P}E_{per}\mathbf{g} = E_{per}D_{per}\mathbf{g} = \lambda_{per}E_{per}\mathbf{g}.$$

Hiermee wordt bewezen dat de matrices $\Sigma\hat{P}$ en D_{per} dezelfde eigenwaarden hebben, daarom gaan we de perturbatie van de matrix D_{per} in plaats van $\Sigma\hat{P}$ analyseren.
Uit vergelijking (6.14) volgt:

$$D_{per} - \underline{D} = \underline{D}X^T(\bar{D}+X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}).$$

Door nogmaals toepassing van driehoeksongelijkheid kunnen we de volgende schatting verkrijgen:

$$||D_{per} - \underline{D}|| \leq ||\underline{D}||\,(||\bar{D}|| + ||\underline{D}||)||(\bar{D}+X\underline{D}X^T)^{-1}||\,||X^T||\,||X||. \qquad (6.16)$$

In appendix A5.3 hebben we laten zien dat de uitdrukking $||(\bar{D}+X\underline{D}X^T)^{-1}||$ naar $||\bar{D}^{-1}||$ gaat, indien $||\mathbf{X}||$ naar nul gaat.
Uit de schatting (6.16) kunnen we afleiden dat $||D_{per} - \underline{D}||$ naar nul convergeert, als $||X||$ naar nul convergeert.
Combinatie van vergelijkingen (7.13) en (7.15) geeft:

$$XD_{per} = \bar{D}X - \bar{D}(\bar{D}+X\,\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D})$$

Voor de vector $\mathbf{g}$ geldt:

$$XD_{per}\mathbf{g} = \bar{D}X\mathbf{g} - \bar{D}(\bar{D}+X\,\underline{D}X^T)^{-1}(\bar{D}X\mathbf{g} - X\underline{D}\mathbf{g})$$

Aangezien $\mathbf{g}$ een eigenvector is en λ_{per} de bijbehorende eigenwaarde:

$$\lambda_{per}X\mathbf{g} = \bar{D}X\mathbf{g} - \bar{D}(\bar{D}+X\,\underline{D}X^T)^{-1}\bar{D}X\mathbf{g} + \bar{D}(\bar{D}+X\,\underline{D}X^T)^{-1}X\underline{D}\mathbf{g}$$

Toepassing van driehoeksongelijkheid hierop levert:

$$\lambda_{per}||X\mathbf{g}|| \leq ||\bar{D}||\,X\mathbf{g}|| + ||\bar{D}||\,||(\bar{D}+X\,\underline{D}X^T)^{-1}||\bar{D}X\mathbf{g} + X\underline{D}\mathbf{g}||$$

$$\lambda_{per}||X\mathbf{g}||\,|| \leq ||\bar{D}||\,X\mathbf{g}|| + ||\bar{D}||\,||(\bar{D}+X\,\underline{D}X^T)^{-1}(||\bar{D}||\,||X\mathbf{g}|| + ||X\underline{D}\mathbf{g}||)$$

We gaan eruit dat er bestaat tenmiste één eigenwaarde λ_{per}, zodanig dat de bijbehorende eigenvector $\mathbf{g}$ van $\Sigma\hat{P}$ voldoet aan: $X\mathbf{g} \neq 0$, anders ligt iedere eigenvector $E_{per}\mathbf{g}$ in de deelruimte opgespannen door de niet kleine eigenvectoren. In dit geval hoeven we verder niets te bewijzen.
Verder er geldt:

$$\lambda_{per} \leq ||\bar{D}||(1+||(\bar{D}+X\,\underline{D}X^T)^{-1}(||\bar{D}|| + \frac{||X\underline{D}\mathbf{g}||}{||X\mathbf{g}||})$$

In appendix A5.3 wordt verder aangetoond dat $||(\bar{D}+X^T\underline{D}X)^{-1}||$ en $\frac{||X\underline{D}\mathbf{g}||}{||X\mathbf{g}||}$ begrensd zijn als $||X||$ naar oneindig convergeert,
waaruit volgt:

$$||\lambda_{per}|| \leq ||\bar{D}||(1+M). \tag{6.17}$$

Op grond van schatting (6.17) kunnen we afleiden dat er bestaat tenminste één kleine eigenwaarde λ_{per} van $\Sigma\hat{P}$ als $||X||$ groot is.

6.6 Conclusies

- Uit paragraaf (7.3) is gebleken dat de projectiematrix P onafhankelijk is van de keuze van de projectievectoren die een basis van de kleine eigenruimte zijn, maar wel van de bijbehorende kleine eigenruimte. Dit resultaat geeft het idee dat we de projectiematrix P bepalen door sparse projectievectoren te construeren, indien deze een basis vormen van de kleine eigenruimte.

- In paragraaf (7.4) is de perturbatie van de kleine eigenruimte in één richting van één van de niet kleine eigenvectoren geanalyseerd. Hieruit is te concluderen dat als de perturbatie zeer klein is, gaat de geperturbeerde eigenwaarde λ_{per} naar een niet kleine eigenwaarde. Omgekeerd gaat λ_{per} naar een kleine eigenwaarde als de perturbatie zeer groot is.

- In paragraaf (7.5) is onze analyse uitgebreid tot de perturbatie van de kleine eigenruimte in verschillende richtingen van de niet kleine eigenvectoren. Hieruit kunnen we concluderen dat de geperturbeerde eigenwaarde λ_{per} naar een niet kleine eigenwaarde gaat, als de perturbatie zeer klein is. Omgekeerd gaat λ_{per} naar een kleine eigenwaarde als de perturbatie zeer groot is.

Hoofdstuk 7

Constructie van projectievectoren

7.1 Inleiding

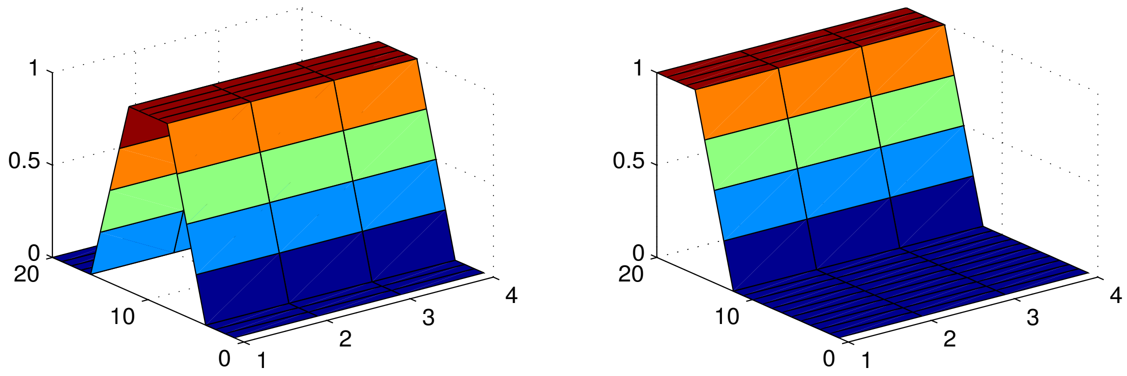
We hebben gezien in hoofdstuk (5) dat de kleine eigenruimte benaderd kan worden door de projectievectoren constant in zandlagen te kiezen en binnen de leilagen een diffusieprobleem op lossen. Bijvoorbeeld in de situatie van een rechte structuur van afwisselend zand en lei is bij eerste benadering de waarde van de projectievectoren 1 in de bijbehorende zandlaag en 0 in de andere zandlagen. De waarde in leilagen is een lineaire interpolatie van de aangenomen waarden in de interface. In de praktijk is deze constructie (zie figuur 8.1a) vaak niet efficiënt. Vooral in het geval van een groot aantal leilagen moeten we meerdere diffusieproblemen oplossen. Daarom wordt ernaar gestreefd om een zeer grove constructie van de de projectievectoren te vinden. Deze nieuwe constructie houdt in dat projectievectoren alleen uit enen en nullen bestaan in de verschillende deelgebieden. Het basisidee daarvan is dat een eigenvector geschreven kan worden als lineaire combinatie van basisvectoren waarvan de componenten alleen enen en nullen zijn , namelijk de eenheidsvectoren:

$$\mathbf{v} = \sum_{i}^{n} \alpha_i \mathbf{e_i}, \text{ hier zijn } \mathbf{e_i} \text{ de eenheidsvectoren}$$

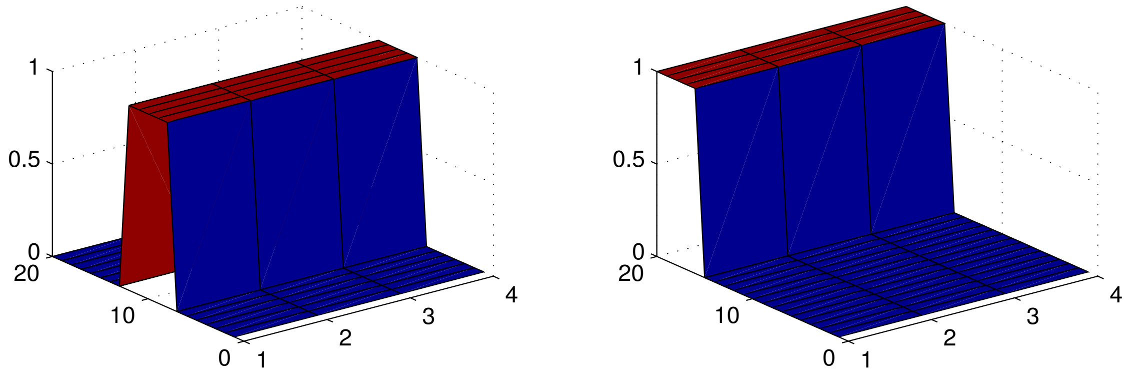
Uit de structuur van de eigenvectoren (bijvoorbeeld constant of lineair) is er een expliciete relatie tussen de coëfficienten α_i , d.w.z. we kunnen door de ene coëfficiënt de anderen bepalen (zie voorbeeld in appendix A2.2). Door deze afhanklijkheid tussen de de coefficiënten α_i kunnen we de ene cöefficiënt vrij kiezen en de andere vastleggen, dan worden de eigenvectoren een lineare combinatie van een klein aantal basisvectoren $\mathbf{w_i}$:

$$\mathbf{v} = \sum_{i}^{m} \beta_i \mathbf{w_i}.$$

Deze basisvectoren kunnen we gebruiken als projectievectoren. Want iedere eigenvector ligt in de ruimte opgespannen door deze vectoren. Bovendien is de kleine eigenruimte een deelruimte van deze laatste ruimte. Maar in de praktijk zullen we ons beperken tot benaderingen van deze basisvectoren. Voor benaderingen van deze basisvectoren komen verschillende constructies van de projectievectoren aan de orde (de benadering via oplossen van het stelsel in het 'lei' gebied zullen we aanduiden als de eerste benadering):



(a):Projectie vectoren van eerste benadering



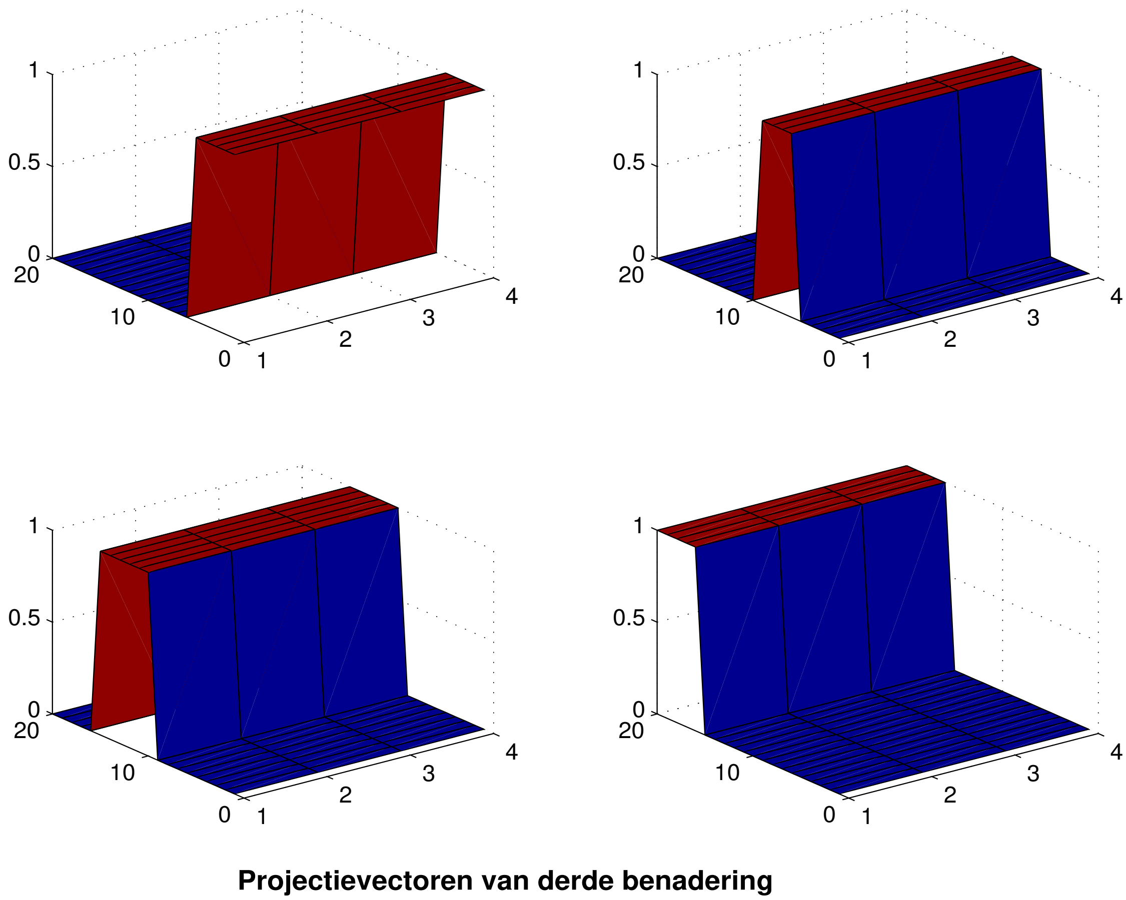
(b) :Projectie vectoren van tweede benadering

Figuur 7.1: Schets van projectievectoren in het geval van 5 lagen

- Bij de tweede benadering is de waarde van de projectievectoren 1 in de bijbehorende zandlaag en 0 in de andere zand-lei lagen (zie figuur 8.1b).
- Bij de derde benadering hoort bij ieder laag een projectievector. De waarde van de projectievectoren in de bijbehorende laag is 1 en 0 in de resterende lagen(zie figuur1.3). We hebben dus veel meer projectievectoren dan de eigenvectoren.

Voordelen van deze constructies zijn de volgende:

- deze constructie van de projectievectoren is erg simpel en dus is de implementatie daarvan zeer eenvoudig.



Figuur 7.2: Schets van de vier projectievectoren in het geval van 4 lagen

- deze constructie is zeer efficiënt qua rekenwerk en geheugenruimte. Want we hoeven geen diffusieprobleem in verschillende gebieden oplossen.

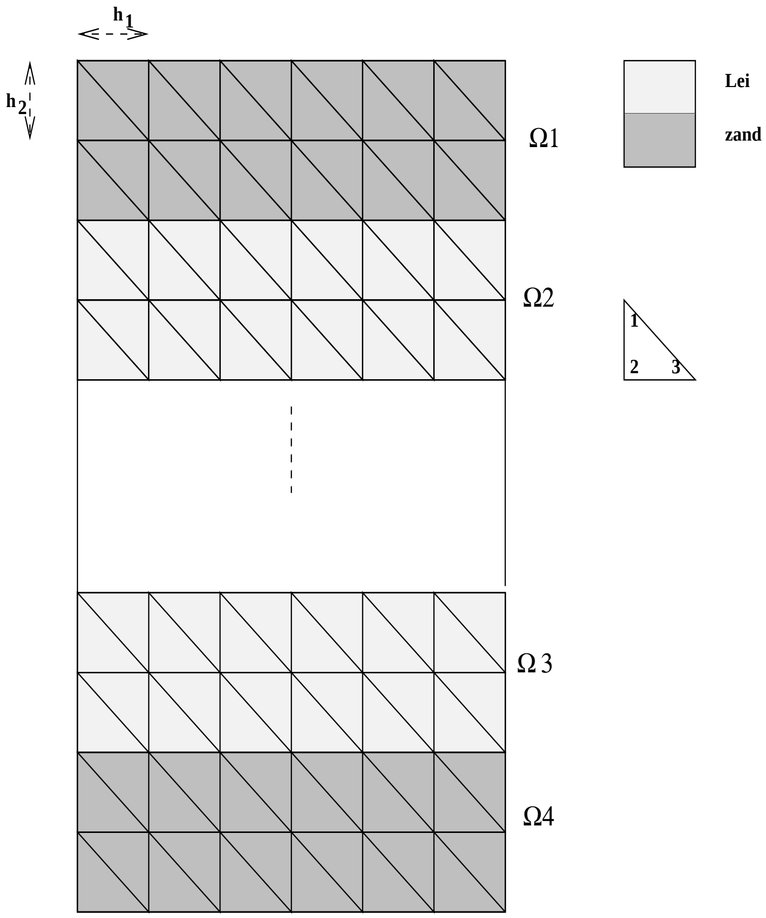
Om de vermelde constructies van de projectievectoren te onderzoeken kijken we naar verschillende structuren van de ondergrond. We beginnen met een eenvoudige structuur en voeren de moeilijkheidsgraad geleidelijk op. Bij iedere structuur van de ondergrond zullen we het convergentiegedrag van DICCG nader onderzoeken. Vooral wordt gekeken naar het aantal iteraties om convergentie te bereiken.
Om te zien in hoeverre onze constructies de DICCG methode beïnvloeden, worden een paar numerieke experimenten met Sepran en Matlab gedaan. Hierbij worden de twee bovengenoemde benaderingen van de kleine eigenruimte getest. Bij het testen wordt een incomplete Cholesky preconditionering gebruikt.

7.2 Toepassingen

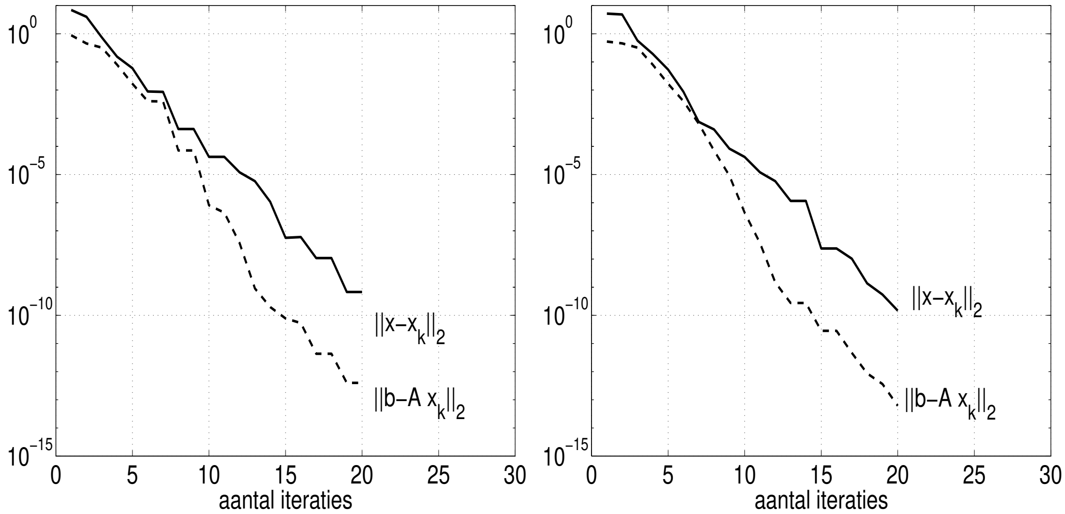
We zullen onze benaderingen toepassen op drie verschillende problemen. Het eerste probleem is een recht gelaagd probleem. Het tweede probleem is een krom gelaagd probleem. Het derde probleem is een probleem met zandinsluitingen.

7.2.1 Afwissende rechte lagen

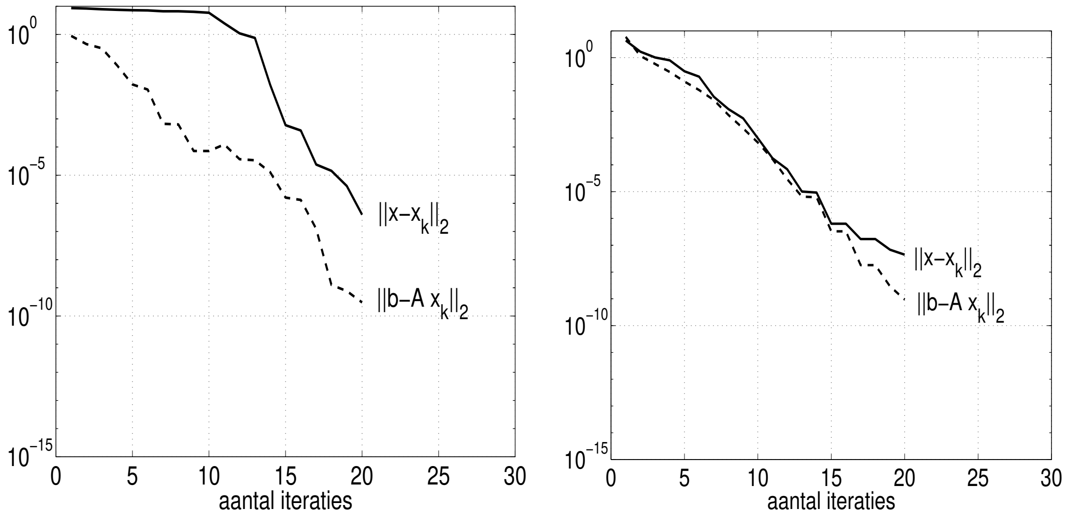
Allereerst gaan we deze constructies toepassen op de eenvoudig structuur bestaande uit rechte lagen van afwisselend zand en lei zoals we die in hoofdstuk (4) in paragraaf 4.3 hebben laten zien (zie figuur 8.3).



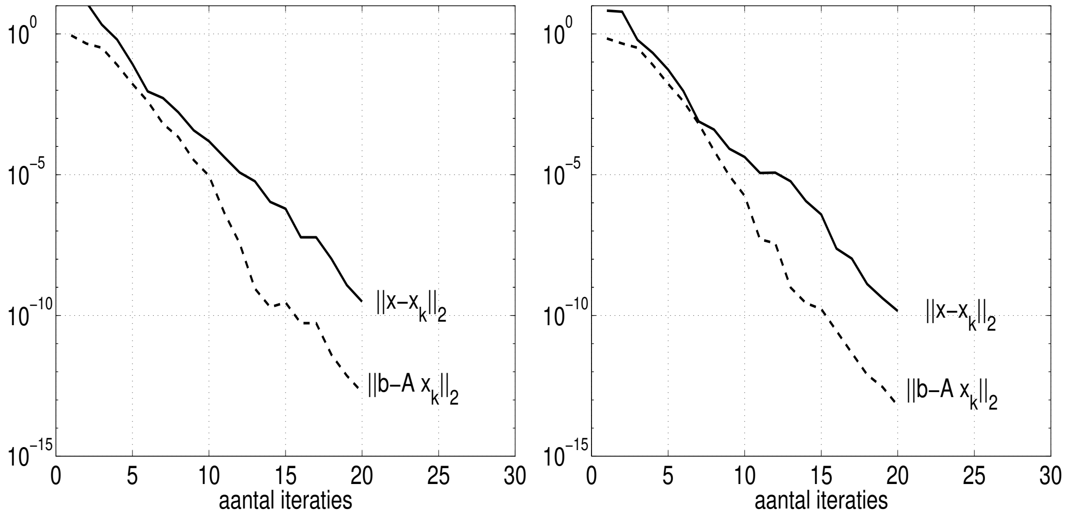
Figuur 7.3: Schets van verdeling van afwisselend zand en lei structuur van de ondergrond in driekhoekselementen



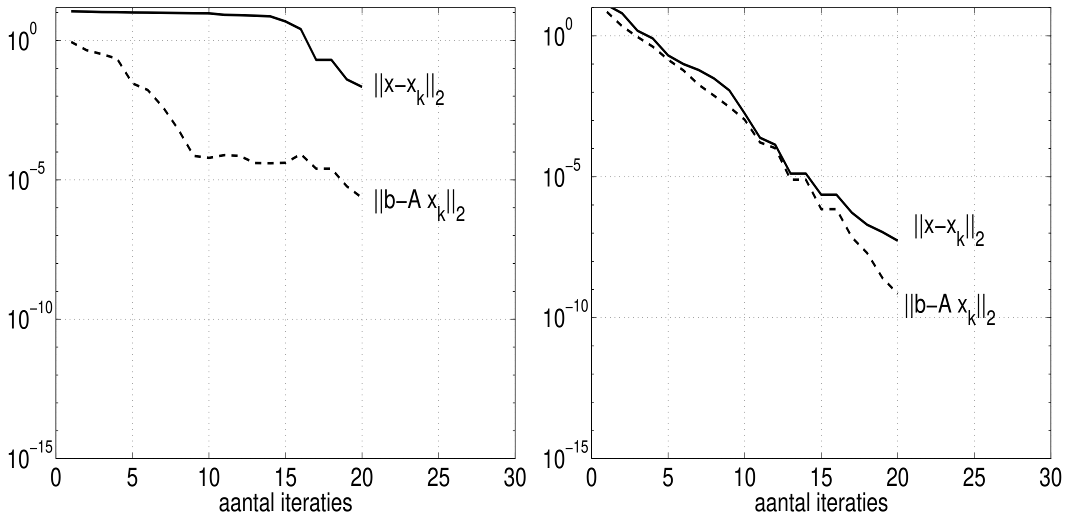
Figuur 7.4: Grafieken in het geval van 7 lagen (links):echte kleine eigenruimte (rechts) DICCG methode eerste benadering



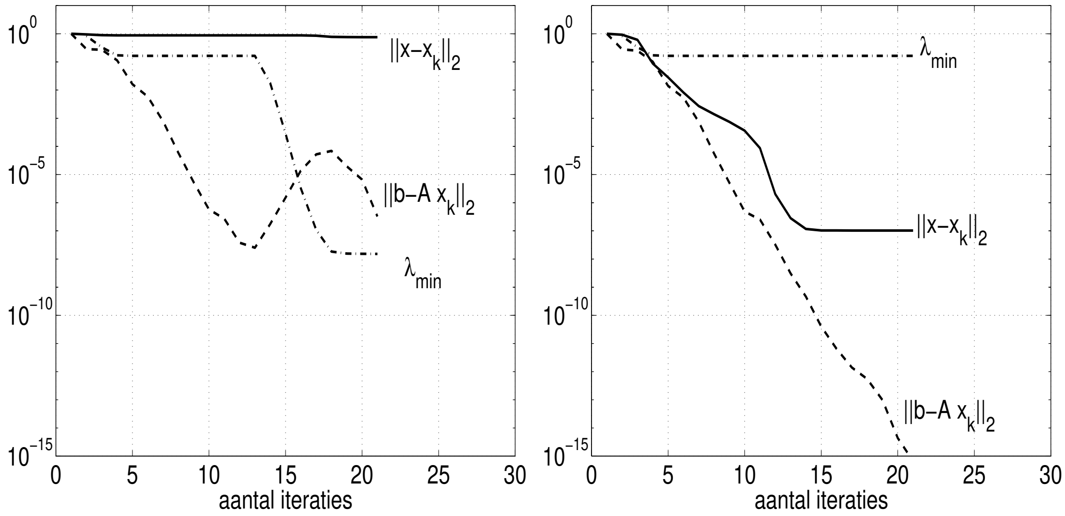
Figuur 7.5: Grafieken in het geval van 7 lagen (links):DICCG methode(tweede benadering) (rechts) DICCG methode (derde benadering)



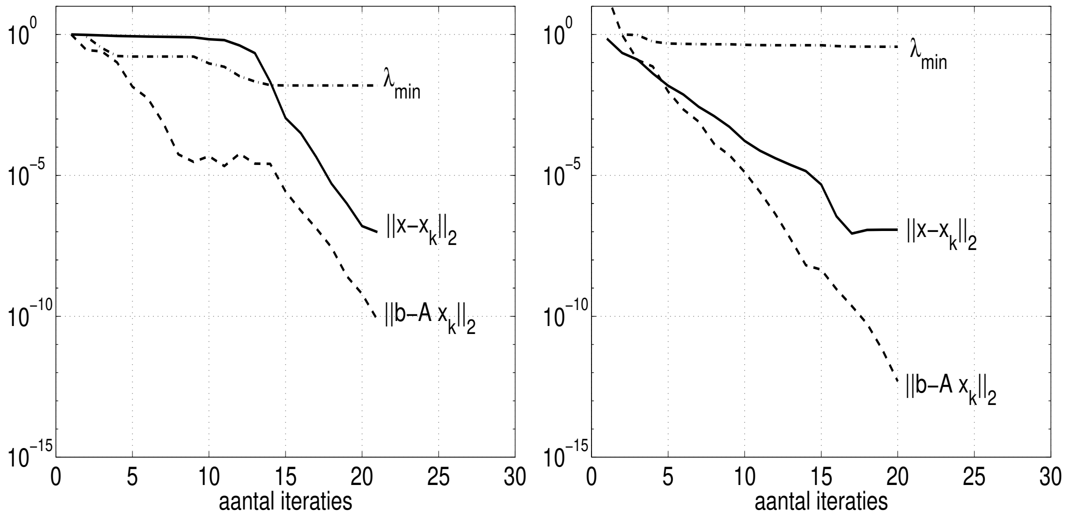
Figuur 7.6: Grafieken in het geval van 11 lagen (links):echte kleine eigenruimte (rechts) DICCG methode eerste benadering



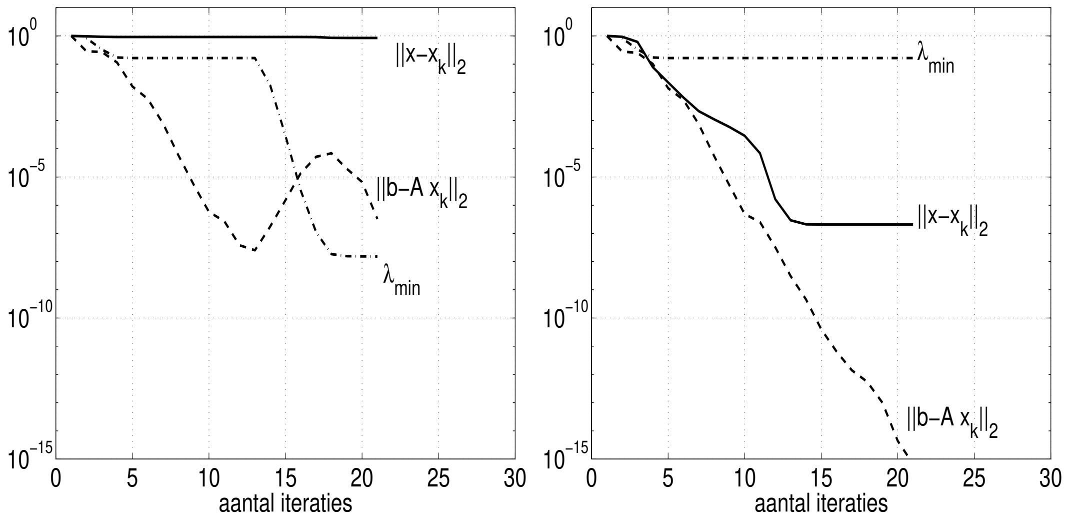
Figuur 7.7: Grafieken in het geval van 11 lagen (links):DICCG methode(tweede benadering) (rechts) DICCG methode (derde benadering)



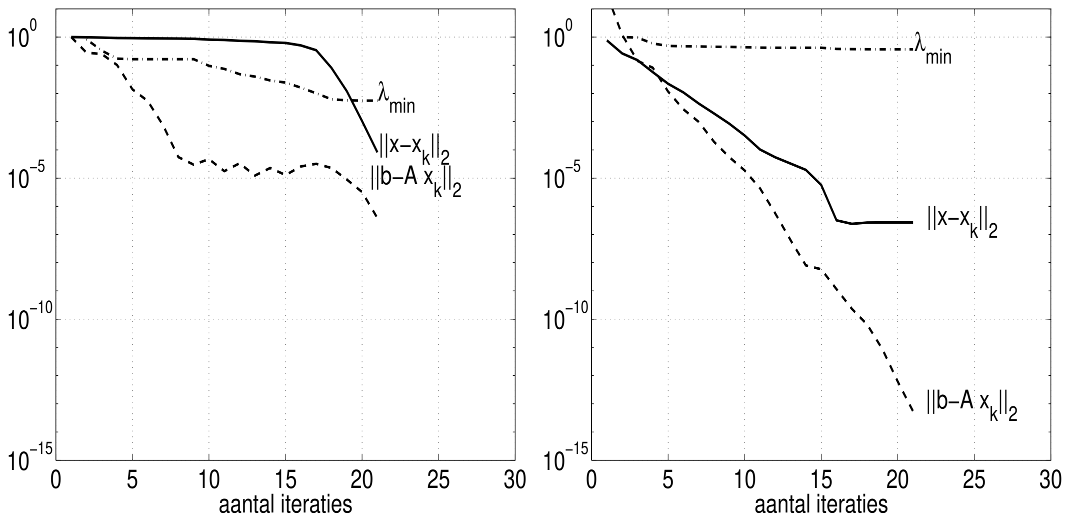
Figuur 7.8: Grafieken in het geval van 7 lagen (links):ICCG methode (rechts) DICCG methode eerste benadering



Figuur 7.9: Grafieken in het geval van 7 lagen (links):DICCG methode(tweede benadering) (rechts) DICCG methode (derde benadering)



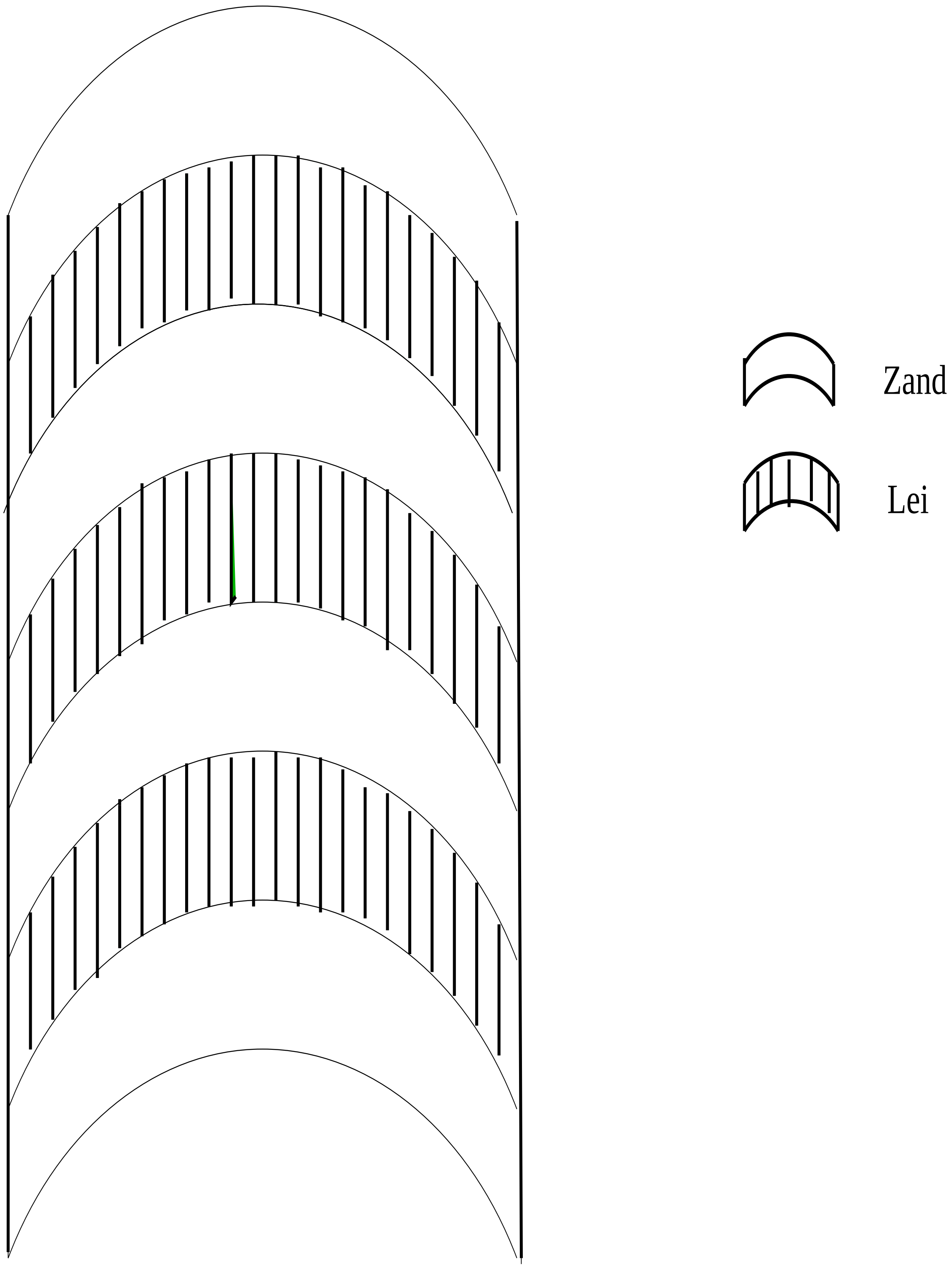
Figuur 7.10: Grafieken in het geval van 11 lagen (links):ICCG methode (rechts) DICCG methode eerste benadering



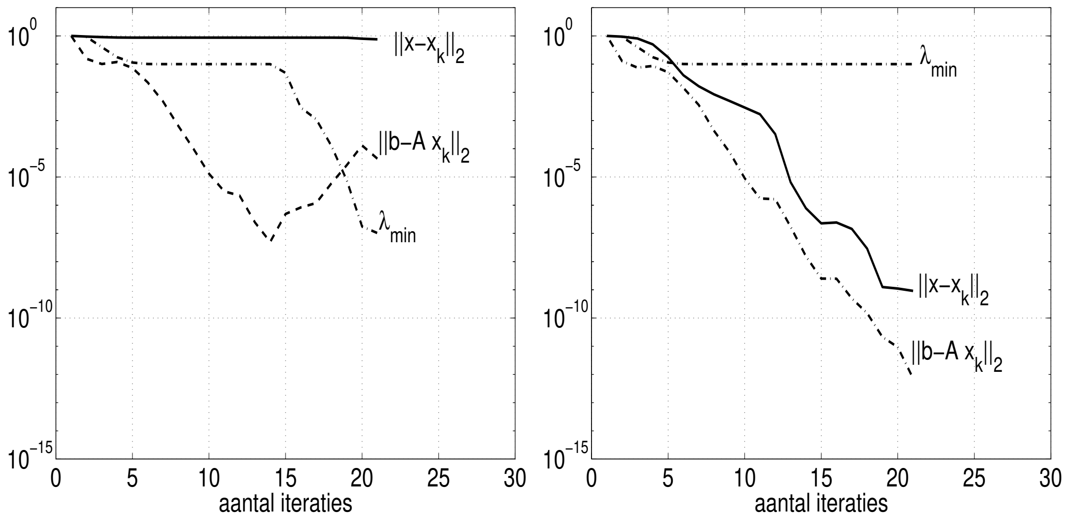
Figuur 7.11: Grafieken in het geval van 11 lagen (links):DICCG methode(tweede benadering) (rechts) DICCG methode (derde benadering)

7.2.2 Afwisselende kromme lagen

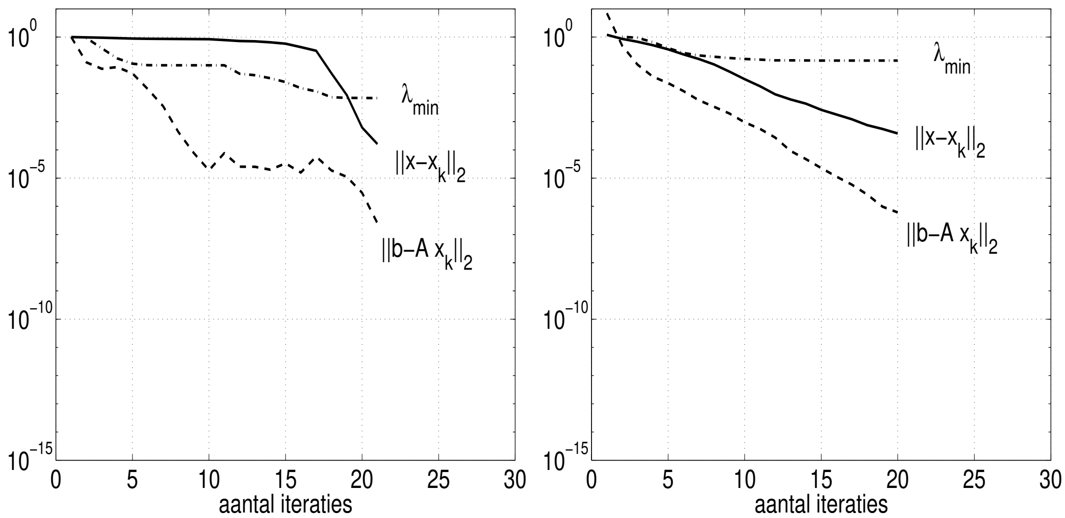
Verder gaan we ons benaderingen toepassen op een gebied met kromme lagen. Hiertoe bekijken we kromme lagengebied met 7 lagen van afwisselend zand en lei (zie figuur (8.12)).



Figuur 7.12: Chets van een afwisselend kromme lagen



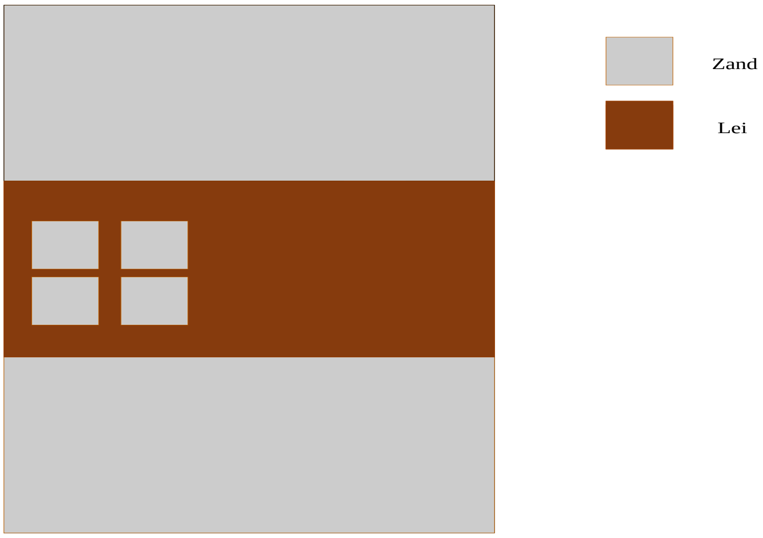
Figuur 7.13: Grafieken in het geval van 7 lagen (links):ICCG methode (rechts) DICCG methode eerste benadering



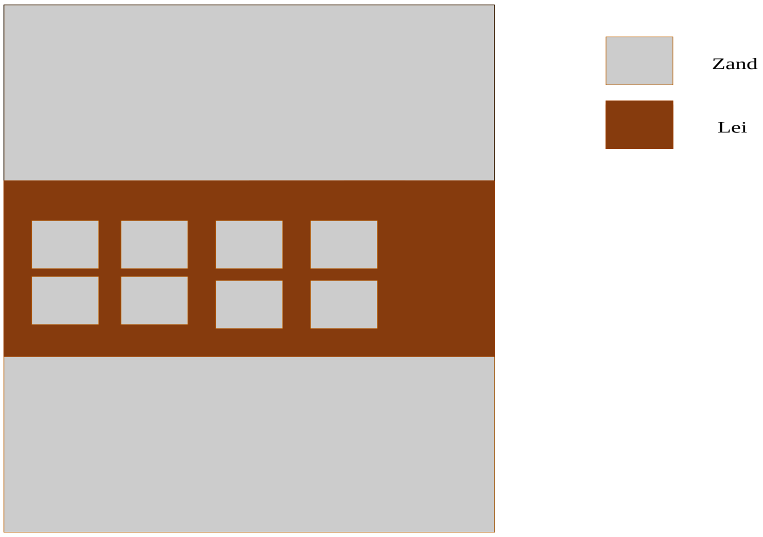
Figuur 7.14: Grafieken in het geval van 7 lagen (links):DICCG methode(tweede benadering) (rechts) DICCG methode (derde benadering)

7.2.3 Afwisselende lagen met gaten

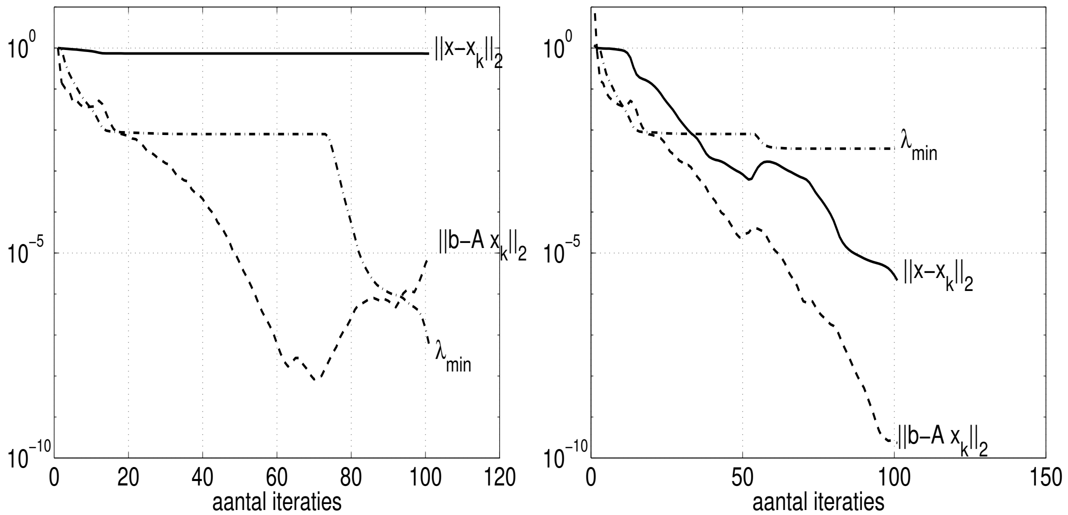
Als laastste toepassing zulen we problemen met 4 (zie figuur (18.15)) zandinsluitingen respectivelijk 8 zandinsluitingen (zie figuur (18.16))bekijken.



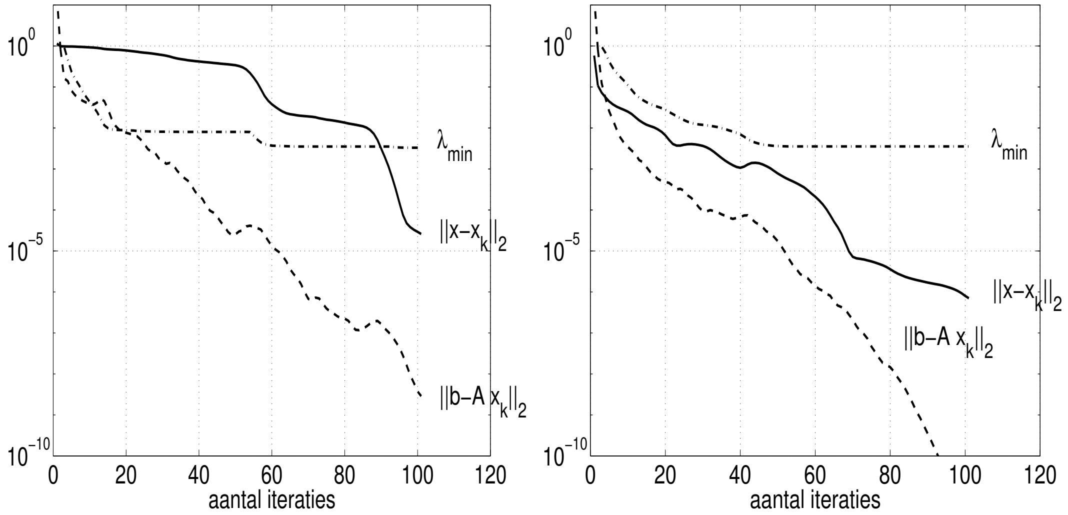
Figuur 7.15: Het gaten probleem met 4 gaten



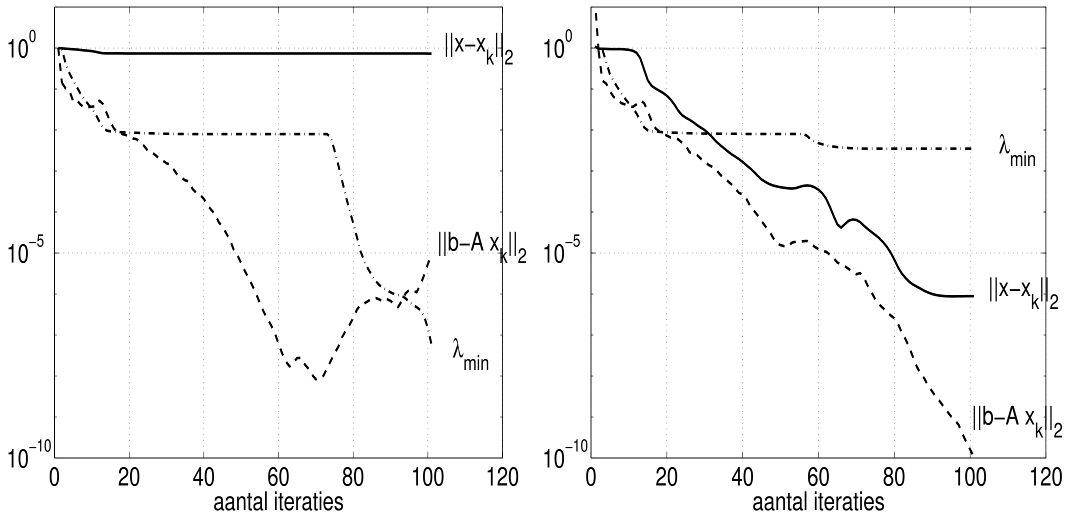
Figuur 7.16: Het gaten probleem met 8 gaten



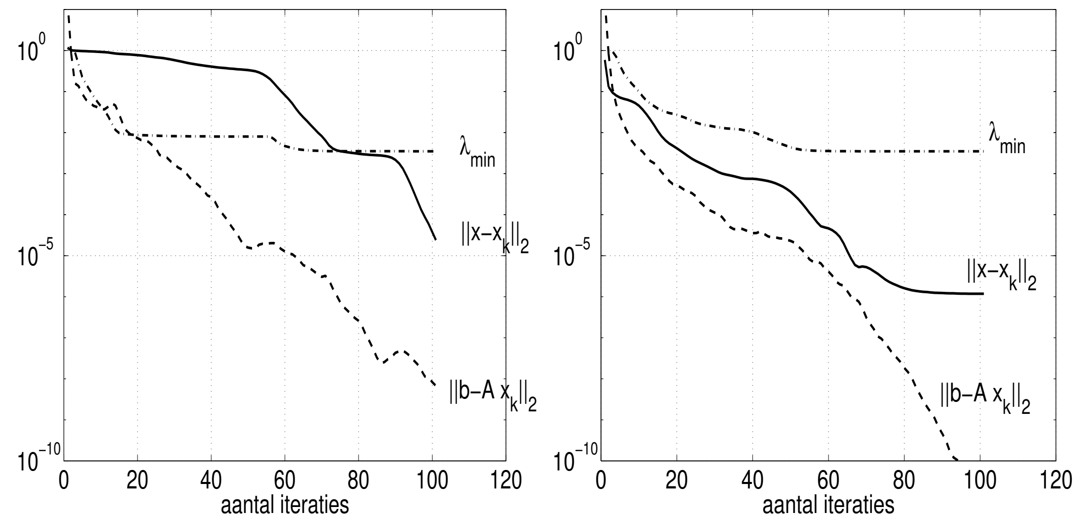
Figuur 7.17: Grafieken in het geval van 4 zandinsluitingen (links):DICCG methode (rechts) DICCG methode (eerste benadering)



Figuur 7.18: Grafieken in het geval van 4 zandinsluitingen (links):DICCG methode (tweede benadering) (rechts) DICCG methode (derde benadering)



Figuur 7.19: Grafieken in het geval van 8 zandinsluitingen(links):DICCG methode (rechts) DICCG methode (eerste benadering)



Figuur 7.20: Grafieken in het geval van 8 zandinsluitingen(links):DICCG methode (tweede benadering) (rechts) DICCG methode (derde benadering)

In de figuren (8.4) tot en met (8.20) worden een aantal grafieken gegeven met daarin de resultaten van deze testen. De fout , het residu en de kleinste effectieve eigenwaarde worden gegeven als functie van het aantal iteraties.

- We kunnen makkelijk zien dat het convergentiegedrag van DICCG methode in het geval van de eerste benadering identiek is aan die in het geval van echte kleine eigenruimte. Uit het vorige hoofdstuk(Gevoeligheidsanalyse)kunnen we dit verklaren door het feit dat de gekozen projectievectoren bij deze benadering een basis van de kleine eigenruimte vormen. Dat impliceert dat in beide gevallen dezelfde projectiematrix wordt gebruikt. In een situatie waarin het aantal lagen oneven is, kunnen we aan de hand van een eenvoudig voorbeeld (zie appendix A6.1) aantonen dat de projectievectoren inderdaad een basis zijn van de kleine eigenruimte.

- In het geval van de tweede constructie krijgen we wat het aantal benodigde iteraties betreft een redelijk verbetering van de convergentie van de DICCG methode in vergelijking met ICCG methode. Deze verbetering in de convergentie kunnen we motiveren doordat de perturbatie van de kleine eigenruimte als gevolg van deze benadering klein genoeg is zodat geen kleine eigenwaarden tot stand kunnen komen. Dit volgt uit het vorige hoofdstuk (Gevoeligheidsanlyse), waarin wordt bewezen dat als de perturbatie zeer klein is, er geen kleine eigenwaarden kunnen optreden. Een relatief slechte convergentie is duidelijk te zien als we de bijbehorende convergentie met die van DICCG methode in het geval van echte kleine eigenruimte vergelijken, is deze wat trager. Want het aantal benodigde iteraties wordt wat groter. Bovendien blijft het residu wat langer hangen.

- In het geval van de derde benadering constateren we dat het convergentiegedrag van de DICCG methode zeer verbeterd is. De convergentie is bijna even goed als die in het geval van de eerste benadering. De bijbehorende grafiek is wat gladder. Dit goede convergentiegedrag kunnen we motiveren door het feit dat de pertubatie in dit geval veel kleiner is dan die bij de tweede benadering. Door de extra projectievectoren op de juiste wijze toe te voegen wordt het effect van de perturbatie in leilagen m.b.t. tweede benadering geannuleerd. De juiste manier van toevoeging betekent dat de toegevoegde projectievectoren overdekken (door enen te kiezen) alle leilagen waar de perturbatie kan optreden. Op deze manier liggen de kleine eigenvectoren bijna in de nieuwe deelruimte opgespannen door de projectievectoren (voor meer toelichting zie het voorbeeld in appendix A3.1). Andere reden voor dit goede convergentiegedrag kan zijn dat de effectieve conditiegetal van 'Deflated' matrix wat kleiner is dan die bij tweede benadering. Dat komt doordat de kleinste niet nul eigenwaarde is van de grootte 10^{-1} is, terwijl voor de andere benadering die van de orde 10^{-2} is.

7.3 Analytisch aanpak

Nu gaan we m.b.v. een analytisch aanpak het convergentiegedrag van DICCG methode in het geval van de derde benadering onderzoeken door een bovengrens van het conditiegetal van de matrix PA te schatten. We zullen hierbij ons beperken tot een rechte lagengebied.

Schatting van de bovengrens van conditiegetal van de matrix PA

Bij de derde benadering worden de projectievectoren $\mathbf{w_i}$, $i = 1...m$ als volgt gedefinieerd:

$$\mathbf{w_i} = \begin{cases} 1 & \text{in } \Omega_i \\ 0 & \text{anders} \end{cases}, \text{voor } i = 1...m$$

De bijbehorende projectiematrix P is de volgende:

$$P = I - W(W^T AW)^{-1} W^T A. \tag{7.1}$$

Hier is $W = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}]$.
Nu gaan we analytisch een bovengrens van het effectieve conditiegetal k_{eff} van de matrix $P^T A$ vinden. Hiertoe zullen we het volgende lemma gebruiken:

Lemma: Laat A symmetrisch positief definiet en P wordt gedefinieerd door 7.1. Neem aan dat er een splitsing bestaat $A = A^* + C$ zodanig dat de matrices A^* en C symmetrisch en positief semidefiniet zijn en bovendien is $span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$ gelijk aan de nulruimte van A^*, dan geldt

$$k_{eff}(L^{-1}PAL^{-T}) \leq \frac{\lambda_n(L^{-1}AL^{-T})}{\lambda_{m+1}(L^{-1}A^*L^{-T})},$$

Hier is $A = LL^T$ een incomplete Cholesky preconditioneringmatrix.
Definities:
$\lambda_i(A)$ is de i^{de} eigenwaarde in de niet dalende ordening van de matrix A.
Het conditiegetal $k_{eff}(A)$ is de verhouding tussen de grooste eigenwaarde en de kleinste eigenwaarde die aan ongelijk nul is, $k_{eff}(A) = \frac{\lambda_n(A)}{\lambda_{m+1}(A)}$.

We hebben gezien in () dat voor een gebied met 2N afwisselende zand-lei lagen (zie figuur 4.7) dat als we een horizontale nummering kiezen en lineaire driehoeken, dan heeft de matrix A de volgende speciale structuur (bloktridiagonal structuur):

$$A = \begin{pmatrix} D_2 & J_1 & & & & & & & & & & & & & & \\ J_1 & D_2 & J_1 & & & & & & & & & & & & & \\ & \ddots & \ddots & \ddots & & & & & & & & & & & & \\ & & J_1 & D_2 & J_1 & & & & & & & & & & & \\ & & & J_1 & D_{1+\epsilon} & J_\epsilon & & & & & & & & & & \\ & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon & & & & & & & & & \\ & & & & & \ddots & \ddots & \ddots & & & & & & & & \\ & & & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon & & & & & & & \\ & & & & & & & J_\epsilon & D_{\epsilon+1} & J_1 & & & & & & \\ & & & & & & & & \ddots & \ddots & \ddots & & & & & \\ & & & & & & & & & \ddots & \ddots & \ddots & & & & \\ & & & & & & & & & & \ddots & \ddots & \ddots & & & \\ & & & & & & & & & & & J_\epsilon & D_{1+\epsilon} & J_1 & & \\ & & & & & & & & & & & & J_1 & D_2 & J_1 & \\ & & & & & & & & & & & & & \ddots & \ddots & \ddots \\ & & & & & & & & & & & & & & J_1 & D_1 \end{pmatrix}.$$

Met:

$$D_\epsilon = \epsilon \begin{pmatrix} a+b & -a & & & \\ -a & 2(a+b) & -a & & \\ & -a & 2(a+b) & \ddots & \\ & & \ddots & \ddots & \\ & & & -a & a+b \end{pmatrix}, \; D_1 = \begin{pmatrix} a+b & -a & & & \\ -a & 2(a+b) & -a & & \\ & -a & 2(a+b) & \ddots & \\ & & \ddots & \ddots & \\ & & & -a & a+b \end{pmatrix},$$

$$J_\epsilon = \epsilon \begin{pmatrix} -b & & & & \\ & -2b & & & \\ & & -2b & & \\ & & & \ddots & \\ & & & & -b \end{pmatrix} \text{ en } J_1 = \begin{pmatrix} -b & & & & \\ & -2b & & & \\ & & -2b & & \\ & & & \ddots & \\ & & & & -b \end{pmatrix}.$$

Hier geldt: $a = \frac{h_1}{2h_2}$en $b = \frac{h_2}{2h_1}$.Voor een definitie van h_1 en h_2 zie figuur 4.6.
De matrix A kunnen we op de volgende wijze splitsen:

$$A = A^* + C.$$

Waarin zijn de blokdiagonale matrices A^* en C gegeven door:

$$A^* = \begin{pmatrix} B_1 & & & & & \\ & B_\epsilon & & & & \\ & & B_1 & & & \\ & & & \ddots & & \\ & & & & B_\epsilon & \\ & & & & & B_1 \end{pmatrix} \text{ en } C = \begin{pmatrix} D_1 & & & & & & \\ & 0 & & & & & \\ 0 & & E_\epsilon & & & & \\ & & & 0 & & & \\ & & & & E_1 & & \\ & & & & & \ddots & \\ & & & & & & E_\epsilon \\ & & & & & & & 0 \end{pmatrix}$$

Gridsize	Aantal lagen	$min\Big(\lambda_2(L_1^{-1}B_1L_1{}^{-T}), \lambda_2(L_\epsilon^{-1}B_\epsilon L_\epsilon^{-T})\Big)$
2×12	4	2.5000e-01
3×16	4	1.4645e-01
4×18	6	1.2500e-01
5×32	8	9.6096e-02

Tabel 7.1: Kleinste effectieve eigenwaarden van de matrices B en B_ϵ

De matrices B_ϵ, B_1, E_ϵ en E_1 worden als volgt gedefinieerd:

$$B_\epsilon = \begin{pmatrix} D_\epsilon & J_\epsilon & & & & & \\ J_\epsilon & D_{2\epsilon} & J_\epsilon & & & & \\ & J_\epsilon & D_{2\epsilon} & J_\epsilon & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & J_\epsilon & D_{2\epsilon} & J_\epsilon & \\ & & & & J_\epsilon & D_{2\epsilon} & J_\epsilon \\ & & & & & J_\epsilon & D_\epsilon \end{pmatrix}, \; B_1 = \begin{pmatrix} D_1 & J_1 & & & & & \\ J_1 & D_2 & J_1 & & & & \\ & J_1 & D_2 & J_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & J_1 & D_2 & J_1 & \\ & & & & J_1 & D_2 & J_1 \\ & & & & & J_1 & D_1 \end{pmatrix}.$$

$$E_\epsilon = \begin{pmatrix} D_\epsilon & J_\epsilon \\ J_\epsilon & D_\epsilon \end{pmatrix} \text{en} E_1 = \begin{pmatrix} D_1 & J_1 \\ J_1 & D_1 \end{pmatrix}.$$

Het is duidelijk te zien dat de matrices A^* en C symmetrisch zijn. Want D_1, D_ϵ, J_1, J_ϵ, B_1, C_1, $B_\epsilon$enC_ϵ zijn symmetrische matrices. In appendix A6.2 wordt bewezen dat de matrices A^* en C positief semidefiniet zijn. Bovendien geldt dat $span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$ gelijk aan de nulruimte van A^* is.
Toepassing van het bovengenoemde lemma geeft:

$$k_{eff}(L^{-1}PAL^{-T}) \leq \frac{\lambda_n(L^{-1}AL^{-T})}{\lambda_{m+1}(L^{-1}A^*L^{-T})}$$

In appendix A6.2 hebben we laten zien dat de matrices B_1 en B_ϵ één nuleigenwaarde hebben, d.w.z $\lambda_1(B_\epsilon) = \lambda_1(B_1) = 0$.
Als preconditionering kiezen we de diagonaalpreconditionering. Daarvoor stellen we

$$L = diag^{-1/2}(A), L_1 = diag^{-1/2}(B_1), \text{en } L_\epsilon = diag^{-1/2}(B_\epsilon).$$

Het is niet moeilijk om in te zien dat:

$$\lambda_{m+1}(L^{-1}AL^{-T}) = min\Big(\lambda_2(L_1^{-1}B_1L_1{}^{-T}), \lambda_2(L_\epsilon^{-1}B_\epsilon L_\epsilon^{-T})\Big),$$

waaruivolgt:

$$k_{eff}(L^{-1}PAL^{-T}) \leq \frac{\lambda_n(L^{-1}AL^{-T})}{min\Big(\lambda_2(L_1^{-1}B_1L_1^{-T}), \lambda_2(L_\epsilon^{-1}B_\epsilon L_\epsilon^{-T})\Big)}.$$

Experimenteel kunnen we laten zien dat de matrix B_1 behalve de nulei-

Gridsize	Aantal lagen	$k_{eff}(L^{-1}PAL^{-T})$	bovenschatting van $k_{eff}(L^{-1}PAL^{-T})$
2×12	4	7.999e+00	8
3×16	4	8.0864e+00	1.3657e+01
4×18	6	8.0468e+00	1.6000e+01
5×32	8	1.6589e+00	2.0813e+01

Tabel 7.2: Bovenschattingen van het effectieve conditiegetal

genwaarde eigenwaarden van de orde 1 heeft (zie tabel (7.1)). Dat geldt uiteraard ook voor de gepreconditioneerde matrix $L_1^{-1}B_1L_1^{-T}$. Aan de andere kant kunnen we ook makkelijk nagaan dat de eigenwaarden van $L_\epsilon^{-1}B_\epsilon L_\epsilon^{-T}$ ook van orde 1 heeft. Hieruit wordt afgeleid dat de effectieve conditiegetal $k_{eff}(L^{-1}PAL^{-T})$ een bovengrens van orde 1. Anderzijds geldt per definitie $k_{eff}(L^{-1}PAL^{-T}) \geq 1$. Aan de hand van deze analyse kunnen we concluderen dat de effectieve conditiegetal $k_{eff}(L^{-1}PAL^{-T})$ van orde 1 is. Dit resultaat geeft een analytische verklaring voor het goede convergentiegedrag van DICCG methode in het geval van derde benadering.

7.4 'IJlheid' van de projectievectoren

Bij de tweede en derde benadering zijn de gekozen projectievectoren ijl. Daarom is het van groot belang om het DICCG algoritme efficiënt te maken door van deze ijlheid gebruik te maken. Later zullen we laten zien dat we hiermee veel rekenwerk en geheugengebruik kunnen besparen. De gebruikte algoritme die we hebben gezien in (5), is de volgende:

DICCG-algoritme
$k = 0, x_0 = 0; r_0 = P^T r_0; p_1 = z_1 = L^{-T}L^{-1}r_0$
while$||r_k|| > \epsilon$ **do**
$k = k + 1;$
$\alpha_k = \frac{(r_{k-1}, r_{k-1})}{(p_k, P^T A p_k)}$
$x_k = x_{k-1} + \alpha_k p_k;$
$r_k = r_{k-1} - \alpha_k P^T A p_k;$
$z_k = L^{-T}L^{-1}r_k$
$\beta_k = \frac{(r_k, z_k)}{(r_{k-1}, z_{k-1})}$
$p_{k+1} = z_k + \beta_k p_k$
endwhile
endDICCG-algoritme

Bij iedere iteratieslag van DICCG algoritme dient de term $P^T A\mathbf{p_k}$ berekend te worden. Omdat de matrix A groot kan zijn en ook kunnen het aantal projectievectoren groot zijn, kan de berekingen van deze term veel rekenwerk en geheugenruimte kosten. Daarvoor zullen we deze term berekenen op een efficiente wijze door van de ijlheid van de matrix A en projectievectoren $\mathbf{w_i}$ gebruik te maken.

We noteren de vector $\mathbf{p_k}$ met $\mathbf{p}$.

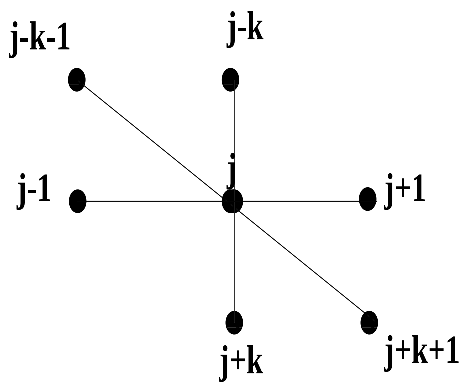
$$P^T A\mathbf{p} = A\mathbf{p} - (AW)E^{-1}W^T A\mathbf{p}. \tag{7.2}$$

De term $P^T A\mathbf{p}$ kan efficient berekend worden in vier fasen:

- In de eerste fase wordt het matrix-vector product $A\mathbf{p}$ berekend.

$$A\mathbf{p} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}p_j \\ \sum_{j=1}^{n} a_{2j}p_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj}p_j \end{pmatrix}.$$

  We stellen $\mathbf{x} = A\mathbf{p}$. In het algemeen zal $\mathbf{x}$ niet ijl zijn. Want de



Figuur 7.21:

  vector $\mathbf{p}$ is niet ijl. Door e.e.m discritisatie is de matrix A sparse. Voor efficientie worden slechts de niet nulelementen van de matrix A opgeslagen in het geheugen. Voor iedere onbekende in de mesh met nummer i definieren we de vezameling I_i van indexen j waarvoor geldt $a_{ij} \neq 0$. Uit de figuur(7.21) is te zien dat in het geval van de keuze van driehoekelementen bij discretizatie bestaat de verzameling $|I_i|$ uit maximaal 7 indices.
  De berekingen van iedere component x_i van $\mathbf{x}$ kan goedkoper gemaakt worden

$$x_i = \sum_{j \in I_i} a_{ij}p_j.$$

  De berekingen van $\mathbf{x}$ per element kost $2|I_i|$ operaties, dus in totaal $\sum_{i=1}^{n} 2|I_i|$, dus maximaal $14n$. $|I_i|$ is het aantal element van I_i. Voor het opslaan van de matrix A hebben we maximaal $\sum_{i=1}^{n} |I_i|$, dus maximaal $7n$ geheugen posities nodig.

- In de tweede fase wordt $W^T\mathbf{x}$ berekend. In het geval van de tweede en derde benadering bestaan de projectievectoren $\mathbf{w_i}$ voor $i = 1...m$ uit veel nullen. We definiëren voor iedere projectievector $\mathbf{w_i}$ de verzameling J_i van indices waarvan het bijbehorende elementen ongelijk aan nul zijn.

$|J_i|$ is het bijbehorende aantal elementen. In ons geval is die precies gelijk aan het aantal punten in een laag Ω_i.

$$\mathbf{y} = W^T \mathbf{x} = \begin{pmatrix} \mathbf{w_1}^T \mathbf{x} \\ \mathbf{w_2}^T \mathbf{x} \\ \vdots \\ \mathbf{w_m}^T \mathbf{x} \end{pmatrix}.$$

De berekening van iedere element $\mathbf{w_i^T x}$ van $\mathbf{y}$ kan gereduceerd worden tot de berekingen van $\sum_{j \in J_i} w_{ij} x_j$. Om die reden gaan we alleen de niet nulelementen van $\mathbf{w_i}$ in het geheugen opslaan. Het aantal benodigde operaties voor het matrix-vector product $W^T \mathbf{x}$ is $\sum_{i=1}^{m} |J_i|$. Voor de opslag van de projectievectoren $\mathbf{w_i}$ hebben we dus $\sum_{i=1}^{m} |J_i|$ geheugen posities nodig.

- In de derde fase gaan we $\bar{W} = AW$ berekenen. W is een matrix die uit m projectievectoren $\mathbf{w_i}$ bestaat. We hebben:

$$\bar{W} = [A\mathbf{w_1} A\mathbf{w_2} ......A\mathbf{w_m}].$$

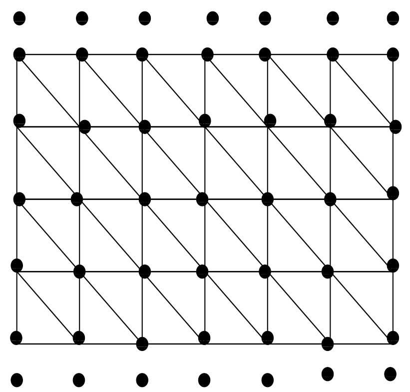
We stellen $\mathbf{\bar{w}_i} = A\mathbf{w_i}$ voor $i = 1..m$.
De berekeningen van $\mathbf{\bar{w}_i}$ kan herleid worden tot

$$\mathbf{\bar{w}_i} = A\mathbf{w_i} = \begin{pmatrix} \sum_{j \in I_1} a_{1j} w_{ij} \\ \sum_{j \in I_2} a_{2j} w_{ij} \\ \vdots \\ \sum_{j \in I_n} a_{nj} w_{ij} \end{pmatrix}.$$

Nu gaan we onderzoeken voor het geval van de tweede en derde benadering wanneer een component van $\mathbf{\bar{w}_i}$ gelijk aan nul zal zijn.
In het geval van de tweede-en de derde benadering kan de berekening van iedere k^{er} component van $\mathbf{\bar{w}_i}$ herleid worden tot

$$\sum_{j \in (I_k \cap J_i)}^{n} a_{kj}.$$

We kunnen makkelijk zien dat voor ieder punt met nummer k van een laag Ω_i er een punt uit dezelfde laag is te vinden die verbonden is met dit punt, d.w.z. $a_{kj} \neq 0$, dus geldt dat $(I_k \cap J_i)$ niet leeg is. Anderzijds geldt voor een laag Ω_i dat als een punt k die niet verbonden is met ieder punt met nummer j uit deze laag via een driehoekelement geldt $a_{kj} = 0$. Derhalve is $I_k \cap J_i$ leeg. De punten van andere lagen die niet verbonden zijn met de randpunten van deze laag Ω_i hebben deze eigenschap, m.a.w de indices van deze punten corresponderen met de indices van de elementen waarvan de vector $\mathbf{\bar{w}_i} = A\mathbf{w_i}$ nul is.
Op basis daarvan kunnen concluderen dat het aantal niet nulelementen van een projectievector $A\mathbf{w_i}$ gelijk is aan het aantal punten in de bijbehorende laag of gebied plus het aantal punten die verboden zijn met de randpunten van deze laag of dit gebied (zie figuur (8.22)).



Figuur 7.22: Schets van niet nulpunten behorend bij projectietvector $A\mathbf{w_i}$

We definieren de verzameling $\bar{J}_i$ van de indexen van de componenten van een vector $\bar{\mathbf{w}}_\mathbf{i} = A\mathbf{w_i}$ die ongelijk aan nul zijn. Het benodigde aantal operaties per component is maximaal gelijk aan 7, dus in total $7|\bar{J}_i|$. Het benodigde rekenwerk voor de uitvoering van AW is $\sum_{i=1}^{m} 7|\bar{J}_i|$.

- In de vierde fase wordt $E^{-1}\mathbf{y}$ berekend. $\mathbf{y}$ is een vector van lengte m. In de eerste stap wordt de matrix E berekend:

$$E = (AW)^T W = \begin{pmatrix} \bar{\mathbf{w}}_\mathbf{1}^T \mathbf{w_1} & \bar{\mathbf{w}}_\mathbf{1}^T \mathbf{w_2} & \cdots & \bar{\mathbf{w}}_\mathbf{1}^T \mathbf{w_m} \\ \bar{\mathbf{w}}_\mathbf{2}^T \mathbf{w_1} & \bar{\mathbf{w}}_\mathbf{2}^T \mathbf{w_2} & \cdots & \bar{\mathbf{w}}_\mathbf{2}^T \mathbf{w_m} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{\mathbf{w}}_\mathbf{m}^T \mathbf{w_1} & \bar{\mathbf{w}}_\mathbf{m}^T \mathbf{w_2} & \cdots & \bar{\mathbf{w}}_\mathbf{m}^T \mathbf{w_m} \end{pmatrix}.$$

Voor de praktische redenen gaan we voor het opslaan van een projectievector $\mathbf{w_i}$ de verzameling $\bar{J}_i$ gebruiken in plaats van J_i. Dat is mogelijk, omdat we hebben gezien dat J_i een deelverzameling is van $\bar{J}_i$.
-De berekeningen van een element $e_{ij} = \bar{\mathbf{w}}_\mathbf{i}^T \mathbf{w_j}$ van de matrix E kost $|\bar{J}_i|$ operaties. We kunnen makkelijk zien dat voor een projectievector $\mathbf{w_i}$ behorend bij een laag Ω_i geldt een projectievector $\mathbf{w_j}$ behorend bij de buurlagen van de laag Ω_i voldoet aan $e_{ij} = \bar{\mathbf{w}}_\mathbf{i}^T \mathbf{w_j} \neq 0$ en omgekeerd. In de situatie van afwisselende lagen heeft een laag alleen met twee buurlagen te maken. Als gevolg daarvan is de matrix E in de gevallen tweede en derde benadering diagonaal respectivelijk tridiagonaal. Het hoeveelheid rekenwerk daarvoor is dus $\sum_{i=1}^{m} |\bar{J}_i|$ als de matrix E diagonaal is en $\sum_{i=1}^{3m} |\bar{J}_i|$ als de matrix E tridiagonaal is.
-Het hoeveelheid rekenwerk voor de bepaling van de inverse van een tridiagonale matrix E door LU-decompositie te gebruiken is $8m^2 - 6m$. De bepaling van de inverse van een diagonale matrix E kost m operaties. De matrix E^{-1} is een m bij m volle matrix, indien de matrix E tridiagonal is en m .De hoeveelheid geheugenruimte voor de opslag daarvan is m^2 .Als de matrix E diagonaal is de hoeveelheid geheugenruimte m geheugen posities.

- In de vijfde fase leidt de berekening van de matrix-vector produkt $E^{-1}\mathbf{y}$ tot een vector $\mathbf{z}$, $\mathbf{z} = E^{-1}\mathbf{y}$. Deze bewerkingen kosten $3m$ operaties voor een tridiagonale matrix en m operaties voor een diagonale matrix.

- Als laatste fase wordt de volgende berekend:

$$\mathbf{s} = P^T A\mathbf{p} = Ap - \bar{W}\mathbf{z} = \mathbf{x} - \begin{pmatrix} \bar{\mathbf{w}}_\mathbf{1} & \bar{\mathbf{w}}_\mathbf{2} & \cdots & \bar{\mathbf{w}}_\mathbf{m} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix}.$$

  Verder uitgewerkt:

$$\mathbf{s} = P^T A\mathbf{p} = \mathbf{x} - z_1\bar{\mathbf{w}}_\mathbf{1} - z_2\bar{\mathbf{w}}_\mathbf{2} - ... - z_m\bar{\mathbf{w}}_\mathbf{m}.$$

  Het is niet moeilijk na te gaan dat het vereiste hoeveelheid rekenwerk $\sum_{i=1}^{m} 2|\bar{J}_i|$ is.

We noteren $|J_i|$ met n_i.
Alle resultaten betreffende rekenwerk en geheugenruimte van verschillende fases worden in de volgende tabellen aangegeven:

Bewerkingen	Rekenwerk(3 benadering)	Rekenwerk(2 benadering)
$\mathbf{x} = A\mathbf{p}$	$14n$	$14n$
$\mathbf{y} = W^T\mathbf{x}$	$\sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} n_i$
$\bar{W} = AW$	$\sum_{i=1}^{m} 7n_i$	$\sum_{i=1}^{m} 7n_i$
$E = (\bar{W}^T W)$ en E^{-1}	$8m^2 + \sum_{i=1}^{m} 3n_i - 6m$	$m + \sum_{i=1}^{m} n_i$
$\mathbf{z} = E^{-1}\mathbf{y}$	m	$3m$
$\mathbf{s} = \mathbf{x} - \bar{W}\mathbf{z}$	$\sum_{i=1}^{m} 2n_i$	$\sum_{i=1}^{m} 2n_i$
$P^T A\mathbf{p}$	$8m^2 - 3m + 12\sum_{i=1}^{m} n_i + 14n$	$2m + 10\sum_{i=1}^{m} n_i + 14n$

Tabel 7.3: Resultaten van het rekenwerk behorend bij tweede benadering en derde benadering

Opslaan	Geheugenposities(3 benadering)	Geheugenposities(2 benadering)
De matrix A	$7n$	$7n$
De matrix W	$\sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} n_i$
De matrix $\bar{W}$,	$\sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} n_i$
De matrix E^{-1}	m^2	m
Totaal	$m^2 + 2\sum_{i=1}^{m} n_i + 7n$	$m + 2\sum_{i=1}^{m} n_i + 7n$

Tabel 7.4: Resultaten van het geheugenruimte behorend bij tweede benadering en derde benadering

7.5 Conclusies

Uit toepassingen van de tweede en derde benadering op verschillende structuren van de ondergrond kunnen we de volgende conclusies trekken:

- Met zowel de tweede benadering als de derde benadering van de kleine eigenruimte kunnen geen kleine eigenwaarden optreden. Dat komt omdat de perturbatie van de kleine eigenruimte klein genoeg is. Dit leidt tot een verbetering van de effectieve conditie van de 'Deflated' matrix. Voor beide benaderingen zien we daarom een merkbare verbetering van het convergentiegedrag van de DICCG methode, vooral in het geval van derde de benadering.

- De tweede benadering toont een redelijk goed convergentiegedrag. Die is wat efficiënter dan de derde benadering als het aantal lagen groter is dan het aantal optredende 'fysische' kleine eigenwaarden. Want de totale rekenwerk en geheugengebruik hangen kwadratisch van het aantal lagen af (zie tabels (7.3)en (7.4)). Een groot nadeel van de tweede benadering is dat voor de constructie van de projectievectoren moet men bepalen welke laag als zandlaag of leilaag beschouwd moet worden. In de situatie van moeilijke structuren van de grond lijkt het detecteren van de lagen geen gemakkelijke zaak te zijn. Dat kan er eventueel voor zorgen dat deze benadering niet toepaasbaar is.

- De derde benadering levert de beste resultaten op qua het aantal iteraties, en de conditie van de 'Deflated' matrix. Een andere Belangrijk voordeel hiervan is dat we geen zandlaag of leilaag hoeven te detecteren voor de constructie van de projectievectoren die de fysische eigenwaarden elimineren. Om die reden is deze het meest aantrekkelijk. Het aantal projectievectoren hierbij is precies gelijk aan het aantal lagen. Dat kan een nadeel zijn als het aantal lagen zeer hoog is. In deze situatie wordt de benadering vanwege grote rekenwerk en geheugenruimte wordt niet efficiënt.

Hoofdstuk 8

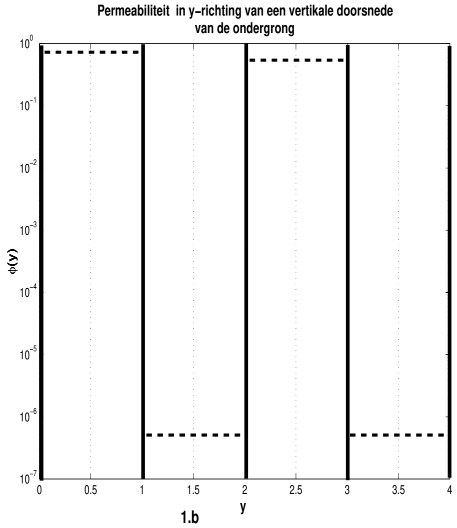
Variatie van de permeabiliteit in de lagen

8.1 Inleiding

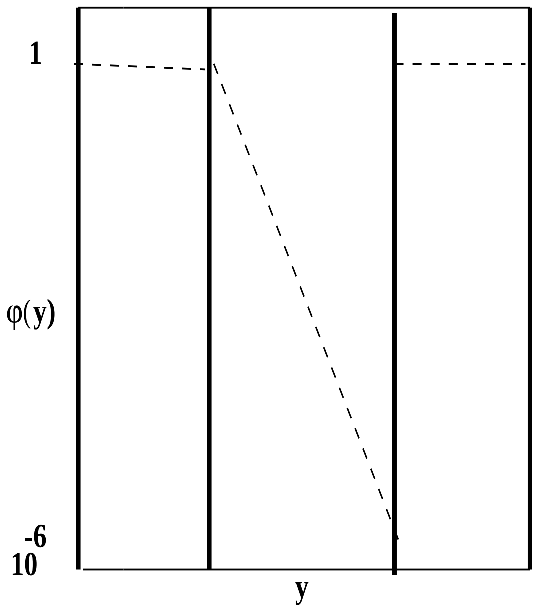
In de vorige hoofdstukken werdt de permeabiliteit in de leilaag en zandlaag constant genomen. In dit hoofdstuk gaan we de permeabiliteit varieren in de lagen. Bij dit onderzoek gaan we er van uit dat de permeabiliteit als een continue functie in een laag varieert. Uit vorige onderzoeken is gebleken dat de 'sterke' discontinuiteit op de randen (van lage permeabiliteit naar hoge permeabiliteit en omgekeerd) de belangrijkste oorzaak is van het ontstaan van kleine eigenwaarden. Het is daarom van belang de problemen te onderzoeken , waarin de discontinuiteit van de permeabiliteit op de randen optreedt. Hiertoe gaan we het volgende basisprobleem bekijken: een laag die ingesloten wordt door twee of meer andere lagen. We gaan de permeabiliteit in de ingesloten laag of in één van zijn buurlagen zo sterk varieren dat er een sterke discontinuiteit op de randen van deze lagen met één van zijn buurlagen is. Hierbij wordt onderscheid gemaakt tussen twee variaties van permeabiliteiten in de betreffende laag, namelijk: de permeabiliteit varieert van hoog permeabel naar laag permeabel en omgekeerd.

We hebben gezien dat een zandlaag met wel of niet Neumann randvoorwaarden , die ingesloten is door leilagen , aanleiding geeft tot het optreden van kleine eigenwaarden (zie voorbeeld in figuur (9.1)). Om die reden zullen we op analoge wijze als in het geval van constante permeabiliteit in de lagen de volgende speciale situaties van problemen bekijken:

- Voor de eerste situatie varieren we de permeabiliteit in de ingesloten zandlaag. We houden de permeabiliteit constant in de buurlagen. De buurlagen worden als leilagen gekozen (zie figuur(9.2)). In deze situatie gaan we onderzoeken het effect van de combinatie van discontinuiteit op één van zijn randen van de ingesloten laag en de sterke variatie hierbinnen .

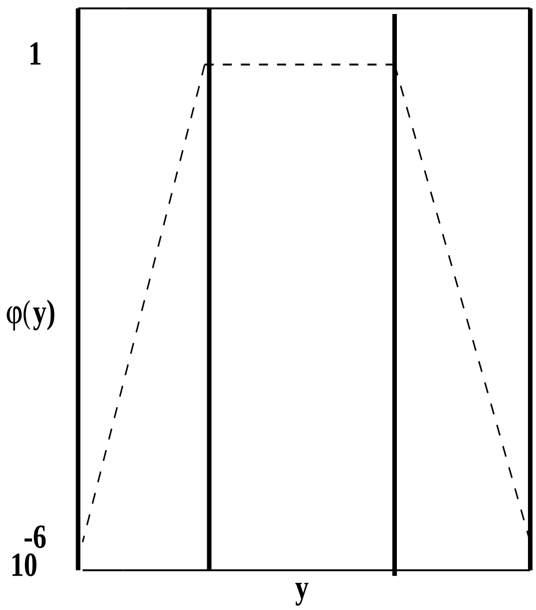


Figuur 8.1: :vier lagen probleem zonder variatie van permeabiliteit



Figuur 8.2:

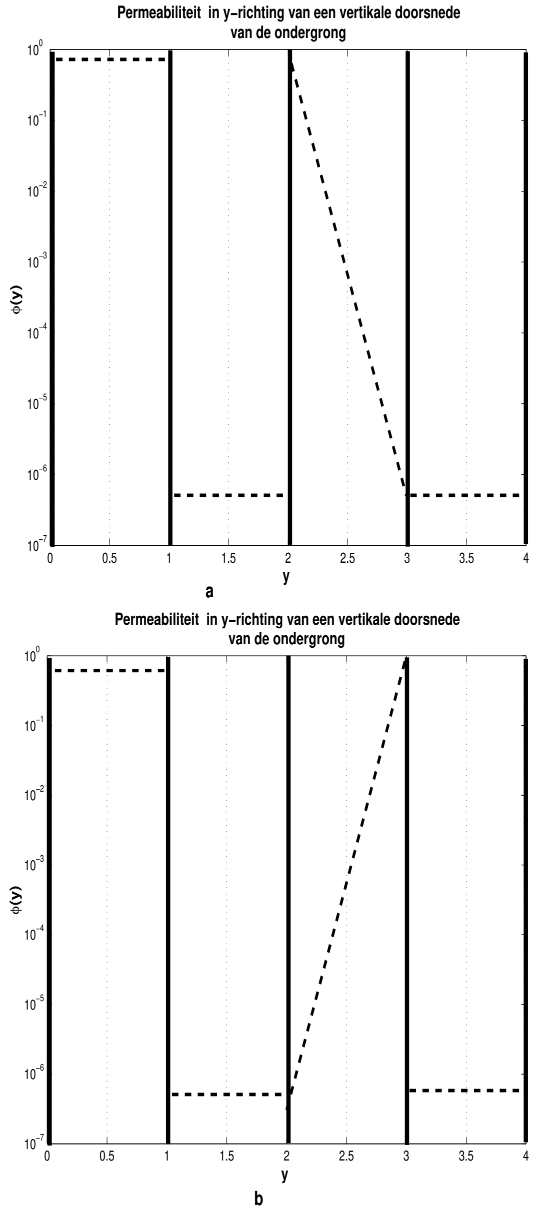
- In de tweede situatie varieert de permeabiliteit als een continue functie in de deelgebied, die uit de ingesloten laag en zijn buurlagen bestaat. Op deze manier hebben we geen discontinuiteit op randen van de ingesloten laag. In de ingesloten laag houden we de permeabiliteit constant, terwijl in de buurlagen zullen we de permeabiliteit sterk varieren (zie figuur (9.3)). In deze situatie gaan we de invloed van sterke variatie van permeabiliteit in buurlagen van de ingesloten laag onderzoeken.



Figuur 8.3:

Voor beide situaties gaan we bekijken of de ingesloten laag als een zandlaag kan beschouwd worden. Eerst gaan we controleren of de discontinuiteit behorend bij deze situaties tot het optreden van fysische kleine eigenwaarden kan leiden. Als gevolg daarvan zal de ICCG methode slecht convergeren. Als dit het geval is, zullen we de DICCG methode hierop toepassen. Daarvoor gaan we de tweede en de derde benaderingen van de kleine eigenruimte op de vermelde basisproblemen gebruiken.
Om de analyse eenvoudiger te maken zullen we ons bij onze experimenten beperken tot het geval van rechte afwisselende zand en lei lagen. Om de sterke variatie te realiseren wordt de permeabiteit in de betreffende laag als een exponentiele functie gevarieerd.

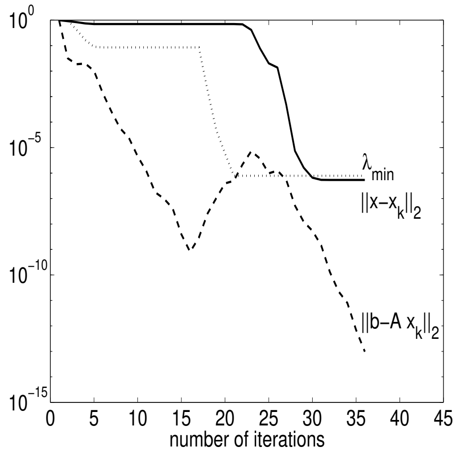
8.2 Eerste situatie



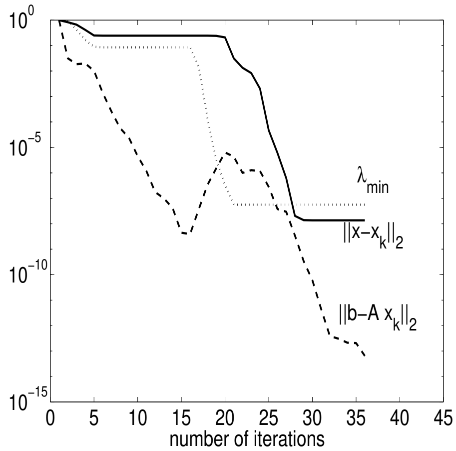
Figuur 8.4: 9.4a:variatie van hoog permeabel naar laag permeabel, 9.4b :variatie van laag permeabel naar hoog permeabel

In eerste instantie zullen we twee basisproblemen bekijken: bij het eerste probleem varieren we exponentieel de permeabiliteit van hoog naar laag in de ingesloten laag (derde laag) van hoog naar laag (zie figuur (9.4a)). Omgekeerd bij het tweede probleem varieren we exponentieel de permeabiliteit van laag naar hoog in de ingesloten laag (derde laag) van laag naar hoog(zie figuur (9.4b)).

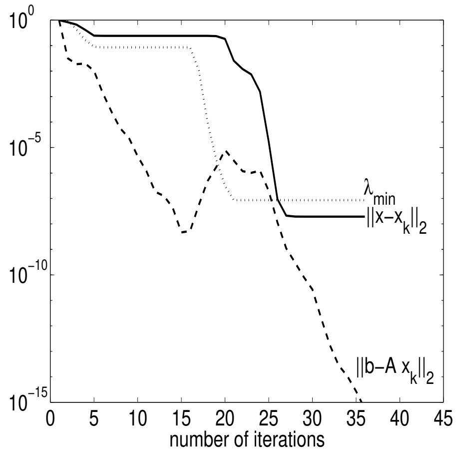
De ICCG methode



Figuur 8.5: Het convergentie gedrag van ICCG methode in het geval van zonder variatie

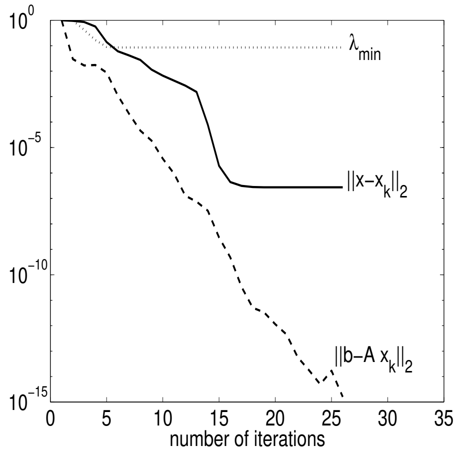


Figuur 8.6: Het convergentie gedrag van ICCG methode in het geval (9.4a)

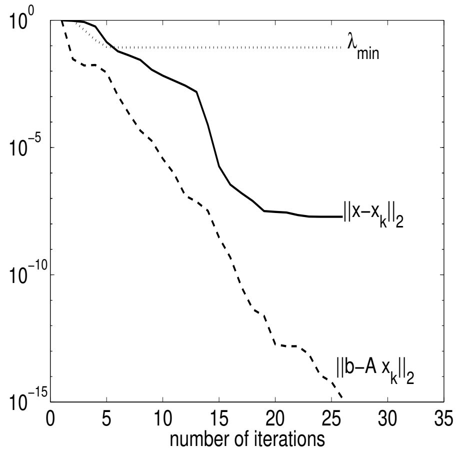


Figuur 8.7: Het convergentie gedrag van ICCG methode in het geval (9.4b)

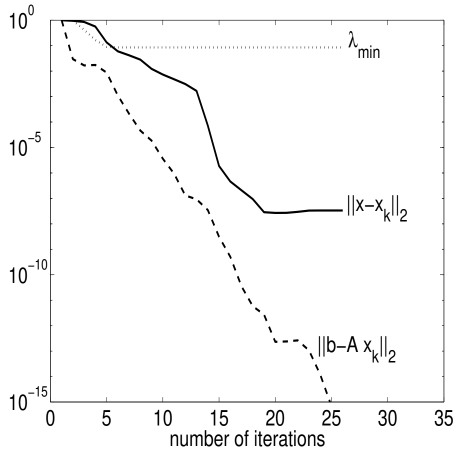
De DICCG methode (eerste benadering)



Figuur 8.8: Het convergentiegedrag van DICCG methode in het geval 4 lagen zonder variatie

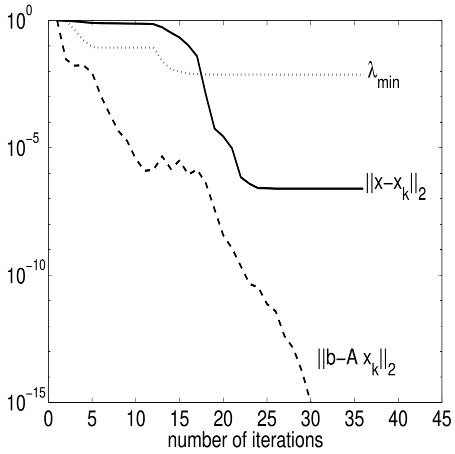


Figuur 8.9: Het convergentiegedrag van DICCG methode in het geval (9.4a)

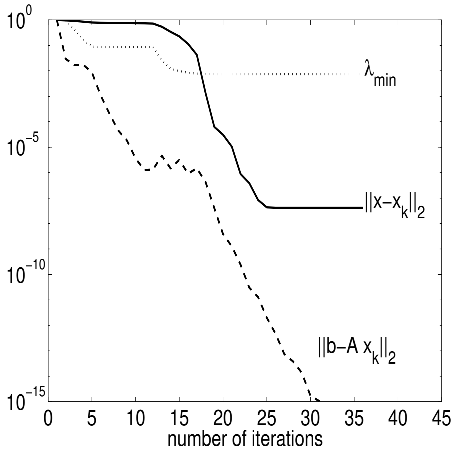


Figuur 8.10: Het convergentiegedrag van DICCG methode in het geval (9.4b)

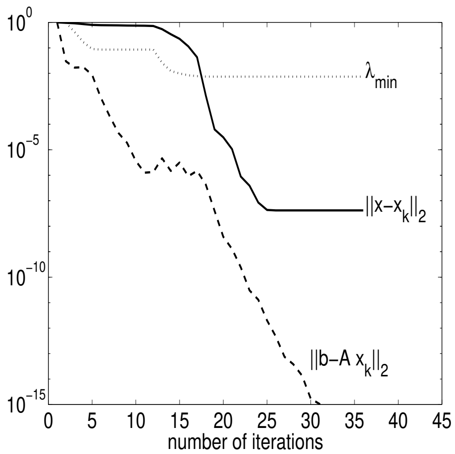
De DICCG methode (tweede benadering)



Figuur 8.11: Het convergentiegedrag van DICCG methode in het geval 4 lagen zonder variatie

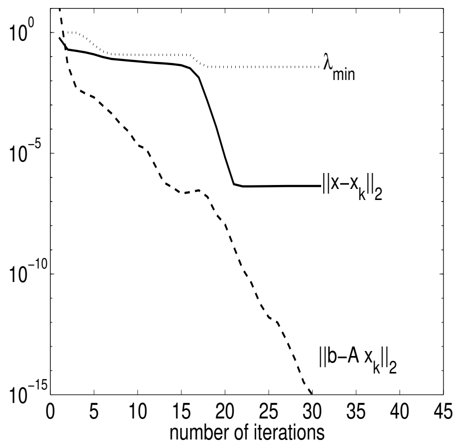


Figuur 8.12: Het convergentiegedrag van DICCG methode in het geval 4 lagen (9.4a) met variatie van hoog naar laag

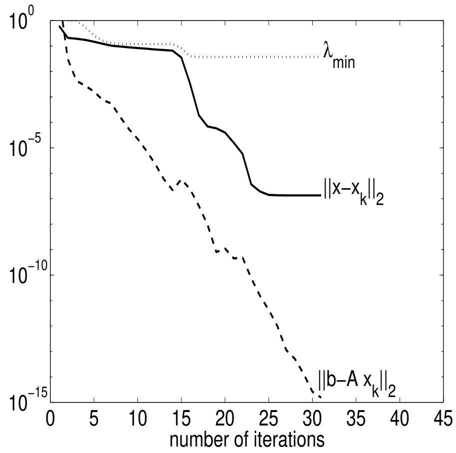


Figuur 8.13: Het convergentiegedrag van DICCG methode in het geval 4 lagen (9.4b) met variatie van laag naar hoog

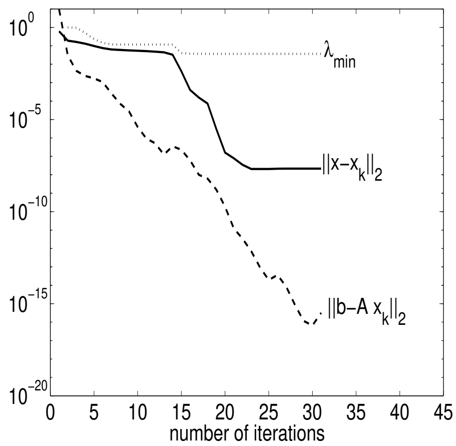
De DICCG methode (Derde benadering)



Figuur 8.14: Het convergentiegedrag van DICCG methode in het geval 4 lagen zonder variatie



Figuur 8.15: Het convergentiegedrag van DICCG methode in het geval 4 lagen (9.4a) met variatie van hoog naar laag



Figuur 8.16: Het convergentiegedrag van DICCG methode in het geval 4 lagen (9.4b) met variatie van laag naar hoog

In de grafieken (9.5) tot met (9.7)wordt het convergentiegedrag van ICCG geplot voor respectivelijk constante permeabiliteit, exponentiele permeabiliteit in de leilaag van hoog naar laag (9.4a) en exponentiele permeabiliteit in de leilaag van laag naar hoog (9.4b). De grafieken (9.8) tot met(9.10) laten het convergentiegedrag zien van DICCG volgens onze eerste benadering, terwijl de grafieken (9.11) tot met(9.13) het convergentiegedrag volgens onze tweede benadering en de grafieken (14) tot met(16) het convergentiegedrag volgens onze derde benadering.
Uit deze grafieken (9.5) tot en met (9.16) kunnen we constateren:

- In het geval van variatie van permeabiliteit van hoog naar laag permeabel of omgekeerd vertoont de ICCG methode bijna dezelfde convergentiegedrag als in het geval van constante permebeliteit. Dit convergentiegedrag kan verklaard worden door het feit dat een combinatie van een sterke continue variatie van de permeabiliteit en de discontinuiteit op één van de randen met zijn buurlagen de oorzaak is van het optreden van kleine eigenwaarden.

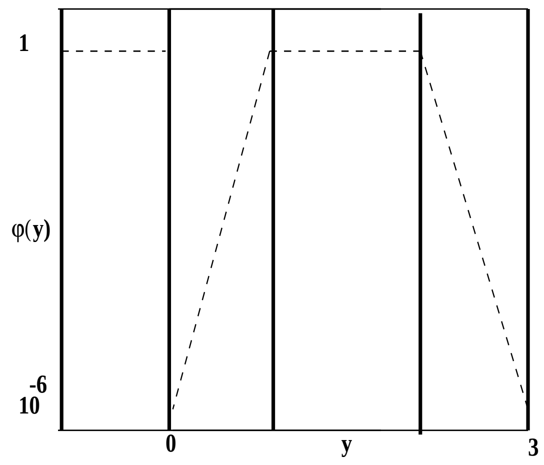
- Zoals in de situatie van constante permeabiliteit levert de eerste benadering, de tweede benadering en de derde benadering van kleine eigenruimte in het geval van variatie van permeabiliteit een aanzienlijk verbetering van het convergentiegedrag van DICCG methode. Dit komt omdat de ingesloten laag waarin aan de éne kant de permeabiliteit sterk varieert en aan de andere kant sterke discontinuiteit op één van zijn randen optreedt, gedrag zich als een zandlaag met constante permeabiliteit. Op deze manier kunnnen we de ingesloten laag beschouwen als een zandlaag met constante permeabiliteit. Als gevolg daarvan zijn de eerste benadering en de tweede benadering toepasbaar in de eerste situatie.

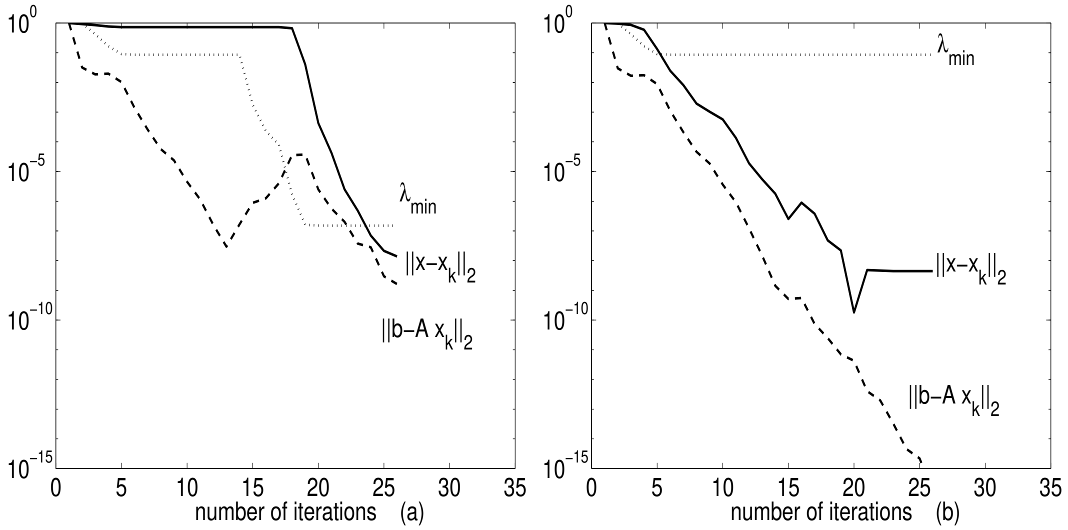
8.3 Tweede situatie

Het bijbehorende basisprobleem is een probleem met vier lagen (zie figuur). De derde laag is de ingesloten laag. De tweede laag en de vierde laag zijn de buurlagen. In de ingesloten laag (tweede laag) is de permeabiliteit constant genomen. In de tweede laag gaan we de permeabiliteit varieren van exponentieel laag naar hoog. In de vierde laag gaan we de permeabiliteit omgekeerd varieren van exponentieel hoog naar laag.
Het convergentiegedrag van ICCG wordt in de grafiek (9.17a) geplot. De grafieken (9.17b) ,(9.18a) resp.(9.18b) laten het convergentiegedrag van DICCG zien volgens onze eerste benadering, tweede benadering respectievelijk onze derde benadering.
Uit deze bovengenoemde grafieken kunnen we constateren:

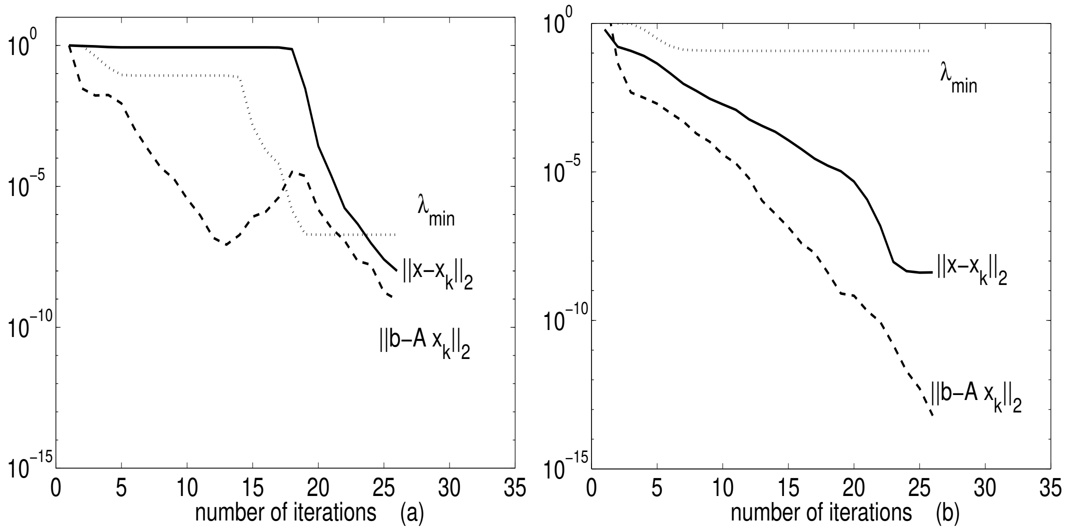
- Een sterke continue variatie van permeabiliteit in buurlagen van de ingesloten laag, zonder dat hij discontinuiteit op zijn randen heeft, leidt tot het optreden van de kleine eigenwaarden. Dit verklaart de slechte convergentiegedrag van de ICCG methode in figuur(9.17a).

- We kunnen uit figuren (9.17b) en (9.18b) ook duidelijk zien dat wat het convergentiegedrag betreft levert de DICCG methode in het geval van zowel de eerste benadering als de derde benadering ook in de tweede situatie goede resultaten. Opvallend is zoals te zien in figuur(9.18a) dat het convergentie gedrag van de DICCG methode in het geval van de tweede benadering lijkt op die van de ICGG methode. Dat houdt in dat in tegenstelling tot de eerste en de derde benadering leidt de tweede benadering niet tot het verwijderen van de kleine eigenwaarden. Dus mogen we de ingesloten laag niet als een zandlaag beschouwen met constante permeabiliteit.





Figuur 8.17: (a): ICCG methode, (b) :DICCG methode(eerste benadering)



Figuur 8.18: (a): DICCG methode (tweede benadering), (b) :DICCG methode(derde benadering)

8.4 Conclusies

- In beide situaties kunnen kleine eigenwaarden tot stand komen. Hierbij hadden we met sterke variatie van de permeabiliteit in de lagen te maken. Hieruit is gebleken dat de sterke variatie van de permeabiliteit in de lagen een belangrijke rol speelt bij het optreden van de kleine eigenwaarden.

- De tweede benadering van de kleine eigenruimte kan wel in de eerste situatie, maar niet in de tweede situatie toepasbaar zijn . De toepassing hiervan blijkt dus afhankelijk te zijn van het verloop van permeabiliteit in de ondergrond. We kunnen suggeren dat de perturbatie van de kleine eigenruimte in deze situatie groot is , zodat het optreden van een kleine eigenwaarde plaatsvindt. Deze perturbatie is een mogelijke oorzaak van de slechte convergentie gedrag van de DICCG methode gebasseerd op tweede benadering.

- De eerste benadering en de derde benadering blijven een goede convergentiegedrag in beide situaties leveren. Hieruit komen we tot het suggestie dat een nauwkeurige constructie van de kleine eigenruimte betreffende constante permeabiliteit in de lagen (bijvoorbeeld eerste benadering en derde benadering) ook gebruikt kan worden voor de benadering van de kleine eigenruimte betreffend beide situaties, m.a.w de perturbatie van de kleine eigenruimte behorend bij de eerste benadering en de tweede benadering blijft zeer klein.

Hoofdstuk 9

Conclusies en aanbevelingen

Conclusies

In het afstudeerwerk is onderzocht of de DICCG methode gevoelig is voor de benaderingen van de kleine eigenruimte en in hoeverre de constructies van de projectievectoren DICCG beïnvloeden. Hierbij komen twee sparse constructies aan de orde. De eerste constructie is aangeduid met de tweede benadering en tweede constructie met derde benadering. Hierbij is ons onderzoek gericht op de volgende onderwerpen met de bijbehorende conclusies.

1. Het analyseren van het effect van de keuze van deflatievectoren op het tot standkomen van nieuwe kleine eigenwaarden. Analytisch wordt geprobeerd aan te tonen of de bijbehorende perturbatie van de kleine eigenruimte de oorzaak is van de 'geperturbeerde' kleine eigenwaarden. Hieruit zijn de volgende conclusies gekomen:
   - Een grote perturbatie veroorzaakt het optreden van nieuwe kleine eigenwaarden.
   - In tegenstelling tot grote perturbatie leidt een kleine perturbatie niet tot het ontstaan van nieuwe kleine eigenwaarden.
2. Het toepassen van deze nieuwe constructies, namelijk tweede benadering en de derde benadering op verschillende structuren van de ondergrond. Daarnaast wordt een analytisch onderzoek gedaan naar derde benadering. Hierbij wordt dit beperkt tot de situatie van afwisselend rechte lagengebied. Door deze constructies toe te passen op drie relevante verschillende structuren van de ondergrond, te weten rechte lagengebied, kromme lagengebied en lagengebied met gaten is gebleken:
   - De DICCG methode die gebaseerd is op tweede benadering laat zien een merkbare verbetering van het gedrag van het residu en de fout. De perturbatie die als gevolg komt van benadering van de kleine eigenruimte is zeer gering. Het grote voordeel daarvan is dat deze constructie van de projectievectoren goedkoper wordt qua rekenwerk en geheugenruimte. Daarmee wordt de DICCG methode

in die zin zeer efficiënt. Het detecteren van lagen maakt echter deze methode niet praktisch en soms niet toepaasbaar.

- De DICCG methode die gebaseerd is op derde benadering vertoont het beste convergentiegedrag. Deze methode is echter wat minder efficiënt, wat rekenwerk en geheugenruimte, indien het aantal lagen hoog is. Behalve het goede convergentiegedrag heeft ook een andere voordeel dat het detecteren van lagen niet nodig is.

3. Het testen van DICCG methode gebaseerd op nieuwe constructies. Maar nu wordt DICCG toegepast op lagengebieden waarin de permeabiliteit zeer sterk varieert in de lagen. Hierbij wordt slechts aandacht besteed aan het afwisselend rechte lagengebied. Uit numerieke experimenten is gebleken:

   - De DICCG methode in het geval van de tweede benadering heeft een slecht convergentiegedrag. De sterke variatie van de permeabiliteit in lagen maakt de tweede benadering in een bepaalde situatie niet toepaa sbaar, en maakt het detecteren van lagen wat ingewikkelder.
   - De DICCG methode in het geval van de derde benadering blijft het convergentiegedrag te houden, en de beste resultaten op te leveren. Deze constructie van de projectievectoren geeft een robuuste en krachtige benadering van de kleine eigenruimte in de zin dat derde benadering voor zowel constante permeabiliteit als variable permeabiliteit in lagen gebruikt kan worden.

Aanbevelingen

Tijdens het onderzoek zijn aantal onderwerpen naar voren gekomen die verder kunnen onderzocht worden:

- Numerieke experimenten laten zien dat de perturbatie m.b.t tweede benadering en derde benadering zeer gering is. De vraag is of dit analytisch bevestigd kan worden gebruikmakend van bijvoorbeeld de structuur van de eigenvectoren.

- Voor een verklaring van het goede convergentiegedrag van DICCG gebaseerd op de derde benadering wordt in het geval van rechte afwisselende rechte lagen een schatting van het effectieve conditiegetal gegeven. Als vervolg daarvan kan een uitbreiding van het bewijs naar andere speciale problemen aan de orde komen. Als voorbeelden : problemen met speciale structuur van ondergrond (afwisselende kromme gebieden, afwisselende gebieden met gaten) of tijdsafhankelijke problemen. Dit bewijs wordt gebaseerd op de eigenschappen van de gediscretiseerde matrix. Dezelfde aanpak kan gelden voor de tweede benadering.

- In de situatie van continue variatie van permeabiliteit in de lagen heeft de DICCG methode in het geval van eerste benadering en derde benadering goede resultaten opgeleverd, terwijl de tweede benadering in een bepaalde situatie niet toepasbaar is. Daarvoor dient een analytische verklaring gevonden te worden. Deze analytisch verklaring kan gebaseerd worden op de perturbatie van de kleine eigenruimte. Hierbij kan men van de structuur gebruikmaken van de kleine eigenvectoren behorend bij deze situaties.

- Bij ons onderzoek worden slechts twee situaties onderzocht. Om meer inzicht te hebben van de invloed van variatie van permeabiliteit kunnen ook andere relevante speciale situatie bekeken worden.

Hoofdstuk 10

Appendices

10.1 Appendix A4

Appendix A4.1

Beschouwen we de volgende diagonale matrix Σ:

$$\Sigma = \begin{pmatrix} \lambda_1 & 0 & \cdots & \\ 0 & \lambda_2 & \cdots & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{pmatrix},$$

Met $\lambda_1 = \lambda_2 = 10^{-6}$, $\lambda_3 = 1$, $\lambda_4 = 2$.
Hierbij is $span\{\mathbf{e_1}, \mathbf{e_2}\}$ de kleine eigenruimte behorend bij Σ.
Nu gaan we verschillende perturbaties van de kleine eigenruimte bekijken:

- Eerste geval: Perturbatie van de kleine eigenruimte uitsluitend in de richting van de kleine eigenvectoren (het aantal projectievectoren is kleiner dan de dimensie van de kleine eigenruimte)

$$\mathbf{w} = 5\mathbf{e_1} + 3\mathbf{e_2}$$

- Tweede geval: Perturbatie in de richting van één eigenvector

$$\mathbf{w_1} = 4\mathbf{e_1} + 3\mathbf{e_2} + 10\mathbf{e_3}$$

$$\mathbf{w_2} = 2\mathbf{e_1} + 5\mathbf{e_2} + \mathbf{e_3}$$

- Derde geval: Perturbatie in meerdere richtingen (in twee verschillende richtingen behorende bij een grote eigenwaarde)

$$\mathbf{w_1} = 7\mathbf{e_1} + \mathbf{e_2} + \mathbf{e_3} + 4\mathbf{e_4}$$

$$\mathbf{w_2} = 8\mathbf{e_1} - 3\mathbf{e_2} + 4\mathbf{e_3} + 5\mathbf{e_4}$$

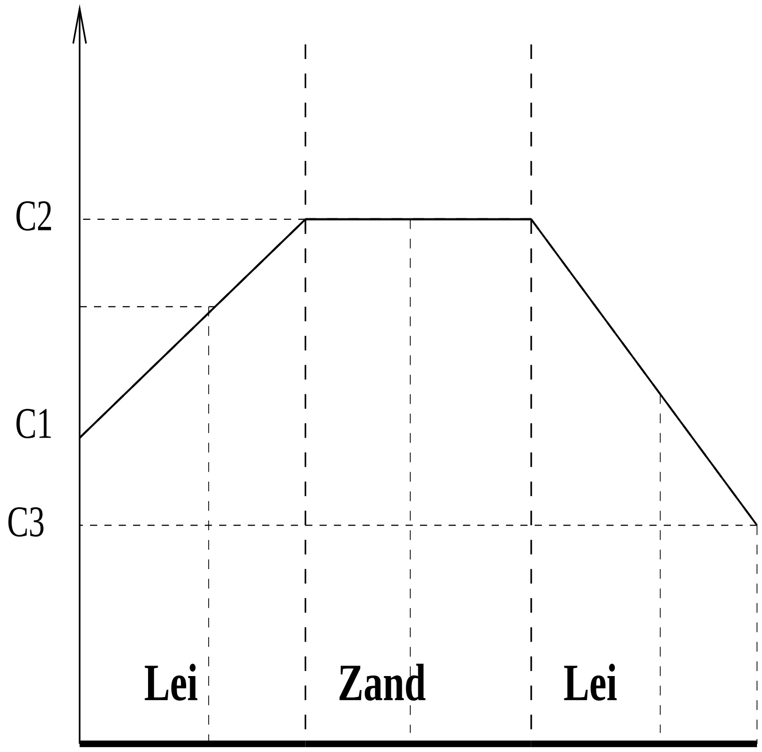
Voor deze gevallen worden de kleine geperturbeerde eigenwaarden van $\hat{P}\Sigma$ en het bijbehorende aantal samengevat in de volgende tabel:

Gevalen	Kleine geperturbeerde eigenwaarden	# kleine geperturbeerde eigenwaarden
Eerste geval	1.0000e-05	1
Tweede geval	1.0795e-05	1
Derde geval	5.6921e-05,2.4812e-05	2

Tabel 10.1:

Appendix A4.2

Eén dimensionaal voorbeeld met drie lagen



Figuur 10.1: Plot van de kleine eigenvector in het geval van 3 lagen

Uit figuur 10.1 blijkt dat een eigenvector $\mathbf{v}$ als volgt geschreven kan worden :

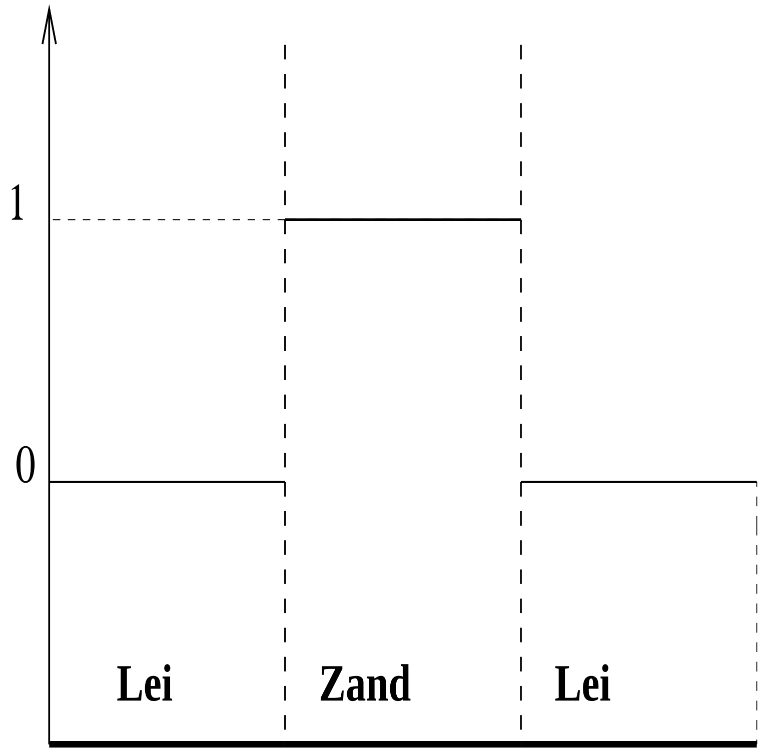
$$\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + ... + \alpha_6 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \alpha_7 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

ofwel:

$$\mathbf{v} = \alpha_1\mathbf{e_1} + \alpha_2\mathbf{e_2} + \alpha_3\mathbf{e_3} + \alpha_4\mathbf{e_4} + \alpha_5\mathbf{e_5} + \alpha_6\mathbf{e_6} + \alpha_7\mathbf{e_7},$$

waarin $\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}, \mathbf{e_4}, \mathbf{e_5}, \mathbf{e_6}$en $\mathbf{e_7}$ de eenheidsvectoren zijn .

De projectievector $\mathbf{w}$ bereffende de tweede benadering (zie figuur 11.4) ziet



Figuur 10.2: Plot van de kleine eigenvector in het geval van 3 lagen

er als volgt uit:

$$\mathbf{w} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \mathbf{e_3} + \mathbf{e_4} + \mathbf{e_5}.$$

De waarde van de eigenvector $\mathbf{v}$ is constant in de zandlaag, dus voldoen de bijbehorende coefficienten aan de volgende betrekkingen:

$$\alpha_3 = \alpha_4 = \alpha_5 = c_2,$$

de waarden van c_1, c_2 en c_3 staan vermeld in figuur (10.1).
En de waarde daarvan is een lineaire interpolatie in leilagen.
We kunnen makkelijk nagaan dat de coëfficiënten α_1 en α_2 aan de volgende relaties voldoen:
In de eerste leilaag:

$$\alpha_1 = c_1 \text{ en } \alpha_2 = \frac{c_2 + c_1}{2}.$$

In de tweede leilaag :

$$\alpha_7 = c_3 \text{ en } \alpha_6 = \frac{c_3 + c_2}{2}.$$

Nu kan de eigenvector op de volgende wijze geformuleerd worden:

$$\mathbf{v} = c_2\mathbf{w} + \alpha_1\mathbf{e_1} + \alpha_2\mathbf{e_2} + \alpha_6\mathbf{e_6} + \alpha_7\mathbf{e_7},$$

ofwel:

$$\mathbf{v} = c_2\mathbf{w} + c_1\mathbf{e_1} + \frac{c_2+c_1}{2}\mathbf{e_2} + \frac{c_3+c_2}{2}\mathbf{e_6} + c_3\mathbf{e_7},$$

De projectievector $\mathbf{w}$ kan dus gesplitst worden in twee vectoren :

$$\mathbf{w} = \mathbf{w^k} + \mathbf{w^g} \text{ met}$$

$$\mathbf{w^k} = \frac{1}{c_2}\mathbf{v} \text{ en } \mathbf{w^g} = c_1\mathbf{e_1} + \frac{c_2+c_1}{2}\mathbf{e_2} + \frac{c_3+c_2}{2}\mathbf{e_6} + c_3\mathbf{e_7}$$

Appendix A4.3

De matrix A is te schrijven als volgt:

$$A = S\Sigma S^T \text{ met } S = [\mathbf{v_1}\mathbf{v_2}....\mathbf{v_n}];$$

Hierin zijn $\mathbf{v_i}$ voor $i = 1..n$ de orthonormale eigenvectoren van $A\,(S^TS = I)$.

De bijbehorende projectiematrix wordt gegeven door:

$$P = I - W((AW)^TW)^{-1}(AW)^T$$

$$W = [\mathbf{w_1}\mathbf{w_2}....\mathbf{w_m}]$$

Er geldt:

$$P = I - W((S\Sigma S^TW)^TW)^{-1}(S\Sigma S^TW)^T$$

$$P = I - W((S^TW)^T\Sigma S^TW)^{-1}(S^TW)^T\Sigma S^T$$

en:

$$P^TA = AP = S\Sigma S^T - S\Sigma S^TW((S\Sigma S^TW)^TW)^{-1}((S\Sigma S^TW)^T$$

$$AP = S\Sigma S^T - S\Sigma S^TW((\Sigma S^TW)^TS^TW)^{-1}((\Sigma S^TW)^T)S^T$$

Als we stellen:

$$\hat{W} = S^TW$$

Dan volgt :

$$P = S(I - \hat{W}((\Sigma\hat{W})^T\hat{W})^{-1}(\Sigma\hat{W})^T)S^T.$$

$$AP = S\Sigma S^T - S\Sigma\hat{W}((\Sigma\hat{W})^T\hat{W})^{-1}(\Sigma\hat{W})^T)S^T.$$

Voor de diagonaal matrix Σ kunnen we de projectie matrix $\hat{P}$ behorend bij $\hat{W}$ definieren als volgt:

$$\hat{P} = I - \hat{W}((\Sigma\hat{W})^T\hat{W})^{-1}(\Sigma\hat{W})^T$$

We kunnen makkelijk nagaan dat:

$$\hat{P} = S^TPS$$

$$AP = S\,(\Sigma\hat{P})\,S^T$$

Het is niet moelijk in te zien dat als λ een eigenwaarde van $\Sigma\hat{P}$ is en ω de bijbehorende eigenvector , dan is λ ook een eigenwaarde van AP en $S\omega$ de bijbehorende eigenvector en omgekeerd. Immers:

$$\Sigma\hat{P}\omega = \lambda\omega \iff S\Sigma\hat{P}S^TS\omega = \lambda S\omega \iff S\Sigma S^TS\hat{P}S^TS\omega = \lambda S\omega \iff APS\omega = \lambda S\omega$$

Appendix A4.4

We gaan bewijzen hier voor het algemene geval de volgende bewering:
als een vector $\mathbf{w}$ in $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$ ligt ,dan ligt $AP^T\mathbf{w}$ ook in $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$ met $p \geq m$ (hierin is m het aantal projectievectoren).
Laat $\mathbf{w}$ een projectievector zijn die tot $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$ behoort, m.a.w:

$$\mathbf{w} = \sum_{j=1}^{p} r_j \mathbf{v_j}.$$

Uiteraard geldt de bewering dan ook voor het geval $(\Sigma, \hat{P})$.
Voor deze vector geldt:

$$PA\mathbf{w} = A\mathbf{w} - W((AW)^T W)^{-1}(AW)^T A\mathbf{w} = A\mathbf{w} - W\mathbf{d},$$

Met:

$$\mathbf{d} = ((AW)^T W)^{-1}(AW)^T A\mathbf{w}$$

We kunnen makkelijk zien dat $W\mathbf{d}$ een lineaire combinatie van de projectie-vectoren $\mathbf{w_j}\, j = 1...m$ is, immers:

$$W\mathbf{d} = \sum_{j=1}^{m} d_j \mathbf{w_j} \text{ met } \mathbf{d} = (d_1, ..., d_m)^T$$

Er geldt:

$$A\mathbf{w} = \sum_{j=1}^{p} r_j A\mathbf{v_j}$$

Op basis van bovengenoemde vergelijkingen geldt:

$$PA\mathbf{w} = \sum_{j=1}^{p} r_j \lambda_j \mathbf{v_j} + \sum_{j=1}^{m} d_j \mathbf{w_j}$$

Duidelijk is dat de eerste deelvector $\sum_{j=1}^{p} r_j \lambda_j \mathbf{v_j}$ in de $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$ zit.
Anderzijds zit de tweede deelvector $\sum_{j=1}^{m} d_j \mathbf{w_j}$ ook in de $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$, want ieder projectievector $\mathbf{w_j}\, j = 1...m$ zit in $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$.
Tenslotte ligt $PA\mathbf{w} = AP^T\mathbf{w}$ in $span\{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_p}\}$.

Appendix A4.5

We hebben de volgende projectiematrix P:

$$\hat{P} = I - \hat{W}\hat{E}^{-1}(\Sigma\hat{W})^T$$

Hierbij worden de matrices $\hat{W}$, $\Sigma\hat{W}$ en $\hat{E}$ gegeven door

$$\hat{E} = ((\Sigma\hat{W})^T\hat{W}),\ \hat{W} = \hat{V}\begin{pmatrix} \bar{R}^T \\ -\mathbf{r^T} \end{pmatrix} \text{ en } \Sigma\hat{W} = \hat{V}\begin{pmatrix} D\bar{R}^T \\ -\lambda_{m+1}\mathbf{r^T} \end{pmatrix}.$$

- Eerst gaan we laten zien dat als $\bar{R}\mathbf{x} = \mathbf{r}$, dan volgt:

$$\hat{P} = I - \hat{V} \begin{pmatrix} (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} D & -\lambda_{m+1}(D + \lambda_{m+1},\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \\ -\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} D & \mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \end{pmatrix} \hat{V}^T.$$

De matrix $\hat{E}$ kan op de volgende wijze geschreven worden:

$$\hat{E} = \begin{pmatrix} D\bar{R} & -\lambda_{m+1}\mathbf{r} \end{pmatrix} \hat{V}^T \hat{V} \begin{pmatrix} \bar{R}^T \\ -\mathbf{r^T} \end{pmatrix}.$$

Na enige berekingen ($V^T V = I$) :

$$\hat{E} = \bar{R} D \bar{R}^T + \lambda_{m+1}\mathbf{r}\,\mathbf{r^T}.$$

Invullen van $\bar{R}\mathbf{x} = \mathbf{r}$ hierin geeft :

$$\hat{E} = \bar{R} D \bar{R}^T + \lambda_{m+1}\bar{R}\mathbf{x}\,\mathbf{x^T}\bar{R}^T.$$

$$\text{ofwel } \hat{E} = \bar{R}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})\bar{R}^T.$$

Nu ziet de projectiematrix er als volgt uit:

$$\hat{P} = I - \hat{V} \begin{pmatrix} \bar{R}^T \\ -\mathbf{r^T} \end{pmatrix} (\bar{R}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})\bar{R}^T)^{-1} \begin{pmatrix} \bar{R}D & -\lambda_{m+1}\mathbf{r} \end{pmatrix} \hat{V}^T.$$

$$\hat{P} = I - \hat{V} \begin{pmatrix} \bar{R}^T \\ -\mathbf{r^T} \end{pmatrix} \bar{R}^{-T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1} \begin{pmatrix} \bar{R}D & -\lambda_{m+1}\mathbf{r} \end{pmatrix} \hat{V}^T.$$

$$\hat{P} = I - \hat{V} \begin{pmatrix} \bar{R}^T \bar{R}^{-T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1} \\ -\mathbf{r^T}\bar{R}^{-T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1} \end{pmatrix} \begin{pmatrix} \bar{R}D & -\lambda_{m+1}\mathbf{r} \end{pmatrix} \hat{V}^T.$$

Omdat er geldt $\mathbf{r^T}\bar{R}^{-T} = \mathbf{x^T}$ (want $\bar{R}$ is inverteerbaar en $\bar{R}\mathbf{x} = \mathbf{r}$), er geldt:

$$\hat{P} = I - \hat{V} \begin{pmatrix} (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1} \\ -\mathbf{x}^T(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1} \end{pmatrix} \begin{pmatrix} \bar{R}D & -\lambda_{m+1}\mathbf{r} \end{pmatrix} \hat{V}^T.$$

$$\hat{P} = I - \hat{V} \begin{pmatrix} (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1}\bar{R}^{1} D & -\lambda_{m+1}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1}\mathbf{r} \\ -\mathbf{x}^T(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1}\bar{R}D & \lambda_{m+1}\mathbf{x}^T(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\bar{R}^{-1}\mathbf{r} \end{pmatrix} \hat{V}^T.$$

Uiteindelijk krijgen we:

$$\hat{P} = I - \hat{V} \begin{pmatrix} (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} D & -\lambda_{m+1}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \\ -\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} D & \lambda_{m+1}\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \end{pmatrix} \hat{V}^T.$$

Appendix A4.6

We gaan aantonen dat de vector

$$\mathbf{e_{per}} = \begin{pmatrix} \mathbf{x} \\ 1 \\ 0 \end{pmatrix}$$

met de deelvector $\mathbf{x}$ die voldoet aan $\bar{R}\mathbf{x} = \mathbf{r}$, een eigenvector is van ΣP. Bovendien is de bijbehorende eigenwaarde λ_{per} als volgt geformuleerd :

$$\lambda_{per} = \lambda_{m+1}\Big((1 + \mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x}))\Big).$$

en

$$\Sigma\hat{P}e_{per} = \begin{pmatrix} D\mathbf{x} - D\Big((D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D\mathbf{x} - \lambda_{m+1}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x}\Big) \\ \lambda_{m+1} - \lambda_{m+1}\Big(\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D\mathbf{x} - \mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\mathbf{x^T})^{-1}\mathbf{x}\Big) \\ 0 \end{pmatrix}$$

Er geldt:

$$\Sigma\hat{P}\mathbf{e_{per}} = \Sigma\mathbf{e_{per}} - \Sigma\hat{V}\begin{pmatrix} (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D & -\lambda_{m+1}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \\ -\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D & \lambda_{m+1}\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\mathbf{x^T})^{-1}\mathbf{x} \end{pmatrix}\hat{V}^T\mathbf{e_{per}}. \tag{10.1}$$

Het is eenvoudig na te gaan:

$$\Sigma\mathbf{e_{per}} = \begin{pmatrix} D & & & \\ 0 & \lambda_{m+1} & & \\ 0 & & \ddots & \\ & & & \lambda_n \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ 1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} D\mathbf{x} \\ \lambda_{m+1} \\ 0 \end{pmatrix},$$

$$\Sigma\hat{V} = \begin{pmatrix} D & \mathbf{0} \\ \mathbf{0} & \lambda_{m+1} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \text{ en } \hat{V}^T\mathbf{e_{per}} = \begin{pmatrix} \mathbf{e_1}^T \\ \mathbf{e_2}^T \\ \vdots \\ \mathbf{e_{m+1}^T} \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ 1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}.$$

Invullen van de bovengenoemde vergelijkingen wordt na enige berekeningen de vergelijking (10.1) als volgt geschreven:

$$\Sigma P\mathbf{e_{per}} = \begin{pmatrix} D\mathbf{x} - D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D\mathbf{x} + \lambda_{m+1}D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}\mathbf{x} \\ \lambda_{m+1} + \lambda_{m+1}\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}D\mathbf{x} - \lambda_{m+1}^2\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\mathbf{x^T})^{-1}\mathbf{x} \\ \mathbf{0} \end{pmatrix}$$

Verder uitgewerkt:

$$\Sigma\hat{P}\mathbf{e_{per}} = \begin{pmatrix} D\mathbf{x} - D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x}) \\ \lambda_{m+1}(1 + \mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x})) \\ \mathbf{0} \end{pmatrix}$$

Nu nemen we als eigenwaarde van $\Sigma\hat{P}$

$$\lambda_{per} = \lambda_{m+1}\Big((1 + \mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x}))\Big).$$

Verder gaan we controleren:

$$\Sigma\hat{P}\mathbf{e_{per}} = \lambda_{per}\mathbf{e_{per}} \text{ ofwel } \Sigma\hat{P}\begin{pmatrix}\mathbf{x}\\ 1\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\lambda_{per}\mathbf{x}\\ \lambda_{per}\\ \mathbf{0}\end{pmatrix}$$

Hieruitvolgt dat het voldoende is te bewijzen:

$$D\mathbf{x} - D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x}) = \lambda_{per}\mathbf{x}$$

Duidelijk is:

$$(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} = I.$$

$$D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} = I.$$

Hieruitvolgt:

$$D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1} = I - \lambda_{m+1}\mathbf{x}\,\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}.$$

Dus er geldt:

$$D\mathbf{x}-D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x}-\lambda_{m+1}\mathbf{x}) = D\mathbf{x}-(I-\lambda_{m+1}\mathbf{x}\,\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1})(D\mathbf{x}-\lambda_{m+1}\mathbf{x}).$$

Na enig reken werk krijgen we:

$$D\mathbf{x}-D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x}-\lambda_{m+1}\mathbf{x}) = \Big(\lambda_{m+1}-\lambda_{m+1}\mathbf{x^T}(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1})(-D\mathbf{x}+\lambda_{m+1}\mathbf{x})\Big)\mathbf{x}.$$

ofwel:

$$D\mathbf{x} - D(D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1}(D\mathbf{x} - \lambda_{m+1}\mathbf{x}) = \lambda_{per}\mathbf{x}$$

Uiteindelijk :

$$\Sigma\hat{P}\mathbf{e_{per}} = \lambda_{per}\mathbf{e_{per}}.$$

Appendix A4.7

- We gaan eerst laten zien dat $||(D + \lambda_{m+1}\mathbf{xx^T})^{-1} - D^{-1}||$ naar nul. convergeert, als $||x||$ naar nul convergeert.
  Stellen we : $\mathbf{Z_1} = (D + \lambda_{m+1}\mathbf{xx^T})^{-1} - D^{-1}$, dan geldt:

  $$\mathbf{Z_1}D+\mathbf{Z_1}\lambda_{m+1}\mathbf{xx^T} = (D + \lambda_{m+1}\mathbf{x}\,\mathbf{x^T})^{-1})(D+\lambda_{m+1}\mathbf{xx^T})-D^{-1}(D+\lambda_{m+1}\mathbf{x}\,\mathbf{x^T}),$$

  waaruitvolgt:

  $$\mathbf{Z_1}D = I - \lambda_{m+1}\mathbf{Z_1}\mathbf{xx^T} - I - \lambda_{m+1}D^{-1}\mathbf{xx^T}.$$

  Vermenigvuldiging van beide termen met D^{-1} levert:

  $$\mathbf{Z_1} = \lambda_{m+1}\mathbf{Z_1}\mathbf{xx^T}D^{-1} - \lambda_{m+1}D^{-1}\mathbf{xx^T}D^{-1}. \qquad (10.2)$$

Toepassing van de driehoeksongelijkheid op (10.2) geeft:

$$||\mathbf{Z_1}|| < \lambda_{m+1}||\mathbf{Z_1}\mathbf{x}\mathbf{x}^{\mathbf{T}}D^{-1}|| + \lambda_{m+1}||D^{-1}\mathbf{x}\mathbf{x}^{\mathbf{T}}D^{-1}||.$$

$$||\mathbf{Z_1}|| < \lambda_{m+1}||\mathbf{Z_1}||\,||\mathbf{x}\mathbf{x}^{\mathbf{T}}||\,||D^{-1}|| + \lambda_{m+1}||D^{-1}||\,||\mathbf{x}\mathbf{x}^{\mathbf{T}}||\,||D^{-1}||.$$

Verder uitgewerkt:

$$(1 - \lambda_{m+1}\,||\mathbf{x}||^2\,||D^{-1}||)||\mathbf{Z_1}|| \leq \lambda_{m+1}||D^{-1}||\,||\mathbf{x}||^2\,||\,||D^{-1}||.$$

Omdat $||\mathbf{x}||$ voldoende klein is, dan geldt

$$(1 - \lambda_{m+1}\,||\mathbf{x}||^2\,||D^{-1}|| > 0),$$

waaruitvolgt:

$$||\mathbf{Z_1}|| < \frac{\lambda_{m+1}\,||\mathbf{x}||^2\,||D^{-2}||}{1 - \lambda_{m+1}\,||D^{-1}||\,||\mathbf{x}||^2}.$$

Hieruit kunnen we afleiden dat als $||\mathbf{x}||$ naar nul convergeert, dan ook $||(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^{\mathbf{T}})^{-1}||$ naar $||D^{-1}||$ convergeert.

- – Eerst laten we zien dat $\mathbf{x}$ een eigenvector van $D(D + \lambda_{m+1}\mathbf{x}\mathbf{x}^{\mathbf{T}})^{-1}$ is met $\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}$ de bijbehorende eigenwaarde. Bovendien gaat $\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}$ naar nul, indien $\mathbf{x}$ zeer groot is.

  We stellen : $\mathbf{Z_2} = D(D + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}$.
  Uitwerken van $\mathbf{Z_2}$ geeft:

  $$\mathbf{Z_2} = (I + \lambda_{m+1}\mathbf{x}\mathbf{x}^{\mathbf{T}}D^{-1})^{-1}.$$

  Het is eenvoudig na te gaan dat:

  $$(I + \lambda_{m+1}\mathbf{x}\mathbf{x}^{\mathbf{T}}D^{-1})\mathbf{x} = (1 + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x})\mathbf{x}.$$

  Hieruit volgt:

  $$\mathbf{Z_2}\mathbf{x} = \frac{1}{1 + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}\mathbf{x}.$$

  Door de driehoeksongelijkheid op : $(\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x} - 1) + 1$ toe te passen hebben we

  $$|(\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x})| \leq 1 + |1 + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}\,D^{-1}\mathbf{x}|.$$

  Verder uitgewerkt:

  $$|(\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x})| - 1 \leq |1 + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}\,D^{-1}\mathbf{x}|.$$

  We hebben

  $$|\lambda_{m+1}\mathbf{x}^{\mathbf{T}}\,D^{-1}\mathbf{x}| = \sum_{i=1}^{m}\frac{\lambda_{m+1}}{\lambda_i}x_i^2 \geq ||\mathbf{x}||^2(\lambda_i < \lambda_{m+1}),$$

waaruitvolgt:

$$||\mathbf{x}||^2 - 1 \leq |1 + \lambda_{m+1}\mathbf{x}^{\mathbf{T}}\, D^{-1}\mathbf{x}|.$$

We kunnen nu de volgende schatting vinden voor de eigenwaarde $\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}$ van $\mathbf{x}$:

$$|\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}| \leq \frac{1}{||\mathbf{x}||^2 - 1},$$

derhalve $\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}$ convergeert naar nul, als $||\mathbf{x}||$ naar oneindig convergeert.

- Nu gaan we bewijzen $||\mathbf{Z_2}|| = ||D(D+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}||$ begrensd is, als $||\mathbf{x}||$ voldoende groot is.

  Door het Gramschmidt proces kunnen we onafhankelijke vectoren $\mathbf{y_i}\,(i=2,..,d)$ construeren die loodrecht op $\mathbf{x}$ staan, m.a.w. $\mathbf{x}^{\mathbf{T}}\mathbf{y_i} = 0$.
  Door $\mathbf{x_i} = D\mathbf{y_i}$ te stellen krijgen we $\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x_i} = 0$.
  Het is niet moeilijk in te zien dat de vectoren $\mathbf{x_i}$ voldoen aan de volgende vergelijking:

$$(I + \lambda_{m+1}\mathbf{x}\mathbf{x}^{\mathbf{T}}D^{-1})\mathbf{x_i} = \mathbf{x_i},$$

  derhalve:

$$\mathbf{Z_2}\mathbf{x_i} = \mathbf{x_i}$$

  waaruitvolgt dat $\mathbf{x_i}\,(i=2,...,d)$ de eigenvektoren van $\mathbf{Z_2}$ zijn met 1 de bijbehorende eigenwaarde.
  Voor een symmetrische matrix $A\,(A = S\Sigma S^T)$ geldt:

$$||A|| = ||\Sigma|| \leq \sum_{i=1}^{n} \lambda_i^2,\ \lambda_i \text{ is een eigenwaarde van A.}$$

  waaruit volgt :

$$||\mathbf{Z_2}|| \leq (d-1) + \left(\frac{1}{1+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}D^{-1}\mathbf{x}}\right)^2.$$

  Indien $\mathbf{x}$ zeer groot is, dan geldt:

$$||\mathbf{Z_2}|| \leq d - 1.$$

  Hieruit kunnen we afleiden dat als $||\mathbf{x}||$ naar oneindig convergeert, dan is $||D(D+\lambda_{m+1}\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}||$ begrensd door een constante M.

10.2 Appendix A5

Appendix A5.1

We hebben de volgende projectiematrix P:

$$\hat{P} = I - \hat{W}\hat{E}^{-1}(\Sigma\hat{W})^T$$

Hierbij worden de matrices $\hat{W}$, $\Sigma\hat{W}$ en $\hat{E}$ gegeven door

$$\hat{E} = ((\Sigma\hat{W})^T\hat{W}),\ \hat{W} = \hat{V}\begin{pmatrix} \bar{R}^T \\ -\underline{R}^T \end{pmatrix} \text{ en } \Sigma\hat{W} = \hat{V}\begin{pmatrix} \bar{D}\bar{R}^T \\ -\underline{D}\,\underline{R}^T \end{pmatrix}.$$

- Eerst gaan we laten zien dat als $\bar{R}X = \underline{R}$, dan volgt:

$$\hat{P} = I - \hat{V}\begin{pmatrix} (\bar{D} + X\,\underline{D}X)^{-1}\bar{D} & -(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \\ -X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D} & X^T(\bar{D} + X\,\underline{D}X)^{-1}X\underline{D} \end{pmatrix}\hat{V}^T.$$

De matrix $\hat{E}$ kan op de volgende wijze geschreven worden:

$$\hat{E} = \begin{pmatrix} \bar{R}\,\bar{D} & -\underline{R}\,\underline{D} \end{pmatrix}\hat{V}^T\hat{V}\begin{pmatrix} \bar{R}^T \\ -\underline{R}^T \end{pmatrix}.$$

Na enige berekeningen ($V^TV = I$) :

$$\hat{E} = \bar{R}\,\bar{D}\,\bar{R}^T + \underline{R}\,\underline{D}\,\underline{R}^T.$$

Invullen van $\bar{R}X = \underline{R}$ hierin geeft :

$$\hat{E} = \bar{R}\bar{D}\bar{R}^T + \bar{R}X\,\underline{D}X^T\bar{R}^T.$$

$$\text{ofwel } \hat{E} = \bar{R}(\bar{D} + X\,\underline{D}X^T)\bar{R}^T.$$

Nu ziet de projectiematrix er als volgt uit:

$$\hat{P} = I - \hat{V}\begin{pmatrix} \bar{R}^T \\ -\underline{R}^T \end{pmatrix}(\bar{R}(\bar{D} + X\,\underline{D}X^T)\bar{R}^T)^{-1}\begin{pmatrix} \bar{R}\,\bar{D} & -\underline{R}\,\underline{D} \end{pmatrix}\hat{V}^T.$$

$$\hat{P} = I - \hat{V}\begin{pmatrix} \bar{R}^T \\ -\underline{R}^T \end{pmatrix}\bar{R}^{-T}(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1}\begin{pmatrix} \bar{R}\bar{D} & -\underline{R}\,\underline{D} \end{pmatrix}\hat{V}^T.$$

$$\hat{P} = I - \hat{V}\begin{pmatrix} \bar{R}^T\bar{R}^{-T}(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1} \\ -\underline{R}^T\bar{R}^{-T}(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1} \end{pmatrix}\begin{pmatrix} \bar{R}\,\bar{D} & -\underline{R}\,\underline{D} \end{pmatrix}\hat{V}^T.$$

Er geldt $\underline{R}^T\bar{R}^{-T} = X^T$ (want $\bar{R}$ is inverteerbaar en $\bar{R}X = \underline{R}$), waaruitvolgt:

$$\hat{P} = I - \hat{V}\begin{pmatrix} (\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1} \\ -X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1} \end{pmatrix}\begin{pmatrix} \bar{R}\bar{D} & -\underline{R}\,\underline{D} \end{pmatrix}\hat{V}^T.$$

$$\hat{P} = I - \hat{V} \begin{pmatrix} (\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1}\bar{R}\bar{D} & -(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1}\underline{RD} \\ -X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1}\bar{R}\bar{D} & X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{R}^{-1}\underline{RD} \end{pmatrix} \hat{V}^T.$$

Uiteindelijk krijgen we:

$$\hat{P} = I - \hat{V} \begin{pmatrix} (\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D} & -(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \\ -X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D} & X^T(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \end{pmatrix} \hat{V}^T.$$

Appendix A5.2

- We gaan eerst aantonen dat voor de vectoren

$$E_{per} = \begin{pmatrix} X \\ 1 \\ 0 \end{pmatrix}$$

met de deelvector X die voldoet aan $\bar{R}X = \underline{R}$ geldt:

$$\Sigma\hat{P}E_{per} = \begin{pmatrix} \bar{D}X - \bar{D}(\bar{D} + X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{D} + \underline{D}X^T(\bar{D} + X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{0} \end{pmatrix}.$$

Uitwerking van $\Sigma\hat{P}E_{per}$ levert:

$$\Sigma\hat{P}E_{per} = \Sigma E_{per} - \Sigma\hat{V} \begin{pmatrix} (\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D} & -(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \\ -X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D} & X^T(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \end{pmatrix} \hat{V}^T E_{per}. \tag{10.3}$$

Het is eenvoudig na te gaan:

$$\Sigma E_{per} = \begin{pmatrix} \bar{D} & & & & \\ 0 & \underline{D} & & & \\ 0 & & \lambda_{m+p+1} & & \\ 0 & & & \ddots & \\ & & & & \lambda_n \end{pmatrix} \begin{pmatrix} X \\ 1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \bar{D}X \\ \underline{D} \\ 0 \end{pmatrix},$$

$$\Sigma\hat{V} = \begin{pmatrix} \bar{D} & \mathbf{0} \\ \mathbf{0} & \underline{D} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \text{ en } \hat{V}^T E_{per} = \begin{pmatrix} \mathbf{e_1}^T \\ \mathbf{e_2}^T \\ \vdots \\ \mathbf{e_p^T} \end{pmatrix} \begin{pmatrix} X \\ I \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} X \\ I \end{pmatrix}.$$

Invullen van de bovengenoemde vergelijkingen wordt na enige berekeningen de vergelijking (10.3) als volgt geschreven:

$$\Sigma\hat{P}E_{per} = \begin{pmatrix} \bar{D}X - \bar{D}(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D}X + \bar{D}(\bar{D} + X\,\underline{D}X^T)^{-1}X\underline{D} \\ \underline{D} - \underline{D}X^T(\bar{D} + X\,\underline{D}X^T)^{-1}\bar{D}X + \underline{D}X^T(\bar{D} + X\underline{D}X^T)^{-1}X\underline{D} \\ \mathbf{0} \end{pmatrix}$$

Verder uitgewerkt:

$$\Sigma\hat{P}E_{per} = \begin{pmatrix} \bar{D}X - \bar{D}(\bar{D} + X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{D} + \underline{D}X^T(\bar{D} + X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D}) \\ \underline{0} \end{pmatrix}. \tag{10.4}$$

- Hieronder gaan we aantonen dat:

$$\Sigma\hat{P}E_{per} = E_{per}D_{per} = \begin{pmatrix} X \\ I \\ \mathbf{0} \end{pmatrix} D_{per}. \qquad (10.5)$$

Dit wordt in hoofdstuk (7) gebruikt om aan te tonen dat de matrices $\Sigma\hat{P}$ en D_{per} dezelfde eigenwaarden hebben.
Volgens de vergelijkingen (10.4)en (10.5) is het voldoende te bewijzen:

$$\bar{D}X - \bar{D}(\bar{D} + X\,\underline{D}X)^{-1}(\bar{D}X - X\underline{D}) = X D_{per}$$

Duidelijk is:

$$(\bar{D} + X\,\underline{D}X)(\bar{D} + X\,\underline{D}X)^{-1} = I.$$

$$\bar{D}(\bar{D} + X\,\underline{D}X)^{-1} + X\,\underline{D}X(\bar{D} + X\,\underline{D}X)^{-1} = I.$$

Hieruitvolgt:

$$\bar{D}(\bar{D} + X\,\underline{D}X)^{-1} = I - X\,\underline{D}X(\bar{D} + X\,\underline{D}X)^{-1}.$$

Dus er geldt:

$$\bar{D}X - \bar{D}(\bar{D} + X\,\underline{D}X)^{-1}(\bar{D}X - X\underline{D}) = \bar{D}X - (I - X\,\underline{D}X(\bar{D} + X\,\underline{D}X)^{-1})(\bar{D}X - X\underline{D}).$$

Na enig rekenwerk krijgen we:

$$\bar{D}X - \bar{D}(\bar{D} + X\,\underline{D}X)^{-1}(\bar{D}X - X\underline{D}) = X\Big(\underline{D} + \underline{D}X^T(\bar{D} + X\underline{D}X^T)^{-1}(\bar{D}X - X\underline{D})\Big).$$

Volgens definitie van D_{per} (zie formule (7.14) in hoofdstuk 7)

$$\bar{D}X - \bar{D}(\bar{D} + X\,\underline{D}X)^{-1}(\bar{D}X - X\underline{D}) = X D_{per}.$$

Hieruitvolgt:

$$\Sigma\hat{P}E_{per} = E_{per}D_{per}. \qquad (10.6)$$

Appendix A5.3

- We gaan eerst laten zien dat $||(\bar{D} + X\underline{D}X^T)^{-1} - \bar{D}^{-1}||$ naar nul. convergeert, als $||X||$ naar nul convergeert.
Stellen we : $Z_1 = (\bar{D} + X\underline{D}X^T)^{-1} - \bar{D}^{-1}$, dan kunnen we op analoge wijze aantonen dat de matrix Z_1 voldoet aan:

$$Z_1 = Z_1 X\underline{D}X^T\bar{D}^{-1} - \bar{D}^{-1}X\underline{D}X^T\bar{D}^{-1},$$

en de toepassing van de driehoeksongelijkheid geeft:

$$\Big(1 - ||\underline{D}||\,||\bar{D}^{-1}||\,||X^T||\,||X||\Big)||Z_1|| \leq ||\bar{D}^{-2}||\,||\underline{D}||\,||X^T||\,||X||.$$

Als $||X||$ voldoende klein is, dan geldt

$$\left(1 - ||\bar{D}^{-1}|| \, ||\underline{D}|| \, ||X^T|| \, ||X||\right) > 0,$$

waaruitvolgt:

$$||Z_1|| < \frac{||\bar{D}^{-2}|| \, ||\underline{D}^{-1}|| \, ||X^T|| \, ||X||}{1 - ||\bar{D}^{-1}|| \, ||\underline{D}|| \, ||X^T|| \, ||X||}.$$

Hieruit kunnen we afleiden dat als $||X||$ naar nul convergeert, dan ook $(\bar{D} + X\underline{D}X^T)^{-1}$ naar $\bar{D}^{-1}$ convergeert.

- Daarna gaan we bewijzen dat $||(\bar{D} + X^T\underline{D}X)^{-1}||$ en $\frac{||X\underline{D}\mathbf{g}||}{||X\mathbf{g}||}$ (voor ieder $\mathbf{g}$) begrensd zijn als $||X||$ naar oneindig convergeert.

  – Aangezien $\bar{D}$ symmetrich en positief definiet is, bestaat er een matrix C zodat $\bar{D} = C^TC$.
  We kunnen makkelijk controleren:

  $$(\bar{D} + X^T\underline{D}X)^{-1} = C^{-1}(I + (XC^{-1})^T\underline{D}(XC^{-1}))^{-1}C^{-T}$$

  Volgens driehoeksongelijkheid geldt:

  $$(\bar{D} + X^T\underline{D}X)^{-1} \leq ||C^{-1}|| \, ||(I + (XC^{-1})^T\underline{D}(XC^{-1}))^{-1}|| \, ||C^{-T}||$$

  Merk op dat de matrix $(XC^{-1})^T\underline{D}(XC^{-1})$ een symmetrische positie matrix is.
  We noteren de eigenwaarden van $(XC^{-1})^T\underline{D}(XC^{-1})$ met α_i $(i = 1..m)$.
  Voor een willekeurige $\mathbf{y}$ geldt dat als we $\mathbf{z} = XC^{-1}\mathbf{y}$ nemen, dan volgt

  $$\mathbf{y}^T(XC^{-1})^T\underline{D}(XC^{-1})\mathbf{y} = \mathbf{z}^T\underline{D}\mathbf{z} \geq 0,$$

  waaruitvolgt dat de matrix $XC^{-1})^T\underline{D}(XC^{-1})$ semi positief definiet is, derhalve $\alpha_i \geq 0$.
  Het is niet moelijk in te zien dat de eigenwaarden van $(I + (XC^{-1})^T\underline{D}(XC^{-1}))^{-1}$ $\frac{1}{1+\alpha_i}$ zijn, daarom er geldt:

  $$||(I + (XC^{-1})^T\underline{D}(XC^{-1}))^{-1}|| = \frac{1}{min_i(1 + \alpha_i)} \leq 1,$$

  waaruitvolgt:

  $$||(\bar{D} + X^T\underline{D}X)^{-1}|| \leq \frac{||C^{-1}|| \, ||C^{-T}||}{min_i(1 + \alpha_i)}. \qquad (10.7)$$

  Als $||X||$ naar oneindig convergeert, dan convergeert $(XC^{-1})^T\underline{D}(XC^{-1})$ ook naar oneindig, vervolgens er bestaat teminste een α_i die ook naar oneindig convergeert.

We laten definieren de verzameling I van indices i waarvoor geldt α_i die niet naar oneindig convergeert.
Uit vergelijking (10.7) volgt:

$$||(\bar{D} + X^T \underline{D} X)^{-1}|| \leq ||C^{-1}|| \, ||C^{-T}||.$$

Hiermee wordt bewezen dat $||(\bar{D} + X^T \underline{D} X)^{-1}||$ begrensd is.

– We kunnen makkelijk controleren:

$$||X\underline{D}\mathbf{g}|| = \sum_{i=m+1}^{p} \left(\sum_{j=1}^{p} \lambda_j g_j x_{ij} \right)^2 \leq max_j \lambda_j^2 \sum_{i=m+1}^{p} \left(\sum_{j=1}^{p} g_j x_{ij} \right)^2.$$

$$||X\underline{D}\mathbf{g}|| \leq max_j \lambda_j^2 \, ||X\mathbf{g}||.$$

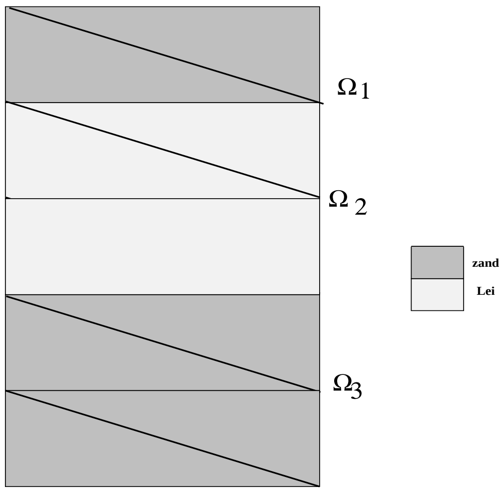
Hiermee wordt bewezen dat $\frac{||X\underline{D}\mathbf{g}||}{||X\mathbf{g}||}$ begrensd is.

10.3 Appendix A6

Appendix A6.1

In het geval van eerste benadering gaan we bewijzen dat als het aantal lagen oneven is, dan er volgt dat de kleine eigenvectoren in de deelruimte opgespannen door de projectievectoren van de eerste benadering liggen. Bijgevolg kan geen perturbatie van de kleine eigenruimte tot standkomen.
In het geval van tweede en derde benadering gaan we aantonen dat de optredende perturbatie gering is en bovendien is deze in het geval van derde benadering wat kleiner.
Hierbij zullen we ons beperken tot een eenvoudige gelaagd probleem met drie lagen.

Een gelaagd probleem met drie lagen



Figuur 10.3: Het probleem met drie lagen

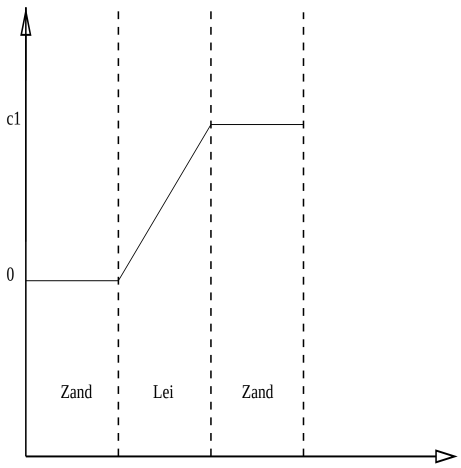
Voor de eenvoud kiezen we slechts twee punten in de horizontale richting.
De bijbehorende matrix A ziet als volgt uit:

$$A = \begin{pmatrix} D_2 & J_1 & & & & \\ J_1 & D_{1+\epsilon} & J_\epsilon & & & \\ & J_\epsilon & D_{2\epsilon} & J_\epsilon & & \\ & & J_\epsilon & D_{1+\epsilon} & J_1 & \\ & & & J_1 & D_2 & J_1 \\ & & & & J_1 & D_1 \end{pmatrix}.$$

In dit geval hebben we alleen één kleine eigenvector $\mathbf{v}$ van de matrix $L^{-1}L^{-T}A$.
Eerste benadering:

De vector $\mathbf{v}$ kan als volgt geschreven worden:

$$\mathbf{v} = \begin{pmatrix} \alpha_{11} \\ \alpha_{12} \\ \alpha_{22} \\ \alpha_{23} \\ \alpha_{31} \\ \alpha_{32} \\ \alpha_{41} \\ \alpha_{42} \\ \alpha_{51} \\ \alpha_{52} \\ \alpha_{61} \\ \alpha_{62} \end{pmatrix} = \sum_{i=1}^{6} \sum_{j=1}^{2} \alpha_{ij} \mathbf{e}_{(\mathbf{j+2(i-1)})}.$$



Figuur 10.4: Plot van de verticale doorsnede van kleine eigenvector in het geval van 3 lagen

Uit figuur (10.4) kunnen we makkelijk nagaan dat de coëfficiënten van $\mathbf{v}$ aan de volgende relaties voldoen:

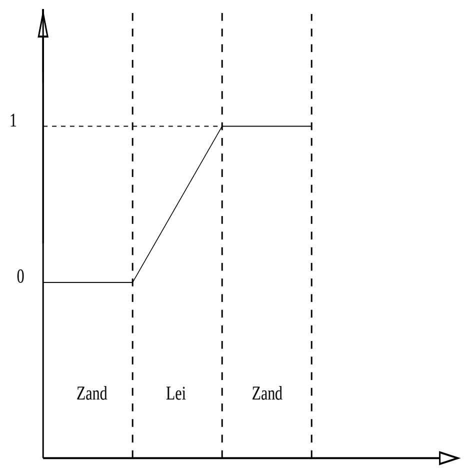
$\alpha_{11} = \alpha_{12} = \alpha_{21} = \alpha_{22} = 0$ (in de eerste zandaag).

$\alpha_{31} = \alpha_{32} = c_1/2$ (in de leilaag).

$\alpha_{41} = \alpha_{42} = \alpha_{51} = \alpha_{52} = \alpha_{61} = \alpha_{62} = c_1$ (in de tweede zandaag).

Nu kan de eigenvector $\mathbf{v}$ op de volgende wijze geformuleerd worden:

$$\mathbf{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ c_1/2 \\ c_1/2 \\ c_1 \\ c_1 \\ c_1 \\ c_1 \\ c_1 \\ c_1 \end{pmatrix} = c_1/2\mathbf{e_4}+c_1/2\mathbf{e_5}+c_1\mathbf{e_6}+\mathbf{e_7}+c_1\mathbf{e_8}+c_1\mathbf{e_9}+c_1\mathbf{e_{10}}+c_1\mathbf{e_{11}}+c_1\mathbf{e_{12}}.$$



Figuur 10.5: Plot van de kleine eigenvector in het geval van 3 lagen

De projectievector $\mathbf{w}$ betreffende de eerste benadering (zie figuur 10.5) ziet er als volgt uit:

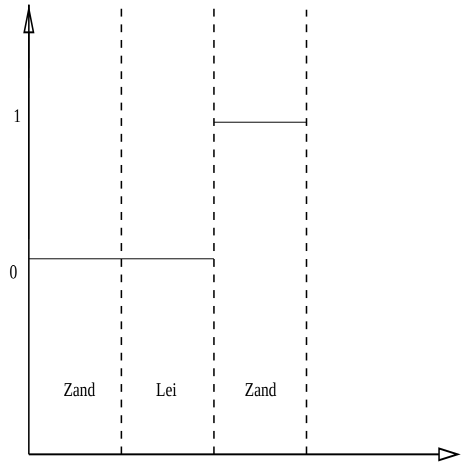
$$\mathbf{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1/2 \\ 1/2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 1/2\mathbf{e_4} + 1/2\mathbf{e_5} + \mathbf{e_6} + \mathbf{e_7} + \mathbf{e_8} + \mathbf{e_9} + \mathbf{e_{10}} + \mathbf{e_{11}} + \mathbf{e_{12}}.$$

Hieruit volgt de volgende verband tussen $\mathbf{w}$ en $\mathbf{v}$:

$$\mathbf{v} = c_1\mathbf{w}.$$

Dus ligt de kleine eigenvector in de deelruimte $span\{\mathbf{w}\}$. Op dezelfde wijze kunnen we aantonen in het geval van het aantal oneven lagen dat de kleine eigenvectoren in de deelruimte opgespannen door de projectievectoren van de eerste benadering liggen.

Tweede benadering:
In het geval van drie lagen hebben we bij tweede benadering één projectie-



Figuur 10.6: Plot van de kleine eigenvector in het geval van 3 lagen

vector, die wordt aangegeven door $\mathbf{w}$ (zie figuur (10.6)).
De projectievector $\bar{\mathbf{w}}$ behorend bij de matrix $\bar{A} = L^{-1}AL^{-T}$ wordt gegeven door:

$$\bar{\mathbf{w}} = L^{-1}\mathbf{w}$$

M.b.v matlab kunnen we de coöfficiënten van de bijbehorende projectievector $\bar{\mathbf{w}}$ in de richtingen van de niet kleine eigenvectoren $\bar{\mathbf{v}}_\mathbf{i}$ van de matrix $\bar{A} = L^{-1}AL^{-T}$ berekenen.
De projectievector $\bar{\mathbf{w}}$ kan geschreven worden als een lineaire combinatie van de eigenvectoren van de matrix:

$$\bar{\mathbf{w}} \approx -1.9013\bar{\mathbf{v}}_\mathbf{1} + 0.0001\bar{\mathbf{v}}_\mathbf{9} - 0.0002\bar{\mathbf{v}}_\mathbf{10}.$$

We kunnen makkelijk zien dat de perturbatie in richting van de kleine eigenvectoren die niet bij de kleine eigenruimte behoren zeer gering is.

Derde benadering:
Analoog kunnen we m.b.v matlab de cöefficiënten van de bijbehorende projectievectoren $\bar{\mathbf{w}}_\mathbf{1}$, $\bar{\mathbf{w}}_\mathbf{2}$, $\bar{\mathbf{w}}_\mathbf{3}$ bepalen van de matrix $\bar{A}$.
Deze projectievectoren zien er als volgt uit:

$$\bar{\mathbf{w}}_\mathbf{1} \approx -0.0314\bar{\mathbf{v}}_\mathbf{3} + 1.7077\bar{\mathbf{v}}_\mathbf{4} + 0.4893\bar{\mathbf{v}}_\mathbf{6} + 0.2172\bar{\mathbf{v}}_\mathbf{7} + 0.0002\bar{\mathbf{v}}_\mathbf{9} - 0.0008\bar{\mathbf{v}}_\mathbf{9}.$$

$$\bar{\mathbf{w}}_\mathbf{2} \approx 0.0002\bar{\mathbf{v}}_\mathbf{9} + 0.0004\bar{\mathbf{v}}_\mathbf{10}.$$

$$\bar{\mathbf{w}}_\mathbf{3} \approx -1.9013\bar{\mathbf{v}}_\mathbf{1} + 0.0001\bar{\mathbf{v}}_\mathbf{9} - 0.0002\bar{\mathbf{v}}_\mathbf{10}.$$

We merken op dat er geldt $\bar{\mathbf{w}} = \bar{\mathbf{w}}_\mathbf{3}$, omdat de projectievectoren $\bar{\mathbf{w}}$ en $\bar{\mathbf{w}}_\mathbf{3}$ behorend tot dezelfde zandlaag.
We kunnen makkelijk controleren:

$$2\bar{\mathbf{w}}_\mathbf{3} + \bar{\mathbf{w}}_\mathbf{2} \approx -3.80026\bar{\mathbf{v}}_\mathbf{1}.$$

Hieruit kunnen we constateren dat de kleine eigenvector $\bar{\mathbf{v}}_\mathbf{1}$ bijna in de nieuwe deelruimte $span\{\bar{\mathbf{w}}_\mathbf{1}, \bar{\mathbf{w}}_\mathbf{2}, \bar{\mathbf{w}}_\mathbf{3}\}$. Als de resultaat daarvan is de perturbatie hierbij bijna nul is.

Appendix A6.2

In deze appendix gaan we aantonen de matrices A^* en C zijn positief semidefiniet. En verder geldt dat $span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$ gelijk aan de nulruimte van A^* is.

We noteren een vector die alleen uit enen bestaat met $\mathbf{1}$.
We beginnen eerst de matrix A^* en de projectievectoren $\mathbf{w_k}$ voor $k = 1...m$ als volgt te schrijven:

$$A^* = \begin{pmatrix} A_1 & & & & & \\ & A_2 & & & & \\ & & A_3 & & & \\ & & & \ddots & & \\ & & & & \ddots & \\ & & & & & A_m \end{pmatrix}, \; C = \begin{pmatrix} C_1 & & & & & \\ & 0 & & & & \\ & & C_2 & & & \\ & & & \ddots & & \\ & & & & 0 & \\ & & & & & C_{m-1} \end{pmatrix},$$

waarbij A_k is op de volgende wijze gedefinieerd voor $k = 1...m$:

$$A_k = \begin{cases} B_1 & \text{als k oneven is} \\ B_\epsilon & \text{anders} \end{cases}, \; C_k = \begin{cases} E_1 & \text{als k oneven is} \\ E_\epsilon & \text{als k even is} \\ D_1 & \text{als } k = 1 \\ 0 & \text{anders} \end{cases}$$

en

$$\mathbf{w_1} = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \mathbf{w_2} = \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \cdots, \mathbf{w_m} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1} \end{pmatrix}.$$

- We gaan bewijzen dat de matrices A^* en C semi positief definiet zijn . Eerst gaan we bewijzen dat de blokmatrices C_k en A_k semi positief definiet zijn . Hiertoe zullen we laten zien dat

$$\mathbf{u^T} C_k \mathbf{u} \geq 0 \text{ voor alle } u.$$

$$\mathbf{v^T} A_k \mathbf{v} \geq 0 \text{ voor alle } v.$$

  - Er geldt na enig uitwerking (als k oneven is):

$$\begin{aligned} \mathbf{u^T} C_k \mathbf{u} = \; & (a+b)(u_1^2 + u_k^2 + u_{k+1}^2 + u_{2k}^2) + \textstyle\sum_{i=2}^{k-1} 2(a+b)(u_i^2 + u_{i+k}^2) \\ & - \textstyle\sum_{i=2}^{k-1} a(u_i u_{i+1} + u_{k+i} u_{k+i+1}) - 2b(u_1 u_{1+k} + u_k u_{2k}) + \\ & - \textstyle\sum_{i=2}^{k-1} 4b u_i u_{k+i}. \end{aligned}$$

We kunnen makkelijk controleren:

$$u_1 + \sum_{i=2}^{k-1} 2u_i^2 + u_k^2 = \sum_{i=1}^{k-1} (u_i^2 + u_{i+1}^2),$$

$$u_{k+1} + \sum_{i=2}^{k-1} u_{i+k}^2 + u_{2k}^2 = \sum_{i=1}^{k-1} (u_{i+k}^2 + u_{i+k+1}^2),$$

waaruitvolgt :

$$\begin{aligned}
\mathbf{u^T} C_k \mathbf{u} = \quad & a\Bigg((\textstyle\sum_{i=2}^{k-1} (u_i^2 + u_{i+1}^2) - \sum_{i=2}^{k-1} u_i u_{i+1}) + \\
& + (\textstyle\sum_{i=2}^{k-1} (u_{i+k}^2 + u_{i+k+1}^2) - \sum_{i=2}^{k-1} u_{i+k} u_{i+k+1}) \Bigg) + \\
& + b\Bigg((\textstyle\sum_{i=2}^{k-1} 2(u_i^2 + u_{k+i}^2) - \sum_{i=2}^{k-1} 4 u_i u_{k+i}) + \\
& + ((u_1^2 + u_{1+k}^2 + u_k^2 + u_{2k}^2) - 2(u_1 u_{1+k} + u_k u_{2k})) \Bigg).
\end{aligned}$$

Uit de volgende regel:

$$2xy \leq x^2 + y^2$$

volgt:

$$\sum_{i=2}^{k-1} u_i u_{i+1} \leq \sum_{i=2}^{k-1} (u_i^2 + u_{i+1}^2).$$

$$\sum_{i=2}^{k-1} u_{i+k} u_{i+k+1} \leq \sum_{i=2}^{k-1} (u_{i+k}^2 + u_{i+k+1}^2).$$

$$\sum_{i=2}^{k-1} u_i u_{k+i} \leq \sum_{i=2}^{k-1} (u_i^2 + u_{k+i}^2).$$

$$(u_1 u_{1+k} + u_k u_{2k}) \leq (u_1^2 + u_{1+k}^2 + u_k^2 + u_{2k}^2).$$

Op grond van de bovengenoemde vergelijkingen, krijgen we de volgende schatting voor $\mathbf{u^T} C_k \mathbf{u}$:

$$\mathbf{u^T} C_k \mathbf{u} \geq 0.$$

Dit resultaat geldt uiteraard ook als k even is.
Als we stellen : $\mathbf{x} = (\mathbf{x_1}, ..., \mathbf{x_{m-1}})^{\mathbf{T}}$, dan volgt:

$$\mathbf{x^T} C \mathbf{x} = \sum_{k=1}^{m-1} \mathbf{x_k^T} C_k \mathbf{x_k}.$$

Vervolgens geldt:

$$\mathbf{x^T} C \mathbf{x} \geq 0, \text{ wegens } \mathbf{x_k^T} C_k \mathbf{x_k} \geq 0 \text{ voor } k = 1, ..., m-1.$$

Hiermee wordt aangetoond dat C semi positief definiet is.

– We hebben (als k oneven is):

$$\mathbf{v^T} A_k \mathbf{v} = \begin{pmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \mathbf{v_3} \\ \vdots \\ \vdots \\ \mathbf{v_r} \end{pmatrix}^T \begin{pmatrix} D_1 & J_1 & & & & & \\ J_1 & D_2 & J_1 & & & & \\ & J_1 & D_2 & J_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & J_1 & D_2 & J_1 & \\ & & & & J_1 & D_2 & J_1 \\ & & & & & J_1 & D_1 \end{pmatrix} \begin{pmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \mathbf{v_3} \\ \vdots \\ \vdots \\ \mathbf{v_r} \end{pmatrix}.$$

Uitwerken van $\mathbf{v^T} A_k \mathbf{v}$ geeft:

$$\mathbf{v^T} A_k \mathbf{v} = \sum_{i=1}^{r-1} (\mathbf{v_i^T} D_1 \mathbf{v_i} + \mathbf{v_{i+1}^T} J_1 \mathbf{v_i} + \mathbf{v_i^T} J_1 \mathbf{v_{i+1}} + \mathbf{v_{i+1}^T} D_1 \mathbf{v_{i+1}}).$$

$$\mathbf{v^T} A_k \mathbf{v} = \sum_{i=1}^{r-1} \begin{pmatrix} \mathbf{v_i} \\ \mathbf{v_{i+1}} \end{pmatrix}^T \begin{pmatrix} D_1 & J_1 \\ J_1 & D_1 \end{pmatrix} \begin{pmatrix} \mathbf{v_i} \\ \mathbf{v_{i+1}} \end{pmatrix}.$$

Door $(\mathbf{v_i}, \mathbf{v_{i+1}})^\mathbf{T}$ met $\mathbf{u_i}$ en $\begin{pmatrix} D_1 & J_1 \\ J_1 & D_1 \end{pmatrix}$ met C_1, te vervangen hebben we

$$\mathbf{v^T} A_k \mathbf{v} = \sum_{i=1}^{r-1} \mathbf{u_i^T} C_1 \mathbf{u_i}$$

Hieruit volgt:

$$\mathbf{v^T} A_k \mathbf{v} \geq 0, \text{ wegens } \mathbf{u_i}^\mathbf{T} C_1 \mathbf{u_i} \geq 0.$$

Dit bewijs geldt ook als k oneven is.
Uitwerken van het linkerlid van $\mathbf{x^T} A^* \mathbf{x}$, indien $\mathbf{x}$ is te schrijven als $\mathbf{x} = (\mathbf{x_1}, ..., \mathbf{x_m})^\mathbf{T}$ geeft:

$$\mathbf{x^T} A^* \mathbf{x} = \sum_{k=1}^{m} \mathbf{x_k^T} A_k \mathbf{x_k}.$$

Wegens

$$\mathbf{x_k^T} A_k \mathbf{x_k} \geq 0,$$

volgt

$$\mathbf{x^T} A^* \mathbf{x} \geq 0.$$

Hiermee wordt bewijzen dat de matrix A^* semi positief definiet is.

- Daarna gaan we bewezen dat

$$\mathcal{N}(A^*) = span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$$

$\Longrightarrow$) : $span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\} \subseteq \mathcal{N}(A^*)$?

Door uitwerken van $A^*\mathbf{w_k}$ hebben we:

$$A^*\mathbf{w_k} = A_k\mathbf{1}.$$

We kunnen makkelijk controleren dat de som van elementen van de rijen van de matrices A_k gelijk aan nul is, waaruitvolgt:

$$A_k\mathbf{1} = 0, \text{ voor } k = 1...m.$$

Bijgevolg:

$$A^*\mathbf{w_k} = \mathbf{0}.$$

Dus ligt $A^*\mathbf{w_k} = 0$ in de nulruimte van A^* voor ieder $k = 1...m$. Vervolgens $span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\} \subseteq \mathcal{N}(A^*)$.
$\Longleftarrow)$: $\mathcal{N}(A^*) \subseteq span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$.
We gaan eruit van dat

$$A^*\mathbf{x} = \mathbf{0}. \tag{10.8}$$

waaruitvolgt, indien x is te schrijven als $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m})^{\mathbf{T}}$, $\mathbf{x_k}$ is een deelvector behorend bij laag Ω_k ,

$$A_k\mathbf{x_k} = \mathbf{0}, \text{ voor} k = 1, ..., m.$$

– Eerst gaan we aantonen in het geval van 3 punten in de horizontale richting dat de elementen van $\mathbf{v}$ gelijk aan elkaar zijn, indien $\mathbf{v}$ voldoet aan de volgende vergelijking:

$$A_k\mathbf{v} = \mathbf{0}. \tag{10.9}$$

In dit geval zien de matrix D_1 en J_1 als volgt er uit:

$$D_1 = \begin{pmatrix} a+b & -a & 0 \\ -a & 2(a+b) & 0 \\ 0 & -a & (a+b) \end{pmatrix}, \; J_1 = \begin{pmatrix} -b & 0 & 0 \\ 0 & -2b & 0 \\ 0 & 0 & -b \end{pmatrix}.$$

-We schrijven $\mathbf{v} = (\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_r})^{\mathbf{T}}$. Hier is $\mathbf{v_i}$ een deelvector.
Uit (11.14) hebben we :

$$\begin{pmatrix} D_1 & J_1 & & & & & \\ J_1 & D_2 & J_1 & & & & \\ & J_1 & D_2 & J_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & J_1 & D_2 & J_1 & \\ & & & & J_1 & D_2 & J_1 \\ & & & & & J_1 & D_1 \end{pmatrix} \begin{pmatrix} \mathbf{v_1} \\ \mathbf{v_2} \\ \mathbf{v_3} \\ \vdots \\ \vdots \\ \mathbf{v_r} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \vdots \\ \mathbf{0} \end{pmatrix},$$

waaruitvolgt:

$$D_1\mathbf{v_1} + J_1\mathbf{v_2} = 0 \text{ voor } i = 1. \tag{10.10}$$

$$J_1\mathbf{v_{i-1}} + D_2\mathbf{v_i} + J_1\mathbf{v_{i+1}} = 0 \text{ voor } i = 2...r-1. \tag{10.11}$$

$$J_1\mathbf{v_{r-1}} + D_1\mathbf{v_r} = 0 \text{ voor } i = r. \tag{10.12}$$

Voor ieder $i = 1..r$ schrijven we $\mathbf{v_i}$ als

$$\mathbf{v_i} = \begin{pmatrix} v_{i1} \\ v_{i2} \\ v_{i3} \\ \vdots \\ \vdots \\ v_{ir} \end{pmatrix}.$$

Uitwerken van $\mathbf{v_i}$ in de vergelijkingen (11.15),(11.16)en (11.17) levert het volgende stelsel van vergelijkingen:

voor $i = 1$

$$\begin{aligned} &(a+b)v_{11} - av_{12} - bv_{21} &&= 0 \\ &-av_{11} + 2(a+b)v_{12} - -av_{13} - 2bv_{22} &&= 0 \\ &-av_{12} + (a+b)v_{13} - bv_{23} &&= 0 \end{aligned}$$

voor $2 \le i \le r-1$

$$\begin{aligned} &-bv_{i-1,1} + 2(a+b)v_{i,1} - 2av_{i,2} - bv_{i+1,1} &&= 0 \\ &-2bv_{i-1,2} - 2av_{i,1} + 4(a+b)v_{i,2} - 2av_{i,3} - 2bv_{i+1,2} &&= 0 \\ &-bv_{i-1,3} - 2av_{i,2} + 2(a+b)v_{i,3} - bv_{i+1,3} &&= 0 \end{aligned}$$

voor $i = r$

$$\begin{aligned} &-bv_{r-1,1} + (a+b)v_{r,1} - av_{r,2} &&= 0 \\ &-2bv_{r-1,2} - av_{r,1} + 2(a+b)v_{r,2} - av_{r,3} &&= 0 \text{ voor } i = 2..r-1 \\ &-bv_{r-1,3} - av_{r,2} + (a+b)v_{r,3} &&= 0 \end{aligned}$$

Door optelling van de eerste vergelijking met de derde vegelijking en daarna door het aftrekken van de tweede vergelijking krijgen we na enige uitwerking:
voor $i = 1$

$$(2a+b)(-v_{i,1} + 2v_{i,2} - v_{i,3}) - b(-v_{i+1,1} + 2v_{i+1,2} - v_{i+1,3}) = 0$$

voor $2 \le i \le r-1$

$$-b(v_{i-1,1}+2v_{i-1,2}+v_{i-1,3})+(4a+2b)(-v_{i,1}+2v_{i,2}-v_{i,3})-b(-v_{i+1,1}+2v_{i+1,2}-v_{i+1,3}) = 0$$

voor $i = r$

$$-b(-v_{i-1,1} + 2v_{i-1,2} - v_{i-1,3}) + (4a+2b)(-v_{i,1} + 2v_{i,2} - v_{i,3}) = 0$$

We definieren voor $i = 1...r$ de variable z_i als volgt

$$z_i = -v_{i,1} + 2v_{i,2} - v_{i,3}.$$

Invullen van z_i in de bovengenoemde vergelijkingen geeft de volgende stelsel van vergelijkingen:

$$\begin{aligned}
&(2a+b)z_1 - bz_2 &&= 0\\
&-bz_{i-1} + (4a+2b)z_i - bz_{i+1} &&= 0\\
&-bz_{r-1} + (2a+b)z_r &&= 0
\end{aligned}$$

In matrix-vector vorm :

$$P\mathbf{z} = \mathbf{0},\ waarbij$$

$$P = \begin{pmatrix} (2a+b) & -b & & & \\ -b & 2(2a+b) & -b & & \\ & -b & 2(2a+b) & \ddots & \\ & & \ddots & \ddots & -b \\ & & & -b & (2a+b) \end{pmatrix} \text{ en } \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ \vdots \\ z_r \end{pmatrix}.$$

Volgens stelling van Gersghorin is de matrix P inverteerbaar, bijgevolg:

$$z_i = 0.$$

$$\text{ofwel } -v_{i,1} + 2v_{i,2} - v_{i,3} = 0 \text{ voor } i = 1...r. \tag{10.13}$$

Door derde vergelijking van de eerste vergelijking af te trekken hebben we na enige uitwerking:

voor $i = 1$

$$(2a+b)(v_{i,1} - v_{i,3}) - b(v_{i+1,1} - v_{i+1,3}) = 0$$

voor $2 \leq i \leq r-1$

$$-b(v_{i-1,1} - v_{i-1,3}) + 2(2a+b)(v_{i,1} - v_{i,3}) - b(v_{i+1,1} - v_{i+1,3}) = 0$$

voor $i = r$

$$-b(v_{i-1,1} - v_{i-1,3}) + 2(2a+b)(v_{i,1} - v_{i,3}) = 0$$

We definieren voor $i = 1...r$ de variable t_i als volgt

$$t_i = v_{i,1} - v_{i,3}.$$

Invullen van t_i in de bovengenoemde vergelijkingen geeft het volgende stelsel van vergelijkingen:

$$\begin{aligned}
&(2a+b)t_1 - bt_2 &&= 0\\
&-bt_{i-1} + 2(2a+b)t_i - bt_{i+1} &&= 0\\
&-bt_{r-1} + (2a+b)t_r &&= 0
\end{aligned}$$

In matrix-vector vorm :

$$Q\mathbf{t} = \mathbf{0},\ waarbij$$

$$Q = \begin{pmatrix} (a+b) & -b & & & \\ -b & 2(a+b) & -b & & \\ & -b & 2(a+b) & \ddots & \\ & & \ddots & \ddots & -b \\ & & & -b & (a+b) \end{pmatrix} \text{ en } \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ \vdots \\ t_r \end{pmatrix}.$$

Volgens stelling van Gersghorin is de matrix Q inverteerbaar, bijgevolg:

$$t_i = 0 \text{ voor } i = 1...r.$$

$$\text{ofwel } v_{i,1} - v_{i,3} = 0 \text{ voor } i = 1...r. \tag{10.14}$$

Uit de vergelijkingen (10.13) en (10.14) volgt:

$$v_{i,1} = v_{i,1} = v_{i,3}, \text{ voor } i = 1...r.$$

– Op analoge wijze kunnen we laten zien in het algemeen voor willekeurige horizontale punten dat uit $A_k\mathbf{v} = \mathbf{0}$, voor$k = 1, ..., m$, volgt dat de elementen van $\mathbf{v}$ gelijk aan elkaar zijn. Als direct gevolg daarvan is $\lambda_1(A_k) = 0$

  De vergelijkingen (11.15),(11.16) en (11.17) worden als volgt geschreven:

$$\begin{array}{ll} D_1\mathbf{v_1} + J_1\mathbf{v_2} & = 0 \\ J_1\mathbf{v_1} + D_2\mathbf{v_2} + J_1\mathbf{v_3} & = 0 \\ J_1\mathbf{v_2} + D_2\mathbf{v_3} + J_1\mathbf{v_4} & = 0 \\ \vdots & \\ J_1\mathbf{v_{r-1}} + D_2\mathbf{v_r} & = 0 \end{array}$$

  Anders geschreven:

$$\begin{array}{ll} (D_1 + J_1)\mathbf{v_1} + J_1(\mathbf{v_2} - \mathbf{v_1}) & = 0 \\ J_1(\mathbf{v_1} - \mathbf{v_2}) + (D_2 + J_2)\mathbf{v_2} + J_1(\mathbf{v_3} - \mathbf{v_2}) & = 0 \\ J_1(\mathbf{v_2} - \mathbf{v_3}) + (D_2 + J_2)\mathbf{v_3} + J_1(\mathbf{v_4} - \mathbf{v_3}) & = 0 \\ \vdots & \\ J_1(\mathbf{v_{r-1}} - \mathbf{v_r}) + (D_1 + J_1)\mathbf{v_r} & = 0 \end{array}$$

  We kunnen makkelijk nagaan dat voor ieder vector $\mathbf{v_i}$ waarvan de elememten gelijk aan elkaar zijn geldt:

$$(D_1 + J_1)\mathbf{v_i} = (D_2 + J_2)\mathbf{v_i} = 0 \text{ voor } i = 1..r$$

  , waaruitvolgt:

$$\begin{array}{ll} J_1(\mathbf{v_2} - \mathbf{v_1}) & = 0 \\ J_1(\mathbf{v_1} - \mathbf{v_2}) + J_1(\mathbf{v_3} - \mathbf{v_2}) & = 0 \\ J_1(\mathbf{v_2} - \mathbf{v_3}) + J_1(\mathbf{v_4} - \mathbf{v_3}) & = 0 \\ \vdots & \\ J_1(\mathbf{v_{r-1}} - \mathbf{v_r}) + (D_1 + J_1)\mathbf{v_r} & = 0 \end{array}$$

Omdat J_1 is inverteerbaar er geldt:

$$\begin{aligned}
&\mathbf{v_2} - \mathbf{v_1} = 0 && = 0\\
&\mathbf{v_1} - 2\mathbf{v_2} + \mathbf{v_3} && = 0\\
&\mathbf{v_2} - 2\mathbf{v_3} + \mathbf{v_4} && = 0\\
&\vdots\\
&\mathbf{v_{r-1}} - \mathbf{v_r} && = 0
\end{aligned}$$

Hieruitvolgt: $\mathbf{v_1} = \mathbf{v_2} = ... = \mathbf{v_r}$
Hiermee wordt bewezen dat alle elementen van $\mathbf{v}$ gelijk aan elkaar zijn.

- De vector $\mathbf{x}$ die aan (10.8) voldoet, kan op de volgende wijze gesplitst worden:

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{0} \end{pmatrix} + \cdots + \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{x_m} \end{pmatrix}.$$

Als we de elementen van $\mathbf{x_k}$ gelijk aan een constante c_k stellen, volgt:

$$\mathbf{x_k} = c_k\mathbf{1}\text{voor}\, k = 1, .., m. \tag{10.15}$$

Uit (10.15) volgt:

$$\mathbf{x} = c_1 \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} + c_2 \begin{pmatrix} \mathbf{0} \\ \mathbf{1} \\ \vdots \\ \mathbf{0} \end{pmatrix} + \cdots + c_m \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{1} \end{pmatrix}.$$

Ofwel:

$$\mathbf{x} = c_1\mathbf{w_1} + c_2\mathbf{w_2} + ... + c_m\mathbf{w_m}.$$

Hiermee wordt bewezen dat : $\mathcal{N}(A^*) \subseteq span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$.
Tenslotte :

$$\mathcal{N}(A^*) = span\{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_m}\}$$

Bibliografie

[1] D.K.Gartling and J.N.Reddy. The finite element method in heat transfer and fluid dynamics. Crc press, Delft, 1994.

[2] J.Bear. Dynamics of fluids in porous media. American elsevier, Dordrecht/Bosten/Londen, 1972.

[3] A.Segal J.van Kan. Numerieke methoden voor partiele differentiaalvergelijkingen. Delfse uitgevers maatschappij, Delft, 1993.

[4] David McGeary and C.Plummer. *Ontdek de Geologie*. 1994.

[5] Melvyn R.Giles. A quantitative perspective, implications for basin modelling and rock property prediction. Kluwer academic publischers, Dordrecht/Bosten/Londen.

[6] C. Vuik. Voortgezette numerieke lineare algebra. Diktaat, Delft, oktober 1996.

[7] C. Vuik, A. Segal, and J.A. Meijerink. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. *J.Comp Phys*, 152:385–403, 1999.

[8] C. Vuik, A. Segal, J.A. Meijerink, and G.T. Wijma. The construction of projection vectors for a Deflated ICCG method applied to problems with extreme contrasts in the coefficients. Report EP2000-8019, Shell, Rijswijk, 2000. **http://ta.twi.tudelft.nl/users/vuik/Shell-2000.pdf**.

[9] G.T. Wijma. Calculations of pressures in a layered structuure with large contrasts in permeability. Report, TU Delft, 1999.

[10] L. El yaakoubi. De numerieke benadering van de overdruk in de ondergrong met grote veschillen in permeabiliteit. Scriptie, TU Delft, 2000.