

U i T

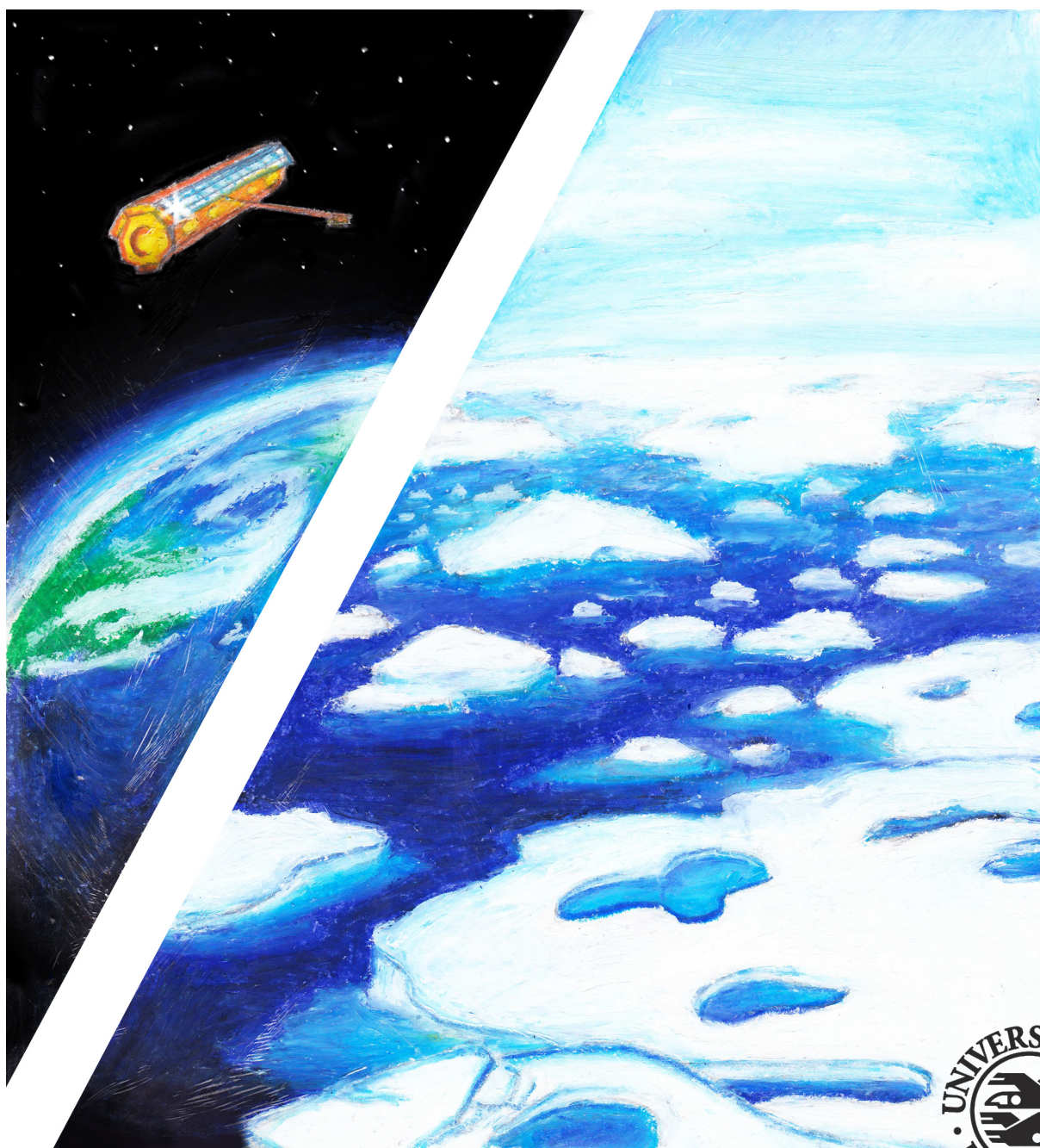
NORGES
ARKTISKE
UNIVERSITET

Fakultet for naturvitenskap og teknologi
Institutt for fysikk og teknologi

Estimering av egenskaper til naturlige klasser i radarbilder med en blandingsmodell

Åse Mari Moe

FYS-3941 Masteroppgave i anvendt fysikk og matematikk, Mai 2018



Sammendrag

I løpet av de siste årene har isdekket i polhavet gått fra å være flerårsis til å bli en sesongbasert istype. Hvordan isen utvikler seg på sommeren vil avhenge av smeltevannandelen, blant annet på grunn av at smeltevannet har mindre albedo enn isen. For å estimere andelen smeltevann er det i denne oppgaven utviklet flere metoder som kan brukes på satellittbilder. Disse metodene bruker en statistisk blandingsmodell og momentmetoden. Momentene som er brukt er lineære- og logaritmiske momenter opp til tredje orden. Totalt er det utviklet fire metoder som ser på intensitetsdata og to som bruker kovariansmatrisa. De statistiske egenskapene varians og forventingsskjevhet er undersøkt gjennom flere simuleringsstudier. Metodene så ut til å fungere så lenge utvalget hadde av en viss størrelse, at ingen av klassene var for små og dersom forventningene til de to fordelingene ikke var for lik. Under sammenligningene så det ut til at estimatorene var konsistente.

Metodene er også blitt testet på et reellt datasett fra satellitten TerraSAR-X, men her så det ikke ut til at metodene fungerte. Estimaten for smeltevannandelen så ut til å generelt ligge for høyt og metodene som brukte kovariansmatrisa så ut til å ha problemer med å skille dataen i to klasser. For et område der bakkesannheten hadde ca. 24 % smeltevannandel estimerte metodene andelen til å være over 67 %. Feilene kan skyldes en kombinasjon av at datasettet lå nær støygulvet og at bildene består av mange blanda piksler og eventuelt flere klasser. Spesielt satellittbildet som lå delvis under støygulvet ga problemer.

Takksigelser

Jeg vil først og fremst takke førsteamanuensis Stian N. Anfinsen for å ha veiledet meg trygt gjennom masteroppgaven. I tillegg vil jeg takke Thomas, Martine og Ane for hjelp med datasettet. Uten hjelp fra Thomas med Gdal, hadde sammenligninga med bakkesannheten vært mangelfull og kanskje ikke fungert. I denne forbindelsen var Martine også til god hjelp. For estimatene av L fikk jeg god hjelp av Johann sin datamaskin og jeg vil takke for lånet.

Jeg vil også takke vennene mine Johann og Margaretha for støtte og hyggelige matpauser. Til slutt vil jeg takke alle som har vært med på å korrekturlese denne oppgaven.



Innhold

1	Introduksjon	7
1.1	Bakgrunn	7
1.2	Problemstilling	8
1.3	Oppbygning av oppgaven	8
2	Teori	11
2.1	De statistiske egenskapene til SAR-målinger	11
2.2	Modell	12
2.2.1	Intensitetsdata	13
2.2.2	Kovariansmatrisa	13
2.2.3	Mulig problem for modellen og alternativer	14
2.3	Momentmetoden	15
2.4	Mellin- og M-transformasjonen	15
2.4.1	Det endimensjonale tilfellet	16
2.4.2	Det flerdimensjonale tilfellet	17
2.5	Momentene til en blandingsmodell	19
2.5.1	Momentene til blandingsmodell med to klasser	19
2.6	Newtons metode	20
2.7	Evaluering av estimatorer	20
3	Metode	23
3.1	Metoder for intensitetsdata	23
3.1.1	Metode 1	24
3.1.2	Metode 2	26
3.1.3	Metode 3	27
3.1.4	Metode 4	27
3.2	Metoder for kovariansmatrisedata	29
3.2.1	Metode 5	29
3.2.2	Metode 6	31
4	Simuleringsstudie: Intensitet	33
4.1	Variierende L	33
4.2	Variierende π_1	37
4.3	Variierende utvalg n	41
4.4	Variierende forhold μ_1/μ_2	44
5	Simuleringsstudie: Kovariansmatrisa	49

5.1	Variierende L	49
5.2	Variierende π_1	52
5.3	Variierende utvalg n	55
5.4	Variierende $ \Sigma_1 / \Sigma_2 $	58
5.5	Variierende korrelasjon	61
6	Reelle data	67
6.1	Datasett	67
6.1.1	Satellittdata	68
6.1.2	Helikopterdata	68
6.2	Satellittbildene T3 og T4	69
6.3	Histogram og estimert blandingsmodell	70
6.4	Resultater sammenlignet med bakkesannhet	73
6.4.1	Histogram til metode 2, 4, 5, 6 og bakkesannhet	73
6.4.2	Sammenligner estimater til metode 2 og 4 med bakkesannheten	76
6.5	Sammenligning med Fors mfl. (2017)	78
7	Konklusjon	81
7.1	Videre arbeid	82
	Bibliografi	85
	Forkortelser	86
	Figurer	88
	Tabeller	89

Kapittel 1

Introduksjon

1.1 Bakgrunn

I denne oppgaven skal det ses på estimering av smeltevannandelen oppå havisen. Denne parameteren er interessant innen klimaforskning og forståelsen av isdynamikken til sommeris. I løpet av de siste årene har mye av isdekket i Polhavet gått fra å være flerårsis til is som er sesongbasert. Is som går under navnet flerårsis har overlevd i minst to somre og er tykkere enn den sesongbaserte (Comiso, 2012; Beckers mfl., 2015). Endringen av isdekket kommer blant annet av varmere klima og lengre smelteperioder (Stroeve mfl., 2014). Stroeve mfl. (2014) fant ut at smelteperioden har økt med 5 dager per tiår fra 1979 til 2013, mens Comiso (2012) fant ut at arealet til havisen har minket med rundt 12% hvert tiår i perioden 1979 til 2011. I tillegg vil mer hav og smeltesdammer oppå isen gi en forsterkende effekt på smeltingen. Denne effekten kommer av at vann reflekterer sollys i mye mindre grad enn det is gjør. Overflatetemperaturen vil da stige og isen smelter forttere (Parkinson og Kellogg, 1979; Comiso, 2012). Albedoen er andelen elektromagnetisk stråling som blir reflektert, og dette varierer med bølgelengdene. Grenfell og Perovich (1984) har sett på albedoen til is ved forskjellige bølgelengder, istyper og årstider. Her var det også sett på hvordan smeltesdammer påvirker isens albedo. Ser man på synlig lys vil albedoen for snø være rundt 0,8, for is vil den være rundt 0,6 og om isen har smeltesdammer vil albedoen variere fra 0,2 til 0,4. Langleben (1971) fant ut at albedoen avhenger lineært med andelen smeltesdammer. Siden man har denne sammenhengen, kan smeltevannandelen være viktig for å forstå isdynamikken om sommeren.

Smeltesdammer er vanddammer oppå isen og det er flere som har studert denne smelteprosessen, blant annet Holt og Digby (1985), Tucker mfl. (1999) og Perovich, Tucker og Ligett (2002). På starten er smeltesdammene grunne og består av smeltevann fra snø. Utover sommeren vil dammene utvikle seg, og etter hvert blir de både større og dypere. Ut over smeltesesongen blir det mer smeltesdammer. Om sesongen er lang og isen ikke er så tykk, kan smeltesdammene smelte gjennom isen. Når det igjen begynner å bli kaldt om høsten, vil smeltesdammen begynne å fryse til (Perovich, Tucker og Ligett, 2002).

For å finne ut hvor stor andel smeltevann som isen består av, kan man bruke satellittbilder. Dette er blant annet gjort i Fors mfl. (2017), Scharien mfl. (2012) og Comiso (2012). Fordelen med satellittdata er at dataen dekker relativt store områder og satellitten går med jevne mellomrom over de samme områdene. En av sensorene som blir brukt på satellitter er syntetisk aparturradar, som ofte forkortes til SAR. Dette er en aktiv sensor, det vil si at systemet selv sender ut elektromagnetisk stråling mot jordoverflaten. På

overflaten vil strålingen bli reflektert og noe av dette vil sensoren motta. Den elektromagnetiske strålingen som brukes er mikrobølger, som i liten grad blir påvirket av atmosfæren rundt jorda (Elachi og Zyl, 2006).

1.2 Problemstilling

Med data fra SAR skal jeg se på om det er mulig å finne en metode for å estimere andelen smeltevann oppå isen. Flere studier har sett på dette, blant annet Fors mfl. (2017) og Scharien mfl. (2012). Metoden som ble brukt i disse studiene var regresjonsmodeller i kombinasjon med isens polarimetriske egenskaper målt med SAR. Dette er modeller som avhenger av at man har bakkesannhet, slik at man kan bestemme parametrene i modellen. Bakkesannheten er direkte observasjoner av egenskaper på overflaten, noe som kan være både dyrt og tidskrevende å samle inn. Dessuten byr det på utfordringer når man ser på is som flyter fritt og skal samlokalisere bakkesannheten og satellittmålinger. Et annet problem er at regresjonsmodellene vil avhenge av vind og innfallsvinkel (Fors mfl., 2017). Yackel og Barber (2000) fant ut at vindforholdene og intensiteten hadde sterkere korrelasjon ved mye vind enn ved lite vind med satellitten RADARSAT-1. Andre faktorer som årstid, istype og temperatur vil også påvirke regresjonsmodellen. Dette kan føre til at regresjonsmodellen blir vanskelig å generalisere og modellen fungerer bare på de forholdene som er representert i bakkesannheten. Det er da ønskelig å finne en metode som er mer overførbar mellom forskjellige forhold. Dette er motivasjonen for å utvikle en metode basert på en statistisk blandingsmodell og de statistiske egenskapene til SAR-data. Fordelen med denne metoden er at den ikke trenger å bli kalibrert mot en bakkesannhet, noe som vil gjøre det lettere å overføre den mellom ulike is- og værforhold. Bakkesannheten vil fortsatt være viktig i denne metoden for å validere hvor godt metoden fungerer.

For å utvikle en metode basert på en blandingsmodell og de statistiske egenskaper til SAR-data over homogene områder, vil jeg bruke estimering basert på momentmetoden. I denne oppgaven er det sett på bruk av både intensitetsdata og kovariansmatrisa. For SAR-sensorer vil det være mulig å bruke forskjellig polarisasjon på signalet som måles og det som blir sendt ut. Ved bruk av kovariansmatrisa er det mulig å bruke flere av disse signalene. Det er også gjort en antagelse om at signalet vil komme fra to klasser og man ser dermed på momentene til en blandingsmodell.

1.3 Oppbygning av oppgaven

I kapittel 2 vil jeg starte med å se på de statistiske egenskapene til SAR-data. Videre i kapitlet vil jeg se på blandingsmodellene mellom to sannsynlighetsfordelinger og hvordan man kan regne ut momentene til slike modeller. Her vil det både bli sett på lineære og logaritmiske momenter. Avhengig av om man ønsker momentene til intensiteten eller determinanten til kovariansmatrisa vil henholdsvis Mellin-transformasjoner og M-transformasjonen bli brukt. Metodene vil bli beskrevet i kapittel 3. For metodene

som er laget er det også sett på flere simuleringstudier og resultatene av disse presenteres i kapittel 4 og 5. Kapittel 4 ser på metodene som bruker intensitetsdata, mens kapittel 5 ser på metoden som bruker kovariansmatrisa. Jeg vil også se på reele data fra satellitten TerraSAR- X. Disse resultatene diskutert i kapittel 6. Kapittel 7 inneholder konklusjonen.

Kapittel 2

Teori

I dette kapitlet ser jeg på bakgrunnen for modellene som skal brukes og hvordan forskjellige momenter til denne statistiske modellen kan bli brukt for å finne smeltevannandelen på overflaten til is. Her blir statistikken bak SAR-data forklart og blandingsmodellen blir presentert. I tillegg går jeg gjennom hvordan man kan finne momenter til blandingsmodellen og hvordan Mellin- og M-transformasjonen kan være nyttig for å finne logaritmiske momenter til en sannsynlighetsfordeling.

2.1 De statistiske egenskapene til SAR-målinger

SAR-sensoren sender ut egne mikrobølger mot bakken. Disse bølgene vil treffe mange små mål og reflekteres, der noe av dette igjen vil mottas av satellitten. Satellitten måler ikke én og én refleksjon, men mange på en gang (Oliver og Quegan, 2004, kap. 4.3). Dette gjør at det man måler blir en sum over mange refleksjoner. Siden man summerer bølger med forskjellige amplituder og faser, vil bølgene noen ganger utslukke hverandre, mens de andre ganger bygger hverandre opp. Dette interferensfenomenet kalles ofte spekkel og gjør at målinga ser ut til å være ganske full av støy, selv om det er dette som er selve signalet (Oliver og Quegan, 2004, kap. 4.3).

Den komplekse spredningen S , som man måler med SAR kan skrives som:

$$S = Ae^{i\phi} = \sum_{k=1}^N A_k e^{i\phi_k}. \quad (2.1)$$

Her er A_k amplituden og ϕ_k er fasen til de enkelte refleksjonene. Ofte kan man anta at fasen er uavhengig fra amplituden, og at den er uniformt fordelt mellom $[-\pi, \pi]$ (Oliver og Quegan, 2004, kap. 4.3). Disse enkeltrefleksjonene summeres sammen, siden sensoren ser et stort område med mange refleksjoner. A og ϕ er henholdsvis amplituden og fasen til det totale signalet man måler. Signalet man måler har både en reell og en imaginær del. Hvis man ser på et større område der man har enkeltrefleksjoner som er statistisk uavhengige, kan man bruke sentralgrenseteoremet til å vise at den imaginære og reelle delen av signalet vil være binormalfordelt. Gjennomsnittet av både den reelle og den imaginære delen vil være null, og variansen vil være lik for begge (Oliver og Quegan, 2004, kap. 4.3). Ser man på amplituden A

til signalet, vil denne være Rayleighfordelt, mens intensiteten I vil være eksponensialfordelt (Oliver og Quegan, 2004, kap. 4.3).

Intensiteten er definert som:

$$I \equiv A^2 = S \cdot S^* \quad (2.2)$$

De fleste SAR-sensorene bruker horisontal og/eller vertikal polarisasjon på mikrobølgene som sendes ut. I tillegg kan man måle den horisontalt og/eller vertikalt polariserte komponenten i det innkomne signalet. Det er da mulig å måle opp til fire spredningskoeffisienter S : to når man måler og sender med samme polarisasjon (S_{hh} og S_{vv}), og to når man måler og sender med forskjellig polarisasjon (S_{hv} og S_{vh}). S_{hh} og S_{vv} kalles ofte kopolariserte målinger, mens S_{hv} og S_{vh} kalles krysspolariserte målinger. De to krysspolariserte målingene er ofte veldig korrelert, og da kan man erstatte S_{hv} og S_{vh} med $S_x = \frac{S_{hv} + S_{vh}}{\sqrt{2}}$. Dette kan også puttes i en vektor S :

$$S = \begin{pmatrix} S_h \\ S_x \\ S_v \end{pmatrix}. \quad (2.3)$$

S kalles ofte spredningsvektoren, og man kan finne kovariansmatrisa til denne vektoren. Kovariansmatrisa til S kan bli funnet ved:

$$\Sigma = E[S \cdot S^*] = E \begin{pmatrix} [S_{hh} \cdot S_{hh}^*] & [S_x \cdot S_{hh}^*] & [S_{hh} \cdot S_{vv}^*] \\ [S_{hh} \cdot S_x^*] & [S_x \cdot S_x^*] & [S_{vv} \cdot S_x^*] \\ [S_{hh} \cdot S_{vv}^*] & [S_x \cdot S_{hh}^*] & [S_{vv} \cdot S_{vv}^*] \end{pmatrix}. \quad (2.4)$$

Siden det er antatt at ϕ_k er uniformt fordelt, vil $E[S] = 0$ (Goodman, 1984). For å estimere kovariansmatrisen C , erstattes forventingen i ligning 2.4 med et middel. Langs diagonalen har man intensitetene. Om man midler over et homogent område med intensitetsverdier, vil middelverdien følge en gammafordeling. Hvor mange verdier man midler over vil reflekteres gjennom formparameteren i denne fordelingen. Når størrelsen på denne øker vil fordelingen gå mot normalfordeling, variansen vil gå ned, mens den romlige oppløsningen vil synke (Oliver og Quegan, 2004, kap. 2.5). Ser man på hele matrisa C under ett, vil denne følge kompleks Wishartfordeling over homogene områder (Oliver og Quegan, 2004, kap. 11.6).

2.2 Modell

For å utvikle en modell av statistiske egenskaper til SAR-signalet, gjør man noen antagelser om området man ønsker å se på. Ut fra egenskapene til SAR-sensoren er gammafordeling naturlig å anta, hvis man ser på det midla intensitetsbildet over homogene områder. Man kan også anta kompleks Wishart-fordeling, om man istedet ser på hele kovariansmatrisa C . I tillegg ønsker jeg å se på en situasjon der man har en blandingsmodell med to klasser. Disse to klassene vil ha to ulike bidrag, og det er disse bidragene man ønsker å finne, siden dette vil gi blandingsforholdet mellom klassene. Her er det i utgangspunktet tenkt å se på smeltevann på is, så de to klassene er da is og smeltevann.

2.2.1 Intensitetsdata

Hvis man bruker en av de polarimetriske kanalene og ser på intensitetbildet, kan man lage en modell. Denne modellen vil si noe om sannsynligheten for at en måling fra et bestemt område vil ta en spesifikk intensitetsverdi. Denne modellen vil være på formen (Frühwirth-Schnatter, 2006, kap. 1):

$$p_X(x) = \pi_1 \cdot p_{X_1}(x) + \pi_2 \cdot p_{X_2}(x), \quad (2.5)$$

hvor π_1 er sannsynligheten for klasse 1, mens π_2 er sannsynligheten for klasse 2. Dette kan man se på som hvor stor andel som er av den ene eller andre klassen. Man har også at de summerte sannsynlighetene til klasse 1 og 2 er $\pi_1 + \pi_2 = 1$. I tillegg vil $0 \leq \pi_i \leq 1$. Modellen inneholder to gammafordelinger, $p_{X_1}(x)$ og $p_{X_2}(x)$, med ulike parametre.

Gammafordelingen er gitt ved (Walploe mfl., 2014, kap. 6):

$$p_{X_i}(x; \mu_i, L) = \left(\frac{L}{\mu_i}\right)^L \cdot \frac{x^{L-1}}{\Gamma(L)} \cdot e^{-\frac{Lx}{\mu_i}}. \quad (2.6)$$

Gammafunksjonen er gitt ved (Walploe mfl., 2014, kap. 6):

$$\Gamma(L) = \int_0^{\infty} t^{L-1} \cdot e^{-t} dt. \quad (2.7)$$

Her antar man at L er den samme for de to klassene, og at μ_1 og μ_2 er forskjellig. Denne antagelsen er basert på at L har sammenheng med hvordan man prosesserer SAR-data og filtrerer bildet. I utgangspunktet vil L være lik antall piksler man midler over, men siden nabopikselene er korrelerte, vil også dette ha påvirkning på L (Oliver og Quegan, 2004, kap. 4.3). I denne oppgaven antar jeg at L er kjent, selv om jeg vil bruke et program til å estimere L og denne også vil ha usikkerhet (Anfinsen, Doulgeris og Eltoft, 2009). For blandingsmodellen vil forventningen være:

$$\mu = E[X] = \pi_1 \mu_1 + \pi_2 \mu_2. \quad (2.8)$$

Ut fra ligning (2.5) kan man finne uttrykkene for forskjellige momenter. Ligning (2.8) er et eksempel på et slikt moment. Momentene kan brukes med momentmetoden for å få estimater for de ukjente parametrene π_1 , π_2 , μ_1 og μ_2 . Denne metoden kommer jeg tilbake til i delkapittel 2.3.

2.2.2 Kovariansmatrisa

Bruker man kovariansmatrisa, vil modellen for overflaten være på formen:

$$p_{\mathbf{C}}(\mathbf{C}) = \pi_1 \cdot p_{\mathbf{C}_1}(\mathbf{C}) + \pi_2 \cdot p_{\mathbf{C}_2}(\mathbf{C}). \quad (2.9)$$

Denne modellen er nesten lik modellen for intensitetbildet. Forskjellen er at de to gammafordelingene er

byttet ut med to komplekse Wishart-fordelinger $p_{C_1}(\mathbf{C})$ og $p_{C_2}(\mathbf{C})$.

Kompleks Wishart-fordeling er gitt ved (Mathai, 1997; Anfinsen, 2016):

$$p_{C_i}(\mathbf{C}; \boldsymbol{\Sigma}_i, L) = \frac{L^{Ld}}{\Gamma_d(L)} \cdot \frac{|\mathbf{C}|^{L-d}}{|\boldsymbol{\Sigma}_i|^L} \cdot e^{-tr(L\boldsymbol{\Sigma}_i\mathbf{C})} \quad (2.10)$$

Kovariansmatrisen \mathbf{C} har dimensjon $d \times d$, $tr(\cdot)$ er sporet til matrisa, $|\cdot|$ er determinanten til matrisa og $\Gamma_d(L)$ er den multivariat gammafunksjonen. Her antas det at L er den samme for de to klassene, mens $\boldsymbol{\Sigma}_1$ og $\boldsymbol{\Sigma}_2$ er forskjellig. Den multivariable gammafunksjonen er gitt ved (Anfinsen, 2016):

$$\Gamma_d(L) = \pi^{\frac{d(d-1)}{2}} \prod_{i=0}^{d-1} \Gamma(L-i) \quad (2.11)$$

Her er $\Gamma(\cdot)$ gammafunksjonen definert i ligning (2.7). Ser man på determinanten til kovariansmatrisa er det mulig å bruke dette sammen med momentmetoden for å estimere de ukjente parametrene.

2.2.3 Mulig problem for modellen og alternativer

Jeg har tenkt at antagelsen om at pikslene er fra enten den ene eller den andre klassen kan være problematisk. Dette kommer blant annet av at modellen forutsetter at oppløsningen på bildene må være såpass god at man får piksler som består av bare is eller vann. Dette vil antagelig ikke stemme for mange av pikslene, der man også vil ha piksler som består av refleksjoner fra både is og vann. Tucker mfl. (1999) observerte smeltesdammer som varierte fra 2 m² til 8000 m². Gjennomsnittet var på 62 m², mens medianen var 14 m². Det tyder på at det er flere små dammer enn store. Hvor mye størrelsen varierer avhenger blant annet av hvor flat isen er (Fetterer og Untersteiner, 1998). Divine mfl. (2015) observerte også at den sene sommerisen kunne ha kompliserte strukturer. Disse observasjonene kan tyde på at det kan oppstå problemer og man kan få blanda piksler. Et resultat av dette er at man også kan få flere klasser, noe metoden ikke er laget for. Hvor mange blanda piksler man har vil også avhenge av oppløsninga (Campbell og Wynne, 2011, kap. 10), der man ved høy oppløsning har mindre andel blanda piksler. Det kan tenkes at en metode som baserer seg direkte på de komplekse dataene, og ikke går veien om estimeringen av intensitet eller kovariansmatrisa, kan gi bedre estimater. Grunnen til dette er at de komplekse dataene vil ha den høyeste tilgjengelige oppløsningen, noe som kan føre til at antagelsen om rene piksler stemmer bedre. Ut fra teorien har man at de målte dataene fra en kanal vil være kompleks binormalfordelt. Dette kan kanskje brukes for å utvikle en nye metode som ligner på den foreslåtte statistiske modellen. Det kan også tenkes at det er mulig å utvikle en statistisk modell som ser på blanda piksler. I en slik modell vil π_1 og π_2 også være stokastiske og følge en fordeling. Her kan for eksempel betafordeling være et alternativ, siden denne fordelingen er definert mellom 0 og 1, som er det samme område hvor π_1 og π_2 er definert. Disse modellene er ikke utviklet eller brukt i denne oppgaven, siden ideen kom som et resultat av at jeg prøvde å forstå hvorfor metoden så ut til å ikke fungere som planlagt. Dette vil jeg gå dypere inn på i kapittel 6.

2.3 Momentmetoden

Momentmetoden er en metode som bruker teoretiske uttrykk for momenter til å skaffe seg ligningsett. De teoretiske momentene vil være en funksjon av parametrene i fordelingen, og hvis man setter uttrykkene for de teoretiske momentene lik de tilsvarende empiriske momentene, kan man løses ut parametrene. De teoretiske momentene blir da erstattet med de estimerte momentene:

$$\begin{array}{lll}
 m_1 = E[X] & \text{erstattes med} & \hat{m}_1 = \frac{1}{n} \sum_{i=1}^n x_i \\
 m_2 = E[X^2] & \text{erstattes med} & \hat{m}_2 = \frac{1}{n} \sum_{i=1}^n x_i^2 \\
 \dots & & \dots \\
 m_a = E[X^a] & \text{erstattes med} & \hat{m}_a = \frac{1}{n} \sum_{i=1}^n x_i^a
 \end{array}$$

Hvis man har et uttrykk for m_a som funksjon av for eksempel μ , kan man erstatte det teoretiske momentet m_a med det estimerte momentet \hat{m}_a . Ut fra dette kan man få en estimator for μ , som er avhengig av \hat{m}_a . Har man flere ukjente, kan man kombinere flere momenter (Upton og Cook, 2014). Disse rå momentene kan man også kombinere for å få sentrale momenter som: varians σ^2 , unormalisert skjevhet γ og unormalisert spissitet (kurtosis) κ .

$$\begin{aligned}
 \mu &= E[X] \\
 \sigma^2 &= E[(X - \mu)^2] = E[X^2] - \mu^2 \\
 \gamma &= E[(X - \mu)^3] = E[X^3] - 3\mu\sigma^2 - \mu^3 \\
 \kappa &= E[(X - \mu)^4] = E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4
 \end{aligned} \tag{2.12}$$

2.4 Mellin- og M-transformasjonen

For å kunne bruke momentmetoden trenger man uttrykkene for noen momenter fra en gitt sannsynlighetsfordeling. Disse momentene kan være de tradisjonelle lineære momentene $E[X^a]$, men man kan også se på de logaritmiske momentene $E[\ln(X)^b]$. For å finne uttrykk for de logaritmiske momentene er Mellin-transformasjonen nyttig, men denne transformasjonen kan også brukes for å finne de lineære momentene. Ser man på matriser som er hermitisk der alle egenverdiene enten er positive, negative eller null, er det mulig å bruke M-transformasjonen (Mathai, 1997). På samme måte som Mellin-transformasjonen, kan denne transformen brukes til å finne momenter til sannsynlighetsfordelinger.

2.4.1 Det endimensjonale tilfellet

Mellin-transformasjonen kan brukes på all funksjoner $f(x)$, der $x > 0$. I denne oppgaven vil jeg begrense bruken til sannsynlighetsfordelingen $p_X(x)$, der transformasjonen vil bli brukt til å finne momentene til fordelingen.

Mellin-transformasjonen kan bli generelt definert som (Anfinsen og Eltoft, 2011):

$$M[f(x)](s) = \int_0^{\infty} x^{s-1} \cdot f(x) dx. \quad (2.13)$$

Bruker man sannsynlighetsfordelingen $p_X(x)$, kan man finne de logaritmiske momentene til denne fordelingen av orden b (Anfinsen og Eltoft, 2011):

$$E[\ln(x)^b] = \frac{d^b}{ds^b} M[p_X(x)](s) \Big|_{s=1}. \quad (2.14)$$

Ut fra Mellin-transformasjonen kan man også gjenkjenne forventningen til de lineære momentene $E[x^a]$ til sannsynlighetsfordelingen. Dette får man ved å sette $s = a + 1$ inn i ligningen for $M[p_X(x)](s)$:

$$E[x^a] = M[p_X(x)](a + 1). \quad (2.15)$$

Mellin-transformen og momentene til gammafunksjon

Bruker man Mellin-transformasjonen på gammafordelingen får man at:

$$M[p_X(x)](s) = \left(\frac{L}{\mu_i}\right)^{1-s} \frac{\Gamma(s + L - 1)}{\Gamma(L)}. \quad (2.16)$$

Ut fra dette kan man finne ulike momenter for gammafordelingen. Ved hjelp av ligning (2.14) kan man finne de logaritmiske momentene og ved hjelp av ligning (2.15) kan man finne de lineære momentene. Dette kan igjen kombineres med ligningene i (2.12), for å finne varians, unormalisert skjevhet og unormalisert spisshet. I tabell 2.1 er disse regnet ut for både de lineære og de logaritmiske momentene. For å regne ut de lineære momentene er det brukt (Walpløe mfl., 2014, kap. 6):

$$\Gamma(L) = (L - 1)\Gamma(L - 1). \quad (2.17)$$

For de logaritmiske momentene er polygammafunksjonen brukt. Polygammafunksjonen av orden 0 er definert som:

$$\psi^{(0)}(L) \equiv \frac{\Gamma'(L)}{\Gamma(L)}. \quad (2.18)$$

Polygammafunksjonen av orden r er:

$$\psi^{(r)}(L) = \frac{d}{dL^r} \psi^{(0)}(L). \quad (2.19)$$

	μ	σ^2	γ	κ
$x:$	μ_i	$\frac{\mu_i^2}{L}$	$\frac{2\mu_i^3}{L^2}$	$\frac{3\mu_i^4(L+2)}{L^3}$
$\ln(x):$	$\psi^{(0)}(L) - \ln\left(\frac{L}{\mu_i}\right)$	$\psi^{(1)}(L)$	$\psi^{(2)}(L)$	$\psi^{(3)}(L) + 3\psi^{(1)}(L)^2$

Tabell 2.1: Utregnede verdier av μ , σ^2 , γ og κ for gammafunksjonen

2.4.2 Det flerdimensjonale tilfellet

Har man et flerdimensjonalt tilfelle med en d ganger d hermitisk matrise, der matrisen også enten er positivt definit, negativt definit, eller null, så kan man bruke en transformasjon som ligner på Mellin-transformasjonen (Mathai, 1997; Anfinsen og Eltoft, 2011). Denne transformasjonen kalles M-transformasjonen, og kan ses på som en matrise-variant av Mellin-transformasjonen. I en hermitisk matrise er alle egenverdiene reelle (Anton og Rorres, 2011, kap. 7.5), og om matrisen er positivt definit er alle egenverdiene til matrisen positive. I en negativt definit matrise er alle egenverdiene negative (Anton og Rorres, 2011, kap. 7.3). Kovariansmatriser oppfyller disse kravene, og med denne metoden er det mulig å finne momentene til determinanten.

Generelt kan matrisen defineres for alle funksjoner $f(\mathbf{C})$, der matrisene oppfyller kravene. M-transformasjonen blir da definert som (Mathai, 1997; Anfinsen og Eltoft, 2011):

$$\mathcal{M}[f(\mathbf{C})](s) = \int_{\Omega^+} |\mathbf{C}|^{s-d} \cdot f(\mathbf{C}) d\mathbf{C} \quad (2.20)$$

Siden jeg vil fokusere på sannsynlighetsfordelinger $p_{\mathbf{C}}(\mathbf{C})$, bruker jeg denne funksjonene i transformasjonen. Ut fra dette kan man finne de logaritmiske momentene av orden b (Anfinsen og Eltoft, 2011):

$$E[\log(|\mathbf{C}|)^b] = \frac{d^b}{ds^b} \mathcal{M}[p_{\mathbf{C}}(\mathbf{C})](s) \Big|_{s=d} \quad (2.21)$$

Ut fra M-transformasjonen kan man også gjenkjenne forventningen til de lineære momentene $E[|\mathbf{C}^a|]$. Disse får man ved å sette $s = a + d$ i ligning (2.20).

$$E[|\mathbf{C}|^a] = \mathcal{M}[p_{\mathbf{C}}(\mathbf{C})](a + d) \quad (2.22)$$

M-transformen og momentene til kompleks Wishart-fordeling

Kovariansmatrisa er hermitisk og positivt semidefinit, og det antas at den er komplekst Wishart-fordelt. Det er da mulig å bruke M-transformasjonen for å finne momentene til determinanten av kovariansmatrisa. M-transformasjonen til kompleks Wishart-fordeling er:

$$M[p_{\mathbf{W}}(\mathbf{W})](s) = \frac{\Gamma_d(L + s - d)}{\Gamma(L)} |\boldsymbol{\Sigma}_i|^{s-d} \quad (2.23)$$

Her er $\mathbf{W} = L \cdot \mathbf{C}$, der L er antatt å være en konstant. For å få uttrykket til \mathbf{C} , kan man bruke ligning:

$$|\mathbf{W}| = |L \cdot \mathbf{C}| = L^d \cdot |\mathbf{C}| \quad (2.24)$$

Ligning (2.22) og (2.21) kan videre brukes i kombinasjon med ligning (2.12), slik at man kan finne uttrykkene for varians, unormalisert skjevhet og unormalisert spisshet. Disse uttrykkene vises i tabellen 2.2. For de lineære uttrykkene er dimensjonen begrenset til $d = 2$ og $d = 3$. Dette er de interessante dimensjonene når man ser på kovariansmatrisa til SAR-data, der kovariansmatrisa ofte er en 3 ganger 3 matrise som i ligning (2.4). Dette gjelder blant annet for full-polariserte radarer som: RADARSAT-2 (MDA, 2016) og ALOS-PALSAR (JAXA, 2016). Data fra TerraSAR-X har som oftest ikke med kryssleddene S_{hv} og S_{vh} (Fritz, 2007), så bruker man data fra denne satellitten vil man bare ha en kovariansmatrise med dimensjon $d=2$.

For å forenkle uttrykkene i tabell 2.2, har jeg valgt å definere $L_3 \equiv (L - 2)(L - 1)^2$. Dette er brukt for de lineære momentene med $d = 3$. Uttrykkene for μ , σ^2 , γ og κ på logaritmisk skala inneholder høyereordenspolygammafunksjon $\psi_d^{(r)}(L)$. Denne funksjonen er definert som:

$$\psi_d^{(r)}(L) \equiv \sum_{i=0}^{d-1} \psi^{(r)}(L - i) \quad (2.25)$$

	μ	σ^2	γ	κ
$ \mathbf{C} $: Ser på $d=2$	$ \boldsymbol{\Sigma}_i \cdot \frac{L-1}{L}$	$2 \cdot \boldsymbol{\Sigma}_i ^2 \cdot \frac{L-1}{L^2}$	$ \boldsymbol{\Sigma}_i ^3 \cdot \frac{10L^2 - 8L - 2}{L^4}$	$12 \cdot \boldsymbol{\Sigma}_i ^4 \cdot \frac{L^4 + 5L^3 - 2L^2 - 3L - 1}{L^6}$
$ \mathbf{C} $: Ser på $d=3$	$ \boldsymbol{\Sigma}_i \cdot \frac{L_3}{(L-1)L^2}$	$3 \cdot \boldsymbol{\Sigma}_i ^2 \cdot \frac{L_3}{L^4}$	$12 \cdot \boldsymbol{\Sigma}_i ^3 \cdot \frac{L_3(2L-1)}{L^6}$	$3 \cdot \boldsymbol{\Sigma}_i ^4 \cdot \frac{L_3(9L^4 + 74L^3 - 19L^2 - 28L + 4)}{L^9}$
$ \ln(\mathbf{C}) $:	$\ln \boldsymbol{\Sigma}_i + \psi_d^{(0)}(L) - d \cdot \ln(L)$	$\psi_d^{(1)}(L)$	$\psi_d^{(2)}(L)$	$\psi_d^{(3)}(L) + 3\phi_d^{(1)}(L)^2$

Tabell 2.2: Utregnede verdier av μ , σ^2 , γ og κ for kompleks Wishart-fordeling

2.5 Momentene til en blandingsmodell

Momentene av orden a til en generell blandingsmodell med n klasse, kan defineres som (Anfinsen, 2016):

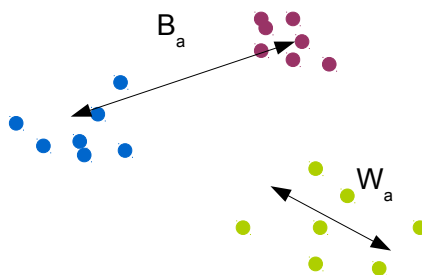
$$E[(X - \mu)^a] = \underbrace{\sum_{i=1}^n \sum_{k=0}^{a-1} \pi_i \binom{a}{k} \delta_i^{a-k} E[(X_i - \mu_i)^k]}_{B_a} + \underbrace{\sum_{i=1}^n \pi_i E[(X_i - \mu_i)^a]}_{W_a} \quad (2.26)$$

I denne ligningen er δ_i definert som $\delta_i \equiv \mu_i - \mu$. Denne formen er generell der n er antall klasser i blandingsmodellen. W_a og B_a blir videre definert som:

$$B_a = \sum_{i=1}^n \sum_{k=0}^{a-1} \pi_i \binom{a}{k} \delta_i^{a-k} E[(X_i - \mu_i)^k], \quad (2.27)$$

$$W_a = \sum_{i=1}^n \pi_i E[(X_i - \mu_i)^a]. \quad (2.28)$$

W_a beskriver variasjonen internt i hver klasse. For en blandingsmodell vil det også være et bidrag til den totale variasjonen som skyldes variasjon mellom klassene. Det er dette B_a måler.



Figur 2.1: W_a måler variasjon innad i egen klasse, mens B_a måler variasjon mellom klassene.

2.5.1 Momentene til blandingsmodell med to klasser

Modellen som skal brukes i denne oppgaven er begrenset til en blandingsmodell med to klasser ($n = 2$). Forenkla uttrykk for B_1 , B_2 , B_3 og B_4 kommer fra Anfinsen (feb. 2016) og er basert på Frühwirth-Schnatter (2006) og Kim og White (2004). For å finne uttrykkene kan man bruke ligninga $\pi_1 + \pi_2 = 1$, ligning (2.27) og ligning (2.8). I de forenklete ligningene er δ definert som $\delta \equiv \mu_1 - \mu_2$, σ_i som variansen til fordelingen p_{X_i} og γ_i som unormalisert skjevhet til fordelingen p_{X_i} . Indeksene i er her 1 eller 2 ut fra hvilken klasse man ser på (Anfinsen, 2016).

$$\begin{aligned} B_1 &= 0, \\ B_2 &= \pi_1 \pi_2 \delta^2, \\ B_3 &= \pi_1 \pi_2 \delta [(\pi_2^2 - \pi_1^2) \delta^2 + 3(\sigma_1^2 - \sigma_2^2)], \\ B_4 &= \pi_1 \pi_2 \delta^2 [(\pi_2^3 + \pi_1^3) \delta^2 + 6(\pi_1 \sigma_2^2 + \pi_2 \sigma_1^2) + 4(\gamma_1 - \gamma_2) / \delta]. \end{aligned}$$

Man kan også finne uttrykkene for W_1 , W_2 , W_3 og W_4 . Disse inneholder, i tillegg til varians σ_i^2 og unormalisert skjevhet γ_i , den unormaliserte spissheten κ_i til p_{X_i} .

$$W_1 = \pi_1\mu_1 + \pi_2\mu_2$$

$$W_2 = \pi_1\sigma_1^2 + \pi_2\sigma_2^2$$

$$W_3 = \pi_1\gamma_1 + \pi_2\gamma_2$$

$$W_4 = \pi_1\kappa_1 + \pi_2\kappa_2$$

2.6 Newtons metode

Momentmetoden er en måte å skaffe ligninger for å løse ut for de ukjente parametrene, men ligningene kan være vanskelig å løse analytisk. Da kan man bruke en numerisk metode, som for eksempel Newtons metode. Denne metoden finner nullpunktene til en funksjon $f(x)$ numerisk. For å bruke denne metoden trenger man også den deriverte $f'(x)$ til funksjonen $f(x)$. Denne metoden konvergerer ofte raskt, og kan brukes når ligningen er vanskelig å løse eksakt (Lindstrøm, 2006).

$$x_0 = \text{startverdi}$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

...

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Metoden stoppes når x_n verdien har blitt nøyaktig nok og dette vil da være en tilnærmet verdi for x . Det er også mulig å sjekke om metoden fungerer ved å teste om: $f(x_{n+1}) \approx 0$. Startverdiene man velger burde velges i nærheten av nullpunktet, hvis ikke kan det hende at metoden ikke konvergerer (Lindstrøm, 2006).

2.7 Evaluering av estimatorer

For å evaluere hvor gode de forskjellige estimatorene er kan man se på egenskaper som varians og forventningsskjevhet. Hvis man antar at estimatoren til parameteren θ er $\hat{\theta}$, så har man at forventningsskjevheten er definert som (Casella og Berger, 2002):

$$\text{forventningsskjevhet}(\hat{\theta}) = E[\hat{\theta}] - \theta. \quad (2.29)$$

For en forventningsrette estimatorer vil dette uttrykket være 0. En konsistent estimator er en estimator som vil bli forventningsrett når antall datapunkter brukt til å estimere $\hat{\theta}$ går mot uendelig. I tillegg vil variansen gå mot 0 (Casella og Berger, 2002). Det er ofte ønskelig å ha en forventningsrett estimator, men også liten varians er en viktig egenskap. MSE er en måte å kombinere disse to egenskapene og er forkortelse for det engelske navnet Mean squared error. På norsk kan dette oversettes med midla kvadratisk feil. MSE er en funksjon av både forventningsskjevhet og variansen, og er definert som (Casella og Berger, 2002):

$$MSE(\hat{\theta}) = E((\theta - \hat{\theta})^2) = \text{Var}(\hat{\theta}) + \left(\text{forventningsskjevhet}(\hat{\theta})\right)^2 \quad (2.30)$$

Kapittel 3

Metode

Til sammen er det i denne avhandling utviklet seks metoder med momentmetoden. I dette kapitlet vil jeg først gå gjennom metode 1 til 4 som tar utgangspunkt i intensiteten til den målte tilbakespredningen. Disse metoden fokuserer på én kanal om gangen. Etterpå vil jeg gå gjennom metode 5 og 6 som tar utgangspunkt i hele kovariansmatrisa.

3.1 Metoder for intensitetsdata

Jeg har laget fire metoder som bruker intensiteten til én kanal. Her er momentmetoden brukt for å finne uttrykk for de fire ukjente parameterne i blandingsmodellen. Når ordenen på momentet øker, vil ofte variansen også øke. Dette er ikke ønskelig, så jeg har prøvd å bruke momenter med lav orden. Jeg har derfor valgt å bruke de tre første logaritmiske momentene av orden 1, 2 og 3, og de to første lineære momentene av orden 1 og 2.

For metodene 1 til 3 blir estimatene π_1 og π_2 funnet på samme måte. Denne metoden gir en analytisk løsning for estimatene til π_1 og π_2 . Forskjellen mellom disse tre metodene er hvordan μ_1 og μ_2 blir estimert. Estimatene μ_1 og μ_2 er ikke hovedfokuset i denne oppgaven, men de er viktige for å bestemme hvilken fysiske klasse estimatene for π_i hører til. Metodene er laget slik at klassen med minst μ_i vil tilsvare klasse 1, slik at $\mu_1 < \mu_2$. I metode 1 til 3 er det mulig å regne ut to estimater for μ_1 og μ_2 , noe som kommer av at estimatet for π_1 har to løsninger. I det første estimatene for μ_1 og μ_2 antas det at π_1 er størst, mens i det andre estimatet antas det at π_1 er minst. For å finne ut hvilken av estimatene som skal brukes, har jeg brukt en ekstra ligning. Under testing så dette ut til å fungere.

Metode 4 estimerer π_1 og π_2 på en annen måte enn metode 1 til 3. Her er π_2 estimert numerisk med Newtons metode. De tre siste estimatene, π_1 , μ_1 og μ_2 , er estimert i etterkant. Metode 4 har ikke det samme problemet som metode 1 til 3 ved estimering av μ_1 og μ_2 . Denne metoden unngår dermed å bruke en ekstra ligning på grunn av tvetydighet. Ellers vil også metode 4 definere klassen med minst μ_i som klasse 1.

Når jeg i de neste avsnittene går dypere inn på de fire metodene, har jeg valgt å bruke x om dataene på lineær skala, og om de er på logaritmisk skala har man: $y = \ln(x)$. Derfor har jeg valgt å bruke $m_{X,a}$ for uttrykket $E[(X - \mu_X)^a]$, der indeksen a angir hvilken orden man ser på. For logaritmiske momenter vil jeg bruke $m_{Y,a}$, for $E[(Y - \mu_Y)^a]$.

3.1.1 Metode 1

Metode 1 er utviklet ved å skaffe seg tre ligninger med momentmetoden. Dette er gjort ved bruk av en blandingsmodell med to gammafordelinger. Det er egentlig fire parametre i modellen som er antatt: π_1 , π_2 , μ_1 og μ_2 , men π_1 og π_2 er lineært avhengige gjennom ligningen: $\pi_2 = 1 - \pi_1$. De tre ligningene er funnet ut fra momentene $m_{X,a}$ og $m_{Y,a}$ til en blandingsmodell. Ligningene som er brukt er:

$$m_{X,2} = W_2[X] + B_2[X] = \frac{(\pi_1\mu_1^2 + \pi_2\mu_2^2)}{L} + \pi_1\pi_2(\mu_1 - \mu_2)^2 \quad (3.1)$$

$$m_{Y,2} = W_2[Y] + B_2[Y] = \psi^{(1)}(L) + \pi_1\pi_2 \ln\left(\frac{\mu_1}{\mu_2}\right)^2 \quad (3.2)$$

$$m_{Y,3} = W_3[Y] + B_3[Y] = \psi^{(2)}(L) + \pi_1\pi_2(\pi_2^2 - \pi_1^2) \ln\left(\frac{\mu_1}{\mu_2}\right)^3 \quad (3.3)$$

Kombinerer man ligning (3.2) og (3.3) kan man få et uttrykk som er uavhengig av μ_1 og μ_2 .

$$\left(\frac{m_{Y,2} - \psi^{(1)}(L)}{\pi_1\pi_2}\right)^3 = \left(\frac{m_{Y,3} - \psi^{(2)}(L)}{\pi_1\pi_2(\pi_2^2 - \pi_1^2)}\right)^2 \quad (3.4)$$

$$\frac{(m_{Y,2} - \psi^{(1)}(L))^3}{(m_{Y,3} - \psi^{(2)}(L))^2} = \frac{\pi_1\pi_2}{(\pi_2^2 - \pi_1^2)^2} \quad (3.5)$$

Den venstre siden av ligning (3.5) avhenger ikke av π_1 og π_2 , og for en gitt L vil dette være konstant. Jeg vil da definere:

$$k_L \equiv \frac{(m_{Y,2} - \psi^{(1)}(L))^3}{(m_{Y,3} - \psi^{(2)}(L))^2} \quad (3.6)$$

I tillegg kan man bruke $\pi_1 + \pi_2 = 1$. Denne ligningen kan kombineres med ligning (3.6), slik at man får et uttrykk for π_1 :

$$\pi_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4k_L + 1}} \quad (3.7)$$

Her kan π_1 være to verdier, og dette gjør at man kan få to svar for μ_1 og to svar for μ_2 . Kombinert med $\pi_1 + \pi_2 = 1$, kan man se at om uttrykket for π_1 bruker det positive fortegnet i \pm , bruker π_2 det negative, og omvendt. For at metoden skal fungere må man også finne uttrykkene til μ_1 og μ_2 . Dette gjøres ved å se på ligning (3.1) og (3.2). Ligning (3.2) kan brukes til å få et uttrykk for forholdet mellom μ_1 og μ_2 :

$$\ln\left(\frac{\mu_1}{\mu_2}\right) = \sqrt{\frac{m_{Y,2} - \psi^{(1)}(L)}{\pi_1\pi_2}} \quad (3.8)$$

Høyre side av ligning (3.8) inneholder ikke μ_1 og μ_2 , så denne er konstant når π_1 , π_2 og L er gitt. Jeg har valgt å definere:

$$k_\pi \equiv \sqrt{\frac{m_{Y,2} - \psi^{(1)}(L)}{\pi_1 \pi_2}} \quad (3.9)$$

Dette gir at forholdet mellom μ_1 og μ_2 er:

$$\frac{\mu_1}{\mu_2} = e^{k_\pi}. \quad (3.10)$$

Kombineres ligningene (3.10) og (3.1), kan man få uttrykkene for μ_1 og μ_2 :

$$\mu_2 = \sqrt{\frac{m_{X,2}}{\frac{\pi_1 e^{2k_\pi} + \pi_2}{L} + \pi_1 \pi_2 (e^{k_\pi} - 1)^2}}, \quad (3.11)$$

$$\mu_1 = \mu_2 e^{k_\pi}. \quad (3.12)$$

Når man har uttrykkene for de tre variablene, kan man bruke momentmetoden. Med momentmetoden erstattes $m_{X,2}$ med et estimatet $\hat{m}_{X,2}$, $m_{Y,2}$ med estimatet $\hat{m}_{Y,2}$ og $m_{Y,3}$ med et estimatet $\hat{m}_{Y,3}$. Dette gir estimatorene:

$$\hat{k}_L = \frac{(\hat{m}_{Y,2} - \psi^{(1)}(L))^3}{(\hat{m}_{Y,3} - \psi^{(2)}(L))^2}$$

$$\hat{\pi}_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4\hat{k}_L + 1}}$$

$$\hat{k}_\pi = \sqrt{\frac{\hat{m}_{Y,2} - \psi^{(1)}(L)}{\hat{\pi}_1 \hat{\pi}_2}}$$

$$\hat{\mu}_2 = \sqrt{\frac{\hat{m}_{X,2}}{\frac{\hat{\pi}_1 e^{2\hat{k}_\pi} + \hat{\pi}_2}{L} + \hat{\pi}_1 \hat{\pi}_2 (e^{\hat{k}_\pi} - 1)^2}}$$

$$\hat{\mu}_1 = \hat{\mu}_2 e^{\hat{k}_\pi}$$

På grunn av muligheten for å få flere svar, vil jeg bruke en ekstra ligning med momentmetoden. Ligningen som er brukt til dette er:

$$m_{X,1} = W_1[X] + B_1[X] = \pi_1 \mu_1 + \pi_2 \mu_2 \quad (3.13)$$

Metoden er laget slik at estimatene $\hat{\mu}_1$ og $\hat{\mu}_2$ først regnes ut med det ene estimatet for $\hat{\pi}_1$. Etterpå regnes det ut to nye estimat for $\hat{\mu}_1$ og $\hat{\mu}_2$, der det andre estimatet for $\hat{\pi}_1$ blir brukt. Hvilken av disse to kombinasjonene man bruker velges ut fra ligningen:

$$\hat{\pi}_1 \hat{\mu}_1 + \hat{\pi}_2 \hat{\mu}_2 - \hat{m}_{X,1}, \quad (3.14)$$

der kombinasjon som gir svaret nærmest null blir valgt.

3.1.2 Metode 2

Metode 2 estimerer π_1 og π_2 på samme måte som metode 1. Metode 1 brukte ligning (3.2) og (3.3), og denne metoden bruker også disse to ligningene. Forskjellen mellom disse to metodene er hvordan man estimerer μ_1 og μ_2 . I stedet for å estimere disse to estimatene ved bruk av estimatet $\hat{m}_{X,2}$, bruker jeg $\hat{m}_{X,1}$. Dette er et moment med lavere orden, noe som kan være en fordel siden dette momentet har lavere varians. Også i denne metoden vil forholdet mellom μ_1 og μ_2 bli som i ligning (3.10). Kombinerer man denne ligningen med (3.13), kan man få følgende ligninger for μ_1 og μ_2 :

$$\mu_2 = \frac{m_{X,1}}{\pi_1 e^{k\pi} + \pi_2}, \quad (3.15)$$

$$\mu_1 = \mu_2 e^{k\pi}. \quad (3.16)$$

Når man har uttrykkene for de tre variablene, kan man bruke momentmetoden. Med denne metoden erstattes $m_{X,1}$, $m_{Y,2}$ og $m_{Y,3}$ med henholdsvis estimatene $\hat{m}_{X,1}$, $\hat{m}_{Y,2}$ og $\hat{m}_{Y,3}$. Dette gir estimatorene:

$$\hat{k}_L = \frac{(\hat{m}_{Y,2} - \psi^{(1)}(L))^3}{(\hat{m}_{Y,3} - \psi^{(2)}(L))^2},$$

$$\hat{\pi}_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4\hat{k}_L + 1}},$$

$$\hat{k}_\pi = \sqrt{\frac{\hat{m}_{Y,2} - \psi^{(1)}(L)}{\hat{\pi}_1 \hat{\pi}_2}},$$

$$\hat{\mu}_2 = \frac{\hat{m}_{X,1}}{\hat{\pi}_1 e^{\hat{k}_\pi} + \hat{\pi}_2},$$

$$\hat{\mu}_1 = \hat{\mu}_2 e^{\hat{k}_\pi}.$$

Siden π_1 og π_2 er estimert på samme måte som i metode 1, trenger også denne metoden en ekstra ligning for å koble sammen riktig π_i og μ_i . Ligningen som er brukt i metode 2 er:

$$m_{Y,1} = W_1[Y] + B_1[Y] = \psi^{(0)}(L) + \pi_1 \ln(\mu_1) + \pi_2 \ln(\mu_2) - \ln(L) \quad (3.17)$$

Metoden regner ut et estimat for $\hat{\mu}_1$ og $\hat{\mu}_2$ med det ene estimatet for $\hat{\pi}_1$, og et estimat for det andre tilfellet. Hvilken av disse to kombinasjonene man bruker velges ut fra ligningen:

$$\psi^{(0)}(L) + \hat{\pi}_1 \ln(\hat{\mu}_1) + \hat{\pi}_2 \ln(\hat{\mu}_2) - \ln(L) - \hat{m}_{Y,1}. \quad (3.18)$$

Kombinasjon som gir svaret nærmest null velges fremfor den andre kombinasjonen.

3.1.3 Metode 3

Metode 3 bruker de tre første logaritmiske momentene til blandingsmodellen $m_{Y,1}$, $m_{Y,2}$ og $m_{Y,3}$. Også her er π_1 og π_2 estimert på samme måte som i metode 1 og forholdet mellom μ_1 og μ_2 er funnet som i ligning (3.10). For å finne μ_1 og μ_2 er moment $m_{Y,1}$ brukt. Med denne metoden får man uttrykkene for μ_1 og μ_2 til å bli:

$$\mu_2 = e^{m_{Y,1} - \psi^{(0)}(L) + \ln(L) - \pi_1 k_\pi}, \quad (3.19)$$

$$\mu_1 = \mu_2 e^{k_\pi}. \quad (3.20)$$

Når man har uttrykkene for de tre variablene kan man bruke momentmetoden, der man erstatter $m_{Y,1}$, $m_{Y,2}$ og $m_{Y,3}$ med henholdsvis $\hat{m}_{Y,1}$, $\hat{m}_{Y,2}$ og $\hat{m}_{Y,3}$. Dette gir estimatorene:

$$\hat{k}_L = \frac{(\hat{m}_{Y,2} - \psi^{(1)}(L))^3}{(\hat{m}_{Y,3} - \psi^{(2)}(L))^2},$$

$$\hat{\pi}_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4\hat{k}_L + 1}},$$

$$\hat{k}_\pi = \sqrt{\frac{\hat{m}_{Y,2} - \psi^{(1)}(L)}{\hat{\pi}_1 \hat{\pi}_2}},$$

$$\hat{\mu}_2 = e^{\hat{m}_{Y,1} - \psi^{(0)}(L) + \ln(L) - \pi_1 \hat{k}_\pi},$$

$$\hat{\mu}_1 = \hat{\mu}_2 e^{\hat{k}_\pi}.$$

For å velge rett μ_1 og μ_2 i metode 3 er ligningen for moment $m_{X,1}$ (ligning 3.13) brukt sammen med momentmetoden. Metode 2 og 3 bruker dermed de samme ligningene, men på ulikt sted i metoden. I metode 3 bruker man:

$$\hat{\pi}_1 \hat{\mu}_1 + \hat{\pi}_2 \hat{\mu}_2 - \hat{m}_{X,1}, \quad (3.21)$$

for å finne de riktige estimatene for $\hat{\mu}_1$ og $\hat{\mu}_2$. Kombinasjonen som gir svaret nærmest null blir valgt fremfor den andre kombinasjonen (se metode 1 og 2).

3.1.4 Metode 4

I metode 4 er moment $m_{X,1}$, $m_{Y,1}$ og $m_{Y,2}$ brukt, dette er henholdsvis ligning (3.13), (3.17) og (3.2). Her er det ikke brukt momenter som har høyere orden enn 2, for å se om det er mulig å lage en metode med bedre variansegenskaper. Ideen bak dette valget er at estimeringsvariansen øker med orden på momentene. Ligning (3.17) har to ledd som bare er avhengig av L. Jeg har dermed valgt å definere:

$$k_{Y1} \equiv m_{Y,1} + \ln(L) - \psi^{(0)}(L) \quad (3.22)$$

For ligning (3.2) definerer jeg:

$$k_{Y2} \equiv \sqrt{m_{Y,2} - \psi^{(1)}(L)} \quad (3.23)$$

På samme måte som i metode 1, 2 og 3, kan jeg finne forholdet mellom μ_1 og μ_2 ut fra ligningen til $m_{Y,1}$ (3.2):

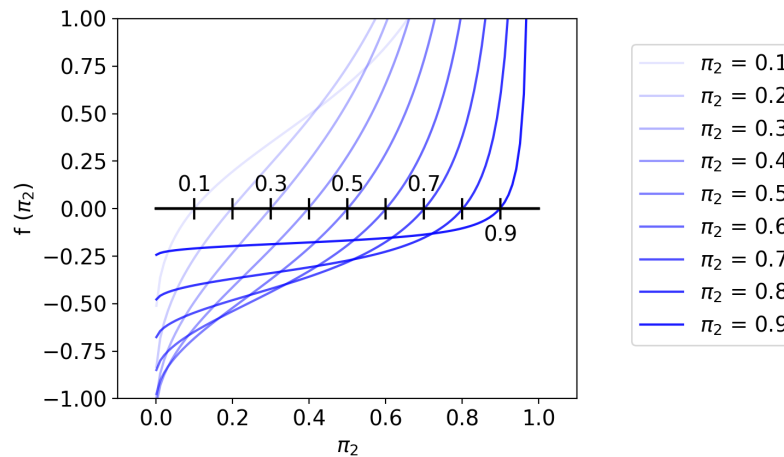
$$\frac{\mu_1}{\mu_2} = e^{k_{Y2}/\sqrt{\pi_1\pi_2}}. \quad (3.24)$$

Kombinert med ligningen for $m_{Y,1}$ (3.17), kan man få et uttrykk for μ_2 :

$$\mu_2 = e^{k_{Y1}-k_{Y2}\sqrt{\frac{\pi_1}{\pi_2}}}. \quad (3.25)$$

Bruker man $\pi_1 + \pi_2 = 1$ og ligning (3.13) for $m_{X,1}$, kan man få en ligning som funksjon av π_2 og L. L antas som kjent. Det er da mulig å løse π_2 ut fra ligningen:

$$f(\pi_2) = e^{k_{Y1}-k_{Y2}\sqrt{\frac{1}{\pi_2}-1}} \left((1-\pi_2)e^{k_{Y2}/\sqrt{\pi_2-\pi_2^2}} + \pi_2 \right) - m_{X,1} = 0 \quad (3.26)$$



Figur 3.1: Funksjonen $f(\pi_2)$, der $m_{X,1}$, $m_{Y,1}$ og $m_{Y,2}$ er estimert ved bruk av 10000 datapunkter for forskjellige verdier av π_2 .

Figur 3.1 viser ni simuleringer med forskjellige estimater for $\hat{m}_{X,1}$, $\hat{m}_{Y,1}$ og $\hat{m}_{Y,2}$. Disse estimatene er gjort fra 10000 datapunkter der π_2 er variert mellom 0,1 og 0,9. For hver av simuleringene er grafene til $f(\pi_2)$ vist for varierende π_2 . Fargen på grafene avhenger av hvor stor π_2 er. Det vil si at for høye verdier av π_2 er grafen mørk, og for lave verdier er den lys. Nullpunktet er marker med en horisontal linje.

Ligningen $f(\pi_2)$ har ikke en analytisk løsning, så jeg har valg å løse ligningen med Newtons metode. Denne metoden er forklart i kapittel 2.6. Metoden ser ut til å ha litt problemer ved noen valg av startverdier, men starter man nærme nullpunktet vil metoden gi svar (Lindstrøm, 2006, kap. 7.3). I utgangspunktet har jeg valgt å alltid starte nær slutten av intervallet ved $\pi_2 = 0,9$. For ligningen $f(\pi_2)$ har det vist seg å gi gode resultater. I tillegg stoppes metoden hvis den ikke har konvergert etter et visst antall runder. Grensen er satt på 50, men metoden har ofte konvergert før 10 runder har gått. Får man likevel problemer vil metoden starte på nytt med en ny startverdi. Estimatene for μ_1 og μ_2 finner man ved ligning (3.25) og (3.24). I disse ligningene bruker man den estimerte verdien av π_2 til å estimere μ_2 og μ_1 .

3.2 Metoder for kovariansmatrisedata

Metode 5 og 6 bruker momentene til determinanten til kovariansmatrisa. I disse metodene brukes de samme ordenene på momentene som i metode 2 og 4. Grunnen til at jeg har valgt å kombinere momentene på samme måte som disse to metodene, er at metode 1, 2 og 3 lignet veldig siden π_1 er estimert på samme måte. Det er ikke så mye å hente ved å se på alle de tre variantene på nytt. Jeg har dermed valgt å bruke bare én av disse variantene. I tillegg vil jeg bruke en metode basert på metode 4, siden denne metoden skilte seg ut ved å estimere $\hat{\pi}_1$ på en ny måte. På bakgrunn av simuleringsstudiene i kapittel 4 har jeg unngått å bruke metode 1, siden den blant annet ga litt dårligere resultater for $\hat{\pi}_1$ nær 0,5. I tillegg virket denne metoden dårligere for lave verdier av n og L . Siden metode 3 var minimalt forskjellig fra metode 2, velger jeg å ikke se på denne metoden heller.

Metodene som er utviklet kan bruke kovariansmatrisa med dimensjon $d = 2$ eller $d = 3$. Disse to dimensjonene er de interessante dimensjonene når man jobber med SAR (se delkapittel 2.1), i tillegg til det endimensjonale tilfellet som ble sett på i metode 1 til 4. Når jeg går dypere inn i metoden fortsetter jeg å bruke \mathbf{C} for kovariansmatrisa, men for determinanten til \mathbf{C} defineres:

$$Z \equiv |\mathbf{C}|.$$

Denne transformasjonen reduserer dataene til én dimensjon, og for kovariansmatrisen vil dette bli et reelt tall. Tar man logaritmen av Z , har jeg definert dette som $\ln(Z) = Q$. For momentene til Z er $m_{Z,a}$ brukt når det er snakk om $E[(Z - \mu_Z)^a]$. Indeksene a sier noen om hvilken orden av momentene som er sett på. For moment på formen $E[(Q - \mu_Q)^a]$, bruker jeg $m_{Q,a}$.

3.2.1 Metode 5

Denne metoden er inspirert av metode 2, der man bruker de samme type momentene. Forskjellen fra metode 2, er at metoden kan brukes på hele kovariansmatrisa. For å gjøre dette bruker man determinanten til matrisa. Ligningene er i denne metoden avhengig av hvor mange dimensjoner d kovariansmatrisa har. Når jeg ser på de lineære momentene, har jeg dermed begrenset dimensjonene til $d = 2$ og $d = 3$. Dette vil si at metodene er laget slik at det er disse dimensjonene den kan ta inn. Ser man på de lineære momentene, har jeg definert:

$$\begin{aligned} k_{d=2} &\equiv \frac{L-1}{L}, \\ k_{d=3} &\equiv \frac{(L-1)(L-2)}{L^2}. \end{aligned} \tag{3.27}$$

Dette gjør at ligningene som skal brukes i momentmetoden blir:

$$m_{Z,1} = W_1[Z] + B_1[Z] = k_{d=2,3} \cdot (\pi_1 |\Sigma_1| + \pi_2 |\Sigma_2|), \quad (3.28)$$

$$m_{Q,2} = W_2[Q] + B_2[Q] = \psi_d^{(1)}(L) + \pi_1 \pi_2 \cdot \ln \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right)^2, \quad (3.29)$$

$$m_{Q,3} = W_3[Q] + B_3[Q] = \psi_d^{(2)}(L) + \pi_1 \pi_2 (\pi_2^2 - \pi_1^2) \cdot \ln \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right)^3. \quad (3.30)$$

På samme måte som i metode 1 til 3, kan man bruke de to logaritmiske momentene med orden 2 og 3 til å få en ligning som ikke avhenger av $|\Sigma_1|$ og $|\Sigma_2|$. Dette tilsvarer å kombinere ligning (3.29) og (3.30):

$$\left(\frac{m_{Q,3} - \psi_d^{(2)}(L)}{\pi_1 \pi_2 (\pi_2^2 - \pi_1^2)} \right)^3 = \left(\frac{m_{Q,2} - \psi_d^{(1)}(L)}{\pi_1 \pi_2} \right)^2 \quad (3.31)$$

$$\frac{(m_{Q,2} - \psi_d^{(1)}(L))^3}{(m_{Q,3} - \psi_d^{(2)}(L))^2} = \frac{\pi_1 \pi_2}{(\pi_2^2 - \pi_1^2)^2} \quad (3.32)$$

Utrykket til venstre er konstant om man antar at L og d er kjent. Dette uttrykket har jeg valgt å definere som:

$$k_{L,d} \equiv \frac{(m_{Q,2} - \psi_d^{(1)}(L))^3}{(m_{Q,3} - \psi_d^{(2)}(L))^2} \quad (3.33)$$

Bruker man også den kjente ligningen: $\pi_1 + \pi_2 = 1$, kan man få et uttrykk som bare er avhengig av π_1 :

$$\pi_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4k_{L,d} + 1}}. \quad (3.34)$$

Med ligning (3.29) kan man få et uttrykk for forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$:

$$\ln \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) = \sqrt{\frac{m_{Q,2} - \psi_d^{(1)}(L)}{\pi_1 \pi_2}} \quad (3.35)$$

Den høyre siden av ligningen inneholder ikke $|\Sigma_1|$, eller $|\Sigma_2|$, så den er konstant om man allerede antar at man har funnet et estimat for π_1 og π_2 . Jeg definerer da $k_{\pi,d}$:

$$k_{\pi,d} = \sqrt{\frac{m_{Q,2} - \psi_d^{(1)}(L)}{\pi_1 \pi_2}}. \quad (3.36)$$

Dette gir forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$:

$$\frac{|\Sigma_1|}{|\Sigma_2|} = e^{k_{\pi,d}}. \quad (3.37)$$

Kombinerer man ligning (3.36) med (3.28), får man ligningen for $|\Sigma_1|$ og $|\Sigma_2|$:

$$|\Sigma_2| = \frac{m_{Z,1}}{\pi_1 e^{k_{\pi,d}} + \pi_2}, \quad (3.38)$$

$$|\Sigma_1| = |\Sigma_2| e^{k_{\pi,d}}. \quad (3.39)$$

Med momentmetoden erstatter man de teoretiske momentene $m_{Z,1}$, $m_{Q,2}$ og $m_{Q,3}$, med de estimerte verdiene $\hat{m}_{Z,1}$, $\hat{m}_{Q,2}$ og $\hat{m}_{Q,3}$. Dette gir estimatorene:

$$\hat{k}_{L,d} = \frac{\left(\hat{m}_{Q,2} - \psi_d^{(1)}(L)\right)^3}{\left(\hat{m}_{Q,3} - \psi_d^{(2)}(L)\right)^2},$$

$$\hat{\pi}_1 = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{4\hat{k}_{L,d} + 1}},$$

$$\hat{k}_{\pi,d} = \sqrt{\frac{\hat{m}_{Q,2} - \psi_d^{(1)}(L)}{\hat{\pi}_1 \hat{\pi}_2}},$$

$$|\hat{\Sigma}_2| = \frac{\hat{m}_{Z,1}}{\hat{\pi}_1 e^{\hat{k}_{\pi,d}} + \hat{\pi}_2},$$

$$|\hat{\Sigma}_1| = \hat{\Sigma}_2 e^{\hat{k}_{\pi,d}}.$$

Akkurat som i metode 1 til 3 er det problemer med at π_1 kan gi to svar. Som i disse metodene bruker jeg en ekstra ligning for å finne ut hvilke estimater som er de rette. I denne metoden bruker jeg moment $m_{Q,1}$:

$$m_{Q,1} = W_1[Q] + B_1[Q] = \pi_1 \ln|\Sigma_1| + \pi_2 \ln|\Sigma_2| + \psi_d^{(0)}(L) - d \cdot \ln(L). \quad (3.40)$$

Metoden regner ut de to estimatene $|\hat{\Sigma}_1|$ og $|\hat{\Sigma}_2|$ på to forskjellige måter og bruker ligning (3.40) for å se hvilken metode som gir best resultat. Dette gjøres ved å sammenligne svarene fra uttrykket (3.41), der kombinasjonen som er nærmest null velges fremfor den andre:

$$\hat{\pi}_1 \ln|\hat{\Sigma}_1| + \hat{\pi}_2 \ln|\hat{\Sigma}_2| + \psi_d^{(0)}(L) - d \cdot \ln(L) - \hat{m}_{Q,1}. \quad (3.41)$$

3.2.2 Metode 6

Denne metoden ligner på metode 4, der Newtons metode ble brukt til å finne estimatet for π_2 numerisk. Også i denne metoden har jeg valgt å begrense antallet dimensjoner i kovariansmatrisa til $d = 2$ og $d = 3$. Momentene som er brukt i denne metoden er $m_{Z,1}$, $m_{Q,1}$ og $m_{Q,2}$ og er gitt henholdsvis i ligning (3.28), (3.40) og (3.29).

For å forenkle ligninga har jeg valgt å definere:

$$k_{Q,1} \equiv m_{Q,1} - \psi_d^{(0)}(L) + d \cdot \ln(L), \quad (3.42)$$

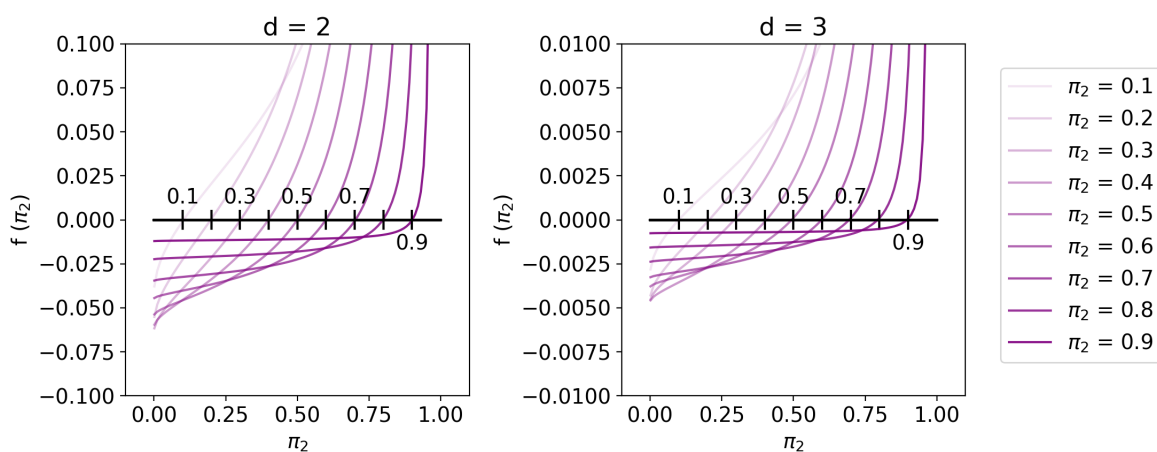
$$k_{Q,2} \equiv \sqrt{m_{Q,2} - \psi_d^{(1)}(L)}. \quad (3.43)$$

Her er $k_{d=2,3}$ definert som i ligning (3.27). Videre finner man forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$, som gjort i metode 5. Dette forholdet er gitt i ligning 3.37. Kombinerer man den med $m_{Q,1}$ (ligning 3.40) kan man finne et uttrykk for $|\Sigma_2|$:

$$|\Sigma_2| = e^{k_{Q,1} - k_{Q,2} \sqrt{\frac{\pi_1}{\pi_2}}}. \quad (3.44)$$

Bruker man de tre ligningene: $\pi_1 + \pi_2 = 1$, (3.28) og (3.44), kan man få et uttrykk som bare avhenger av π_2 , d og L :

$$f(\pi_2, d) = e^{k_{Q,1} - k_{Q,2} \sqrt{\frac{1}{\pi_2} - 1}} \left((1 - \pi_2) e^{k_{Q,2} / \sqrt{\pi_2 - \pi_2^2}} + \pi_2 \right) - m_{Z,1} \cdot \frac{1}{k_{d=2,3}} = 0. \quad (3.45)$$



Figur 3.2: Funksjonen $f(\pi_2)$, der $m_{Z,1}$, $m_{Q,1}$ og $m_{Q,2}$ er estimert ved bruk av 10000 datapunkter. Figuren inneholder de to tilfellene der dimensjonen på matrisene er $d = 2$ og $d = 3$. Figurene viser flere funksjoner der π_2 er variert fra 0,1 til 0,9.

Dimensjonen d er kjent og L antas å være kjent. For å finne π_2 har jeg valgt å bruke Newtons metode, der $m_{Z,1}$, $m_{Q,1}$ og $m_{Q,2}$ estimeres på forhånd. Figur 3.2 viser to tilfellene der dimensjonene på kovariansmatrisa er $d = 2$ og $d = 3$. π_2 er også variert og grafen til $f(\pi_2, d)$ er vist. Figurene ligner på figur 3.1. Jeg har dermed valgt å lage den nye metoden, metode 6, med inspirasjon fra metode 4. Jeg har blant annet valgt å starte i det samme område som metode 4, der standard startpunkt i metode 6 er satt til $\pi_2 = 0,99$.

De seks metodene er implementert i python 3 og er vedlagt oppgaven.

Kapittel 4

Resultat:

Simuleringsstudier med intensitetsdata

I dette kapitlet vil jeg gå gjennom flere simuleringstudier der metode 1 til 4 blir sammenlignet. Data vil bli simulert fra en blandingsmodell, der jeg vil se hvordan de fire metodene oppfører seg i fire tilfeller: når L i blandingsmodellen øker, sannsynligheten til klassene endres, antall datapunkter n øker og når forholdet mellom μ_1 og μ_2 blir mindre. For hver av simuleringstudiene er det blitt sett på forventnings-skjevhet, varians og MSE (se delkapittel 2.7). Siden jeg ikke har teoretiske uttrykk for forventningen til estimatorene $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\mu}_1$ og $\hat{\mu}_2$, har jeg brukt Monte Carlo simuleringer. Jeg har heller ikke klart å finne noe teoretisk uttrykk for varians til de samme estimatorene, så også denne vil bli estimert. I simuleringene har jeg valgt å bruke 1000 estimater når variansen og forventningen blir estimert.

	Metode 1	metode 2	metode 3	metode 4
Bruker :	$m_{X,2}, m_{Y,2}, m_{Y,3}$	$m_{X,1}, m_{Y,2}, m_{Y,3}$	$m_{Y,1}, m_{Y,2}, m_{Y,3}$	$m_{X,1}, m_{Y,1}, m_{Y,2}$
Ekstra ligning ved tvetydighet:	$m_{X,1}$	$m_{Y,1}$	$m_{X,1}$	
Metode:	Momentmetode	Momentmetode	Momentmetode	Momentmetode og Newtons metode

Tabell 4.1: Oppsummering av metode 1 til 4

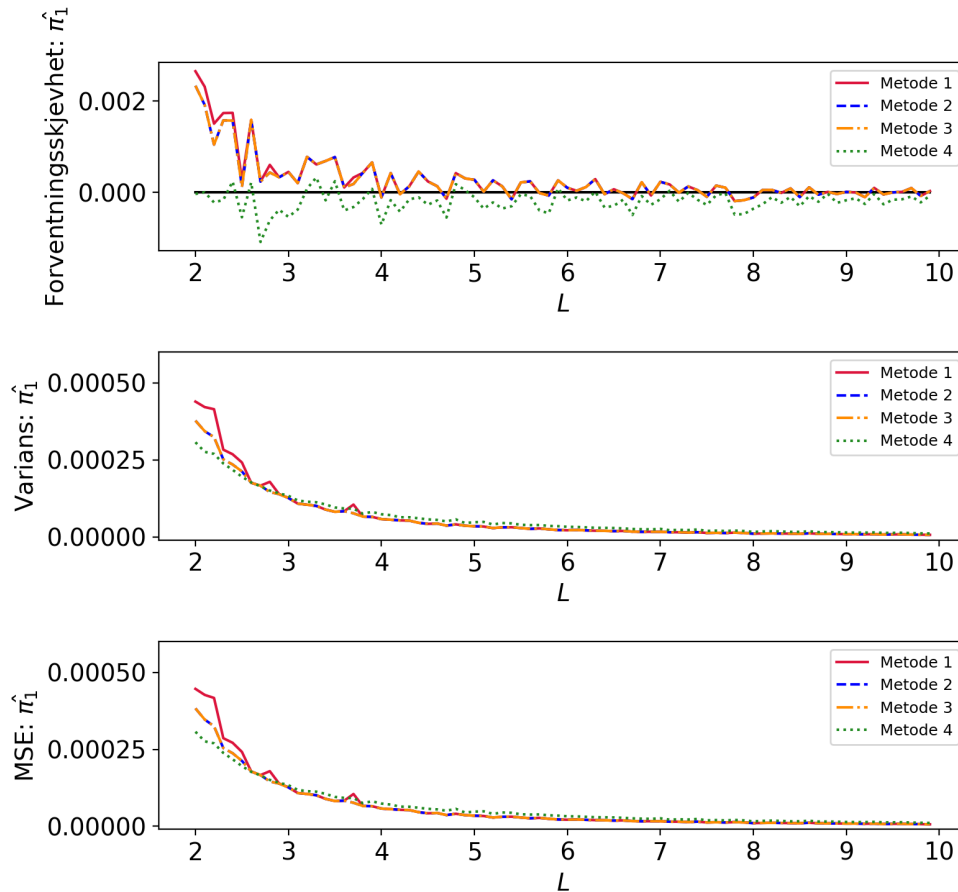
4.1 Varierende L

I denne simuleringstudien har jeg sammenlignet metode 1 til 4 når L er den eneste variabelen som er blitt endret. For hver verdi av L som er sett på, er det beregnet 1000 estimater. Estimatet til variansen og gjennomsnittet til estimatene er funnet fra disse realisasjonene. Variablene som er brukt kan man se i tabell 4.2. μ_1 og μ_2 er valgt ut fra observasjoner beskrevet i kapittel 6.

L	π_1	n	μ_1/μ_2
Varierer fra 2 til 10	0,4	$15^2 = 225$	0,001/ 0,03

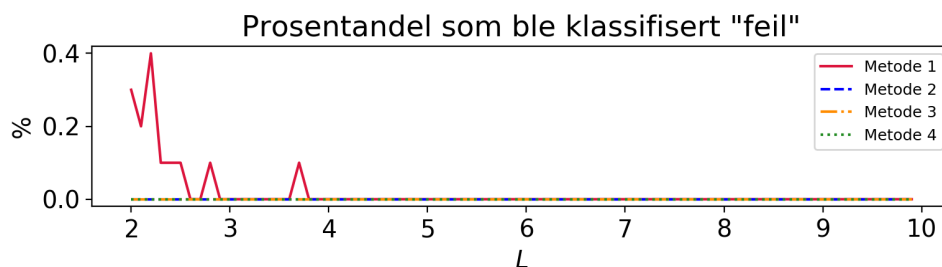
Tabell 4.2: Verdier som er brukt i simuleringstudie med varierende L .

Den øverste figuren i figur 4.1 viser forventningsskjevhet til estimatet $\hat{\pi}_1$ for de fire metodene 1 til 4. I figuren er metodene 1 til 4 henholdsvis rød, blå, gul og grønn. For metode 1 til 3 kan det se ut til at metodene kanskje overestimerer π_1 for lave verdier av L, mens metode 4 ser kanskje ut til å være litt underestimert. Når L øker, ser metodene ut til å gi mindre feil og variansen ser ut til å gå ned. MSE ser ut til å få et større bidrag fra variansen, enn fra forventningsskjevheten.



Figur 4.1: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ ved varierende L

De tre første metode 1 til 3 estimerer $\hat{\pi}_1$ på samme måte, så eventuelle forskjeller her kommer av at estimatene er plassert i feil klasse. Dette vil avhenge av $\hat{\mu}_1$ og $\hat{\mu}_2$. I figur 4.1 kan det se ut til at metode 1 avviker litt fra metode 2 og 3 for lave verdier av L. For å se om dette henger sammen med mulige feilplasseringer, som at μ_1 er koblet med π_2 , har jeg sammenlignet den estimerte verdien $\hat{\pi}_1$ med den sanne π_1 . Er denne avstanden mindre enn avstanden mellom $\hat{\pi}_2$ og π_1 , antas det at klassene er riktig. Dette vil antagelig gi et godt bilde på hvor mange estimerer som er byttet om på, siden det er relativt stor avstand mellom π_1 og π_2 . Figur 4.2 viser hvor mye feil metodene ser ut til å gjøre. I denne figuren ser det ut til at metode 1 er den eneste metoden som gjør feil med variablene som er brukt. Feilene ser ut til å være plassert på de samme verdiene av L der det var avvik mellom metode 1 og metode 2 og 3. Bruker man derimot $\pi_1 = 0,6$ isteden for 0,4, hadde ikke metode 1 problemer med å plassere klassene rett.



Figur 4.2: Figuren viser hvor mye som er gjort feil på grunn av ombytting av klasse 1 og 2

Siden $\pi_2 = 1 - \pi_1$, så har π_2 den samme variansen som π_1 :

$$\text{Var}(\hat{\pi}_2) = \text{Var}(1 - \hat{\pi}_1) = \text{Var}(\hat{\pi}_1).$$

Dette er grunnen til at jeg ikke har gått videre inn på estimatene til π_2 . I tillegg er:

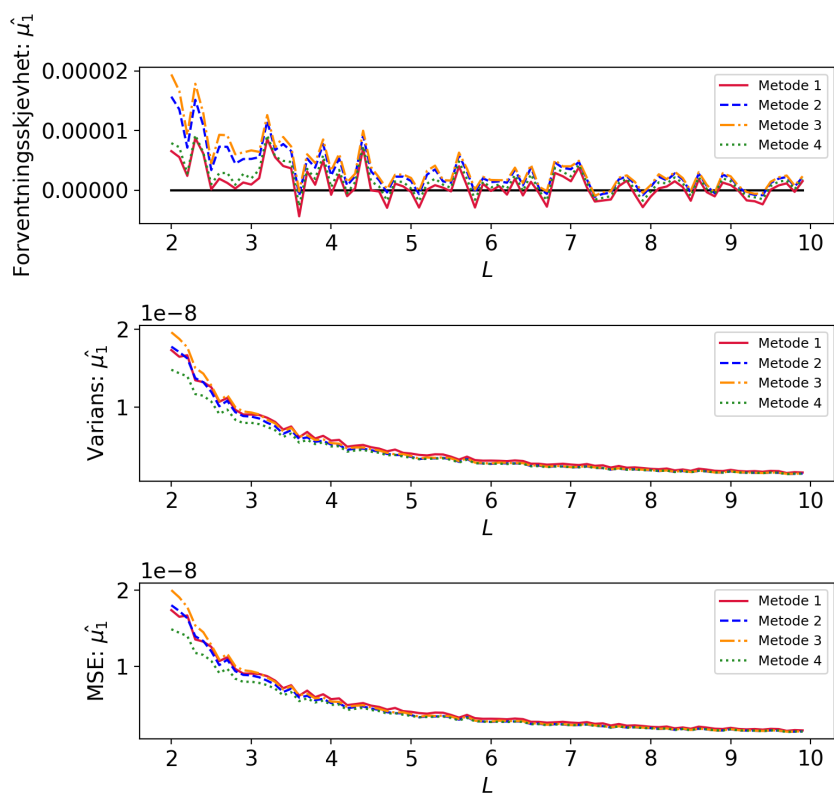
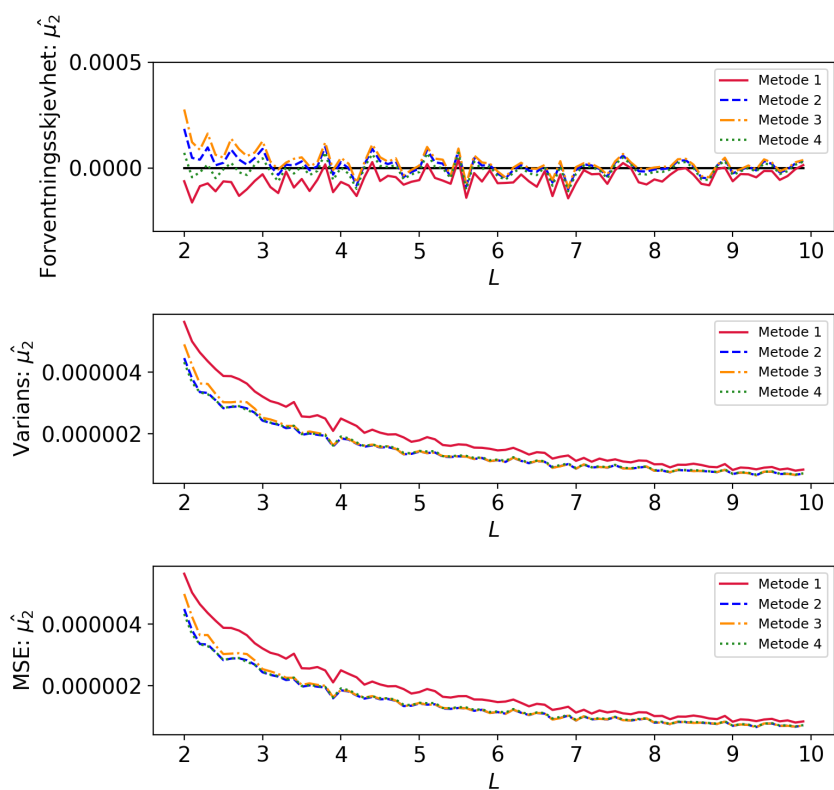
$$E[\hat{\pi}_2] = E[1 - \hat{\pi}_1] = 1 - E[\hat{\pi}_1].$$

Ut fra disse to ligningene kan man både få egenskapene til varians, forventningskjevhet og MSE, der variansen og MSE vil være lik, mens forventningskjevheten vil ha motsatt fortegn.

Verdien L , som er variert i denne simuleringen, har sammenheng med midlingen over intensitetsbildet. Ved å ta gjennomsnittet av et større vindu, vil man få en større verdi L . Dataene vil da få mindre varians, og fordelingen vil ifølge sentralgrenseteoremet gå mot normalfordeling (Walpløe mfl., 2014, kap. 8.3). Det vil kunne føre til at data lettere skilles fra hverandre, forutsatt at man ikke filtrerer klassene sammen. Dette er en tendens som er felles for alle de fire metodene, og kan være grunnen til at metoden ser ut til å gi estimater med mindre varians for høyere verdi av L . I tillegg vil selve blandingsmodellen også få mindre varians, siden variansen innad i klassen minker.

Figur 4.3 viser de samme egenskapene som var undersøkt for estimatene av π_1 , men nå blir det sett på estimatene av μ_1 . For denne parameteren ser det ut til at alle metodene gir ganske like svar, i tillegg ser feilen ut til å være ganske liten. Ut fra forventningskjevheten til $\hat{\mu}_1$, kan det kanskje se ut til at metodene har en tendens til å overestimere $\hat{\mu}_1$, men ved økende verdi av L ser det ut til at denne tendensen forsvinner. De fire estimatorene ser også ut til å bli mer lik ved økende L , i tillegg går variansen ned. MSE, som sier noe hvor god en estimator er, har minimale forskjeller mellom metodene. I tillegg ser det ut til at variansen til $\hat{\mu}_1$ har størst innvirkning på MSE.

For estimatoren $\hat{\mu}_2$ er de tre egenskapene forventningskjevhet, varians og MSE, vist i figur 4.4. Forventningskjevheten ser ut til å være liten og ganske lik for de fire metodene. Variansen til metode 1 ser ut til å være høyere enn for de tre andre metodene, noe som ser ut til å gjelde hele området for L . Endrer man derimot verdien på variabelen $\pi_1 = 0,4$ med for eksempel $0,6$, har man ikke lenger denne tendensen. I dette tilfellet var estimatene mer lik de for metode 2 og 3. I simuleringene med varierende L , kan det se ut til at metode 1 har litt dårligere egenskaper.

Figur 4.3: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_1$ ved varierende LFigur 4.4: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_2$ ved varierende L

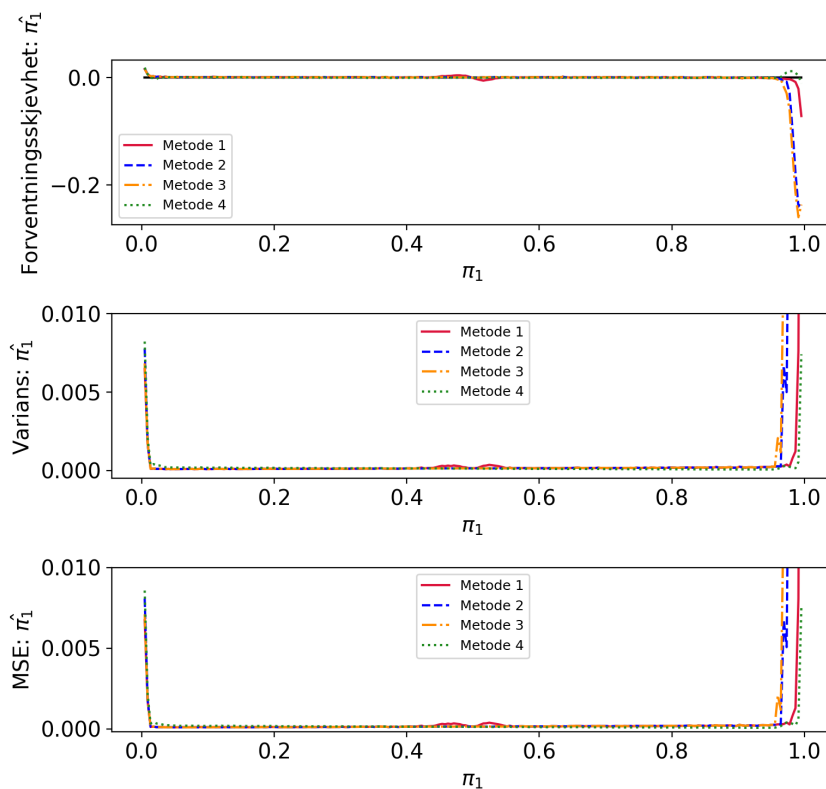
4.2 Varierende π_1

Variablene π_1 og π_2 kan bare ta verdier mellom 0 og 1. Det er da mulig å undersøke hele intervallet og se hvordan de fire metodene fungerer. I denne simuleringstudien blir π_1 variert, mens π_2 følger fra ligningen $\pi_1 + \pi_2 = 1$. Resten av variablene ble holdt konstant. Estimatenes for varians og gjennomsnitt ble funnet ved å estimere 1000 verdier for hver verdi av π_1 som det ble sett på. Tabell 4.3 viser hvilke verdier som ble brukt i denne simuleringstudien. Variablene er nesten de samme som i simuleringen med varierende L (se delkapittel 4.1).

L	π_1	n	μ_1/μ_2
3	Varierer fra 0 til 1	$15^2 = 225$	0,001/ 0,03

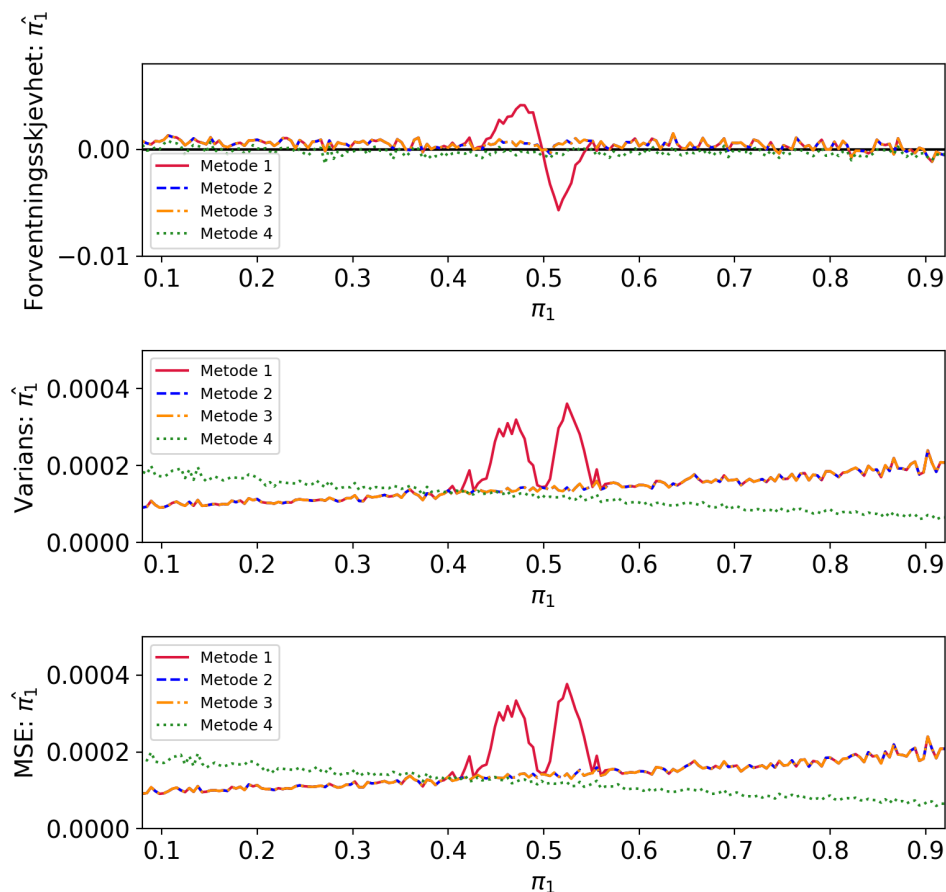
Tabell 4.3: Verdier som ble brukt i simuleringstudie med varierende π_1 .

I figur 4.5 ser det ut til at alle metodene hadde litt problemer når π_1 nærmet seg 0 eller 1. Dette er i grenseland for antagelsen om at dataene er fra en blandingsmodell. Det er derfor ikke så overraskende at problemene er størst her. Nær randpunktene er både forventningsskjevheten og variansen størst. Metode 4 er den metoden som har hatt mest problemer og forbedringspotensial, men i de tre figurene i 4.5 skiller den seg ikke ut og alle metodene ser ut til å gi ganske like estimater. Problemene til metode 4 har bestått i blant annet større varians nær $\pi_1 = 0,8$. Dette problemet så ut til å bli løst ved å sette startverdien nærmere dette punktet. Det kan også oppstå problemer som følge av at Newtons metode konvergerer sakte og dermed stoppet man metoden for fort. Dette førte til større estimert forventningsskjevhet enn hva som er reelt.



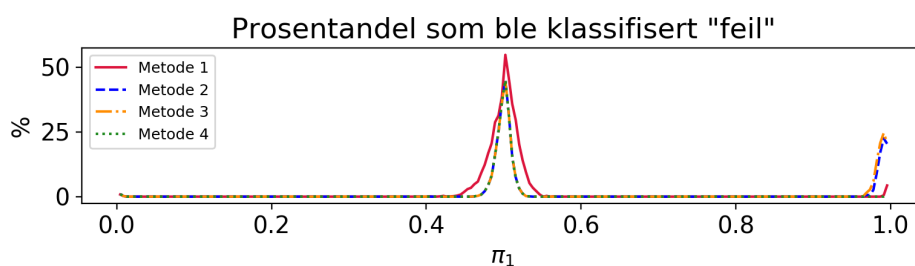
Figur 4.5: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ ved varierende π_1

Går man nærmere inn på intervallet mellom $\pi_1 = 0,1$ og $\pi_1 = 0,9$ kommer forskjellene tydeligere frem. Dette er gjort i figur 4.6. Alle metodene ser ut til å gi svar rundt den sanne verdien, men metode 1 har litt problemer rundt $\pi_1 = 0,5$. Er man under 0,5, ser det ut som om metode 1 overestimerer $\hat{\pi}_1$. Er man derimot over 0,5, ser det ut til at variabelen $\hat{\pi}_1$ blir underestimert. Grunnen til dette er antagelig at metode 1 bytter om på klassene, slik at metoden klassifiserer π_2 til klasse 1. Dette gjør at gjennomsnittet blir høyere for $\pi_1 < 0,5$, på grunn av ligningen $\pi_2 = 1 - \pi_1$. Er man over $\pi_1 = 0,5$, vil man få et lavere gjennomsnitt av samme årsak. Dette mønstret ser bare ut til å gjelde i intervallet $0,4 < \pi_1 < 0,6$. I dette området vil det være litt tvetydig hvilken klasse estimatene vil tilhøre, siden forskjellene på π_1 og π_2 er liten. Det vil heller ikke være like avgjørende, siden feilene er mindre på grunn av likheten mellom π_1 og π_2 . Metode 2 til 4 ser likevel ut til å håndtere dette området bedre. Ser man på variansen til metode 1 i området mellom $\pi_1 = 0,4$ og $0,6$, er dette området med mest varians (om man ser bort fra når π_1 er nær 0 og 1). Alle metodene ser ut til å ha varians som avhenger av π_1 . For metode 4 ser det ut til at variansen går ned med økende π_1 , men for alle de andre metodene ser det ut til at variansen øker. Estimatet for MSE ser ut til å være dominert av variansen. Ut fra den nederste figuren i figur 4.6, kan det se ut til at metode 4 kanskje har litt bedre egenskaper for lave verdier av π_1 . Har man derimot høye verdier av π_1 , ser det ut til at metode 2 og 3 har bedre egenskaper. Metode 1 ser ut til å være litt dårligere på grunn av den høyere variansen i intervallet nær $\pi_1 = 0,5$.



Figur 4.6: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ i intervallet mellom 0,1 og 0,9

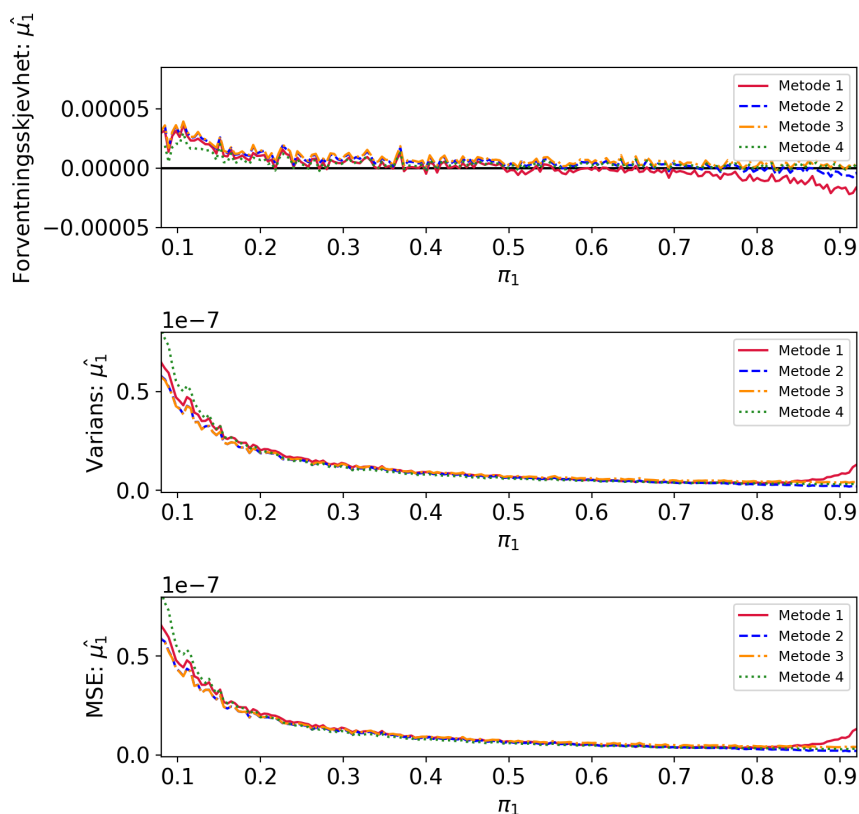
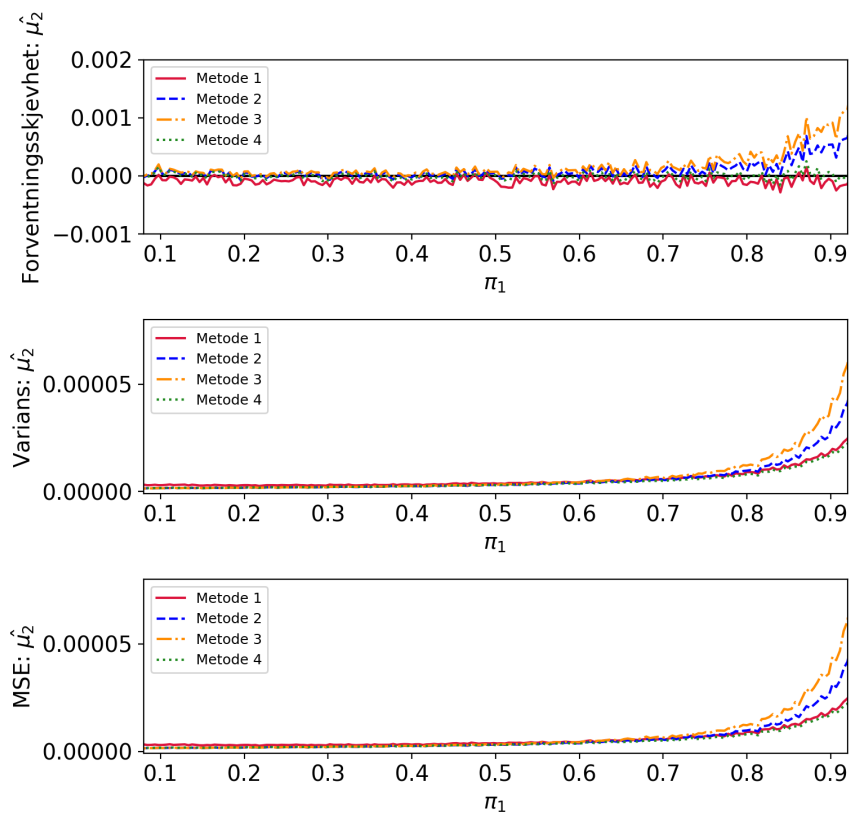
Det er også sett på hvor stor prosentandel som blir plassert i feil klasse. Dette er vist i figur 4.7. Her er det litt feil ved kantene, der metode 2 og 3 ser ut til å ha mest problemer når π_1 er rundt 1. Når sannsynligheten er nær 0,5 er det flest bytter mellom klassene som ser ut til å være feil. Feilene er regnet ut fra hvor mange ganger avstandene er mindre mellom $\hat{\pi}_2$ og π_1 enn mellom $\hat{\pi}_1$ og π_1 . Er noen plassert feil vil det i området rundt $\pi_1 = 0,5$ ikke være like avgjørende, siden verdiene for π_1 og π_2 er så like. Dette kan man også se på figur 4.5 og 4.6, der de estimerte verdiene avviker lite fra den sanne verdien. Det kan likevel se ut som om metode 1 gjør litt mer feil. Dette til tross for at estimatet for $\hat{\pi}_1$ er lik for metode 1 til 3. Dette kan tyde på at metode 2 og 3 er bedre i dette området.



Figur 4.7: Figuren viser hvor mye som er blitt feil på grunn av ombytting av klasse 1 og 2

Videre i simuleringen har jeg valgt å se på egenskapene til $\hat{\mu}_1$ og $\hat{\mu}_2$ i intervallet mellom 0,1 og 0,9. Grunnen til dette er at metodene gir veldig mye mer feil for π_1 nær 0 og 1 siden antagelsen om en blandingsmodell er dårlig. I dette området vil estimatene være mer påvirket av hva metoden gjør når estimatene ikke kan regnes ut, og dette er ikke hovedoppgavene til metodene.

Variansen til $\hat{\mu}_1$ (midterste figur i figur 4.8) for de fire metodene ser ut å være større for lave verdier av π_1 . Når π_1 er lav vil dette si at fordelingen er dominert av klasse 2. Den samme tendensen ser man i figur 4.9, der man ser på de estimerte verdiene $\hat{\mu}_2$. Her er det derimot større varians for høye verdier av π_1 . Dette vil si at klasse 1 dominerer, og π_2 er lite. Grunnen til dette kan komme av at det er mindre data å estimere μ_1 og μ_2 fra når den motsatte klassen er størst. Antall data er konstant $n = 225$ for hele simuleringen, så når for eks. $\pi_1 = 0,2$, vil klasse 2 bestå av 45 datapunkter. Er π_1 derimot $\pi_1 = 0,8$, vil klasse 2 bestå av 180 datapunkter. Ellers ser det ut til at metoden gir estimater med liten forventningsskjevhet for både $\hat{\mu}_1$ og for $\hat{\mu}_2$. Det var heller ikke noen metoder som ga betydelig mer feil enn de andre metodene.

Figur 4.8: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_1$ ved varierende π_1 Figur 4.9: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_2$ ved varierende π_1

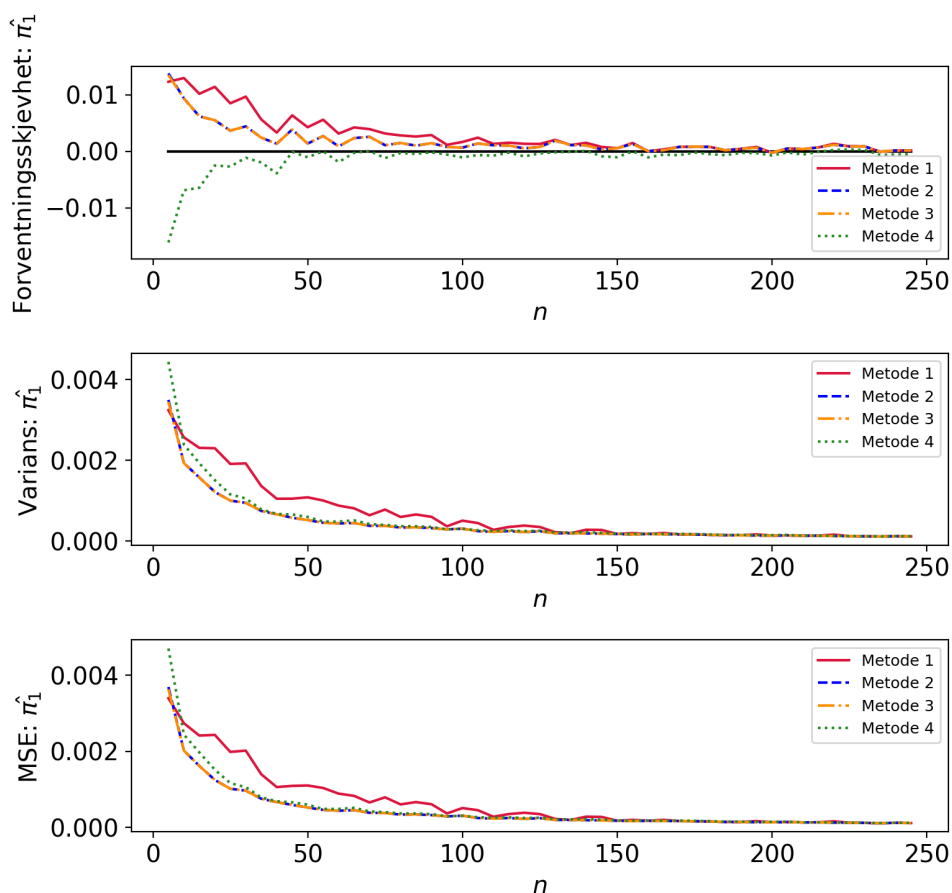
4.3 Varierende utvalg n

Det er mulig å variere utvalget n , som blir brukt til å estimere momentene. Dette antallet kan påvirke hvor godt modellen fungerer. Siden $\pi_1 = 0,4 = 2/5$, har jeg valgt å bruke verdiene $n = 5, 10, 15, \dots$. Dette vil sørge for at π_1 holder seg konstant, siden antallet simulerte verdier fra klasse 1 ($\pi_1 \cdot n$) og klasse 2 ($\pi_2 \cdot n$) vil være heltall. Tabell 4.4 inneholder de forskjellige variablene som ble brukt i denne simuleringstudien.

L	π_1	n	μ_1/μ_2
3	0,4	Varierer fra 5 til 250	0,001/ 0,03

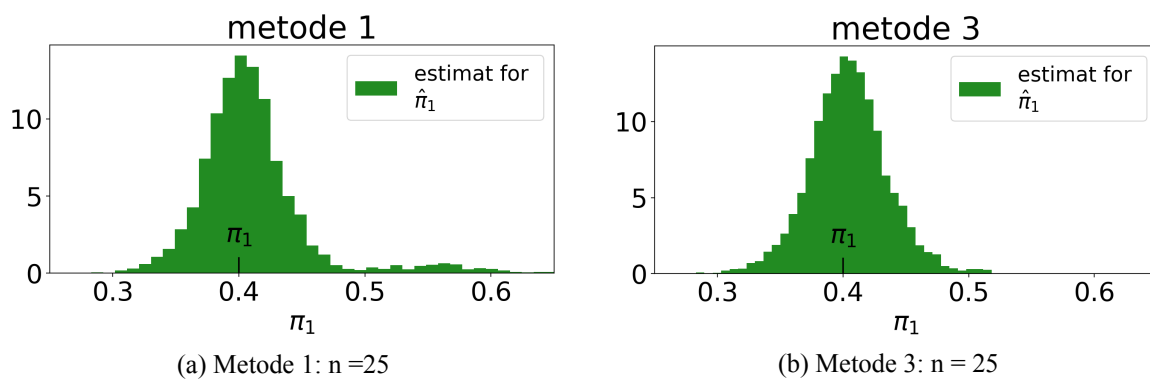
Tabell 4.4: Verdier som er brukt i simuleringstudien med varierende utvalg n .

I figur 4.10 er utvalget n økt fra 5 til 225, noe som har resultert i gradvis mindre forventningsskjevhet og varians. Alle metodene gjør mest feil for små utvalg n , og de forskjellige estimatene er også mest forskjellig i dette området. Metode 1 til 3 ser kanskje ut til å gi litt for høye estimater av $\hat{\pi}_1$, mens metode 4 gir litt lave. Når utvalget n øker, ser det ut til at alle metodene gir estimater rundt den sanne verdien π_1 , noe som kan tyde på at estimatet for π_1 er konsistent (Se delkapittel 2.7). For MSE ser det ut til at det er variansen som dominerer den kvadratiske avstanden mellom π_1 og $\hat{\pi}_1$.



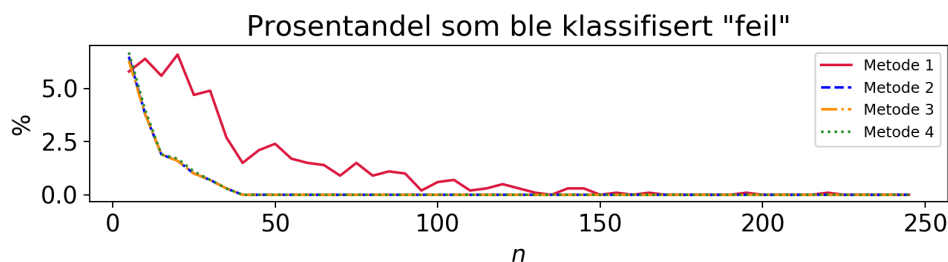
Figur 4.10: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ ved varierende n

Metode 1, som estimerer $\hat{\pi}_1$ på samme måte som metode 2 og 3, har en god del høyere varians. Det kan også se ut som om metoden er mer forventningsskjev. Jeg har dermed sett nøyere på estimatene til $\hat{\pi}_1$ med metode 1 og metode 3. For disse estimatene er utvalget n holdt konstant på $n = 25$. I figuren 4.11a, kan det se ut til at estimatene fra metode 1 danner to grupper. Den lille gruppen rundt 0,55 er antagelig estimerer av $\hat{\pi}_2$ som er plassert i klasse 1. Ser man på figur 4.11b, har ikke det samme skjedd her. Grunnen til dette er ikke måten π_1 er estimert, siden de er estimert på samme måte. Den eneste forskjellen på disse to metodene er at man har brukt et moment med lavere orden ved estimering av μ_1 og μ_2 i metode 3. Siden estimatene fra metode 1 er mer spredd på grunn av feilplassering, vil også variansen til denne metoden bli høyere.



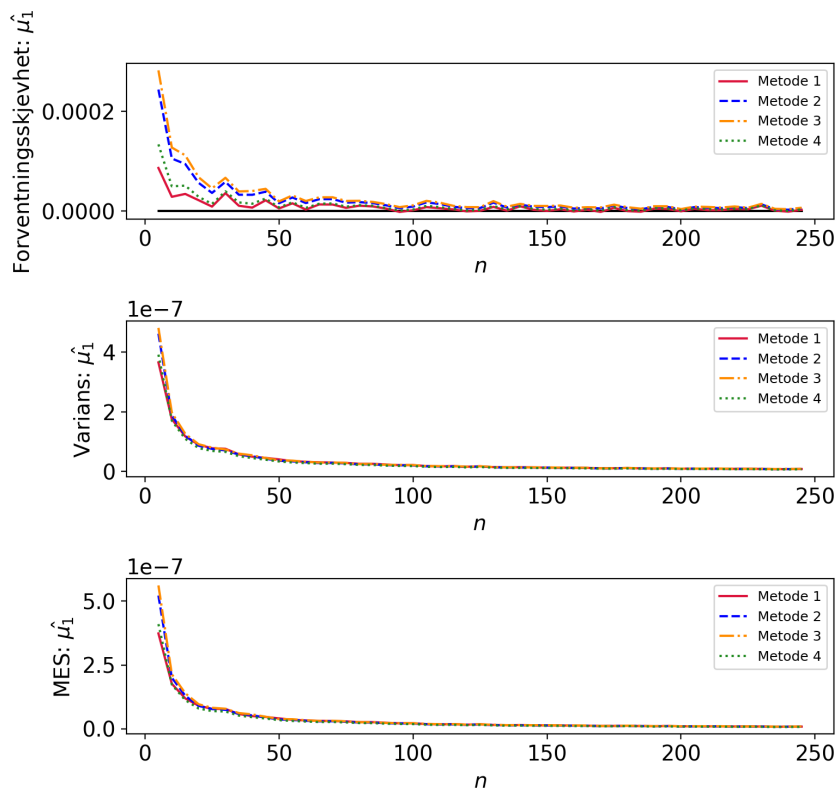
Figur 4.11: Sammenligning av metode 1 og 3, for å undersøke hvorfor metode 1 ser ut til å gi dårligere resultater til tross for at π_1 blir estimert på samme måte. For begge figurene er alle variablene holdt konstant, og det er til sammen simulert 5000 datasett som π_1 er estimert fra.

Det er også mulig å se på hvor mange ganger metodene har byttet om på estimatene for π_1 og π_2 . Dette er vist i figur 4.12. Her gjør alle metodene litt mer feil ved små utvalg, men metode 1 ser ut til å gjøre mest av disse feilene. For metode 2 til 4 ser det ikke ut til at denne typen feil skjer etter $n = 50$ med de variablene som gitt i tabell 4.4. Metode 1 derimot ser ut til å ha litt feil også for verdier over $n = 100$.

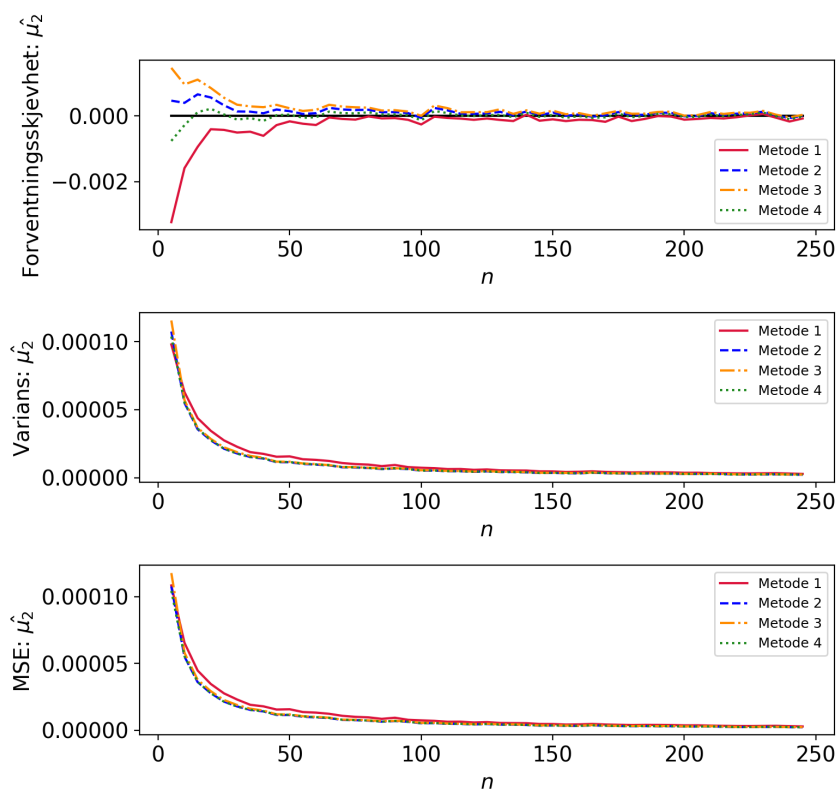


Figur 4.12: Figuren viser hvor mye som er gjort feil på grunn av ombytting av klasse 1 og 2

Egenskapene til estimatorene $\hat{\mu}_1$ og $\hat{\mu}_2$ kan også være av interesse. Disse estimatene er vist i henholdsvis figur 4.13 og 4.14. I figurene ser det ut til at forventningsskjevheten og variansen går ned for både $\hat{\mu}_1$ og $\hat{\mu}_2$ når utvalget n øker. Dette kan tyde på at også disse to estimatorene er konsistente. Ser man på forventningsskjevheten i figur 4.13, ser det ut til at metode 1 og 4 gir minimalt mindre feil enn metode 2 og 3 for lave utvalg av n . Variansen, vist i den nederste figuren i figur 4.13, ser ut til å være ganske lik for alle de fire metodene. Dette gjelder også variansen til $\hat{\mu}_2$ i figur 4.14. For estimatene til $\hat{\mu}_2$, ser det ut til at metode 1 underestimerer μ_2 for lave utvalg av n . Metode 2 og 3 ser derimot ut til å overestimere μ_2 .



Figur 4.13: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_1$ ved varierende n



Figur 4.14: Egenskapene skjevhet, varians og MSE til $\hat{\mu}_2$ ved varierende n

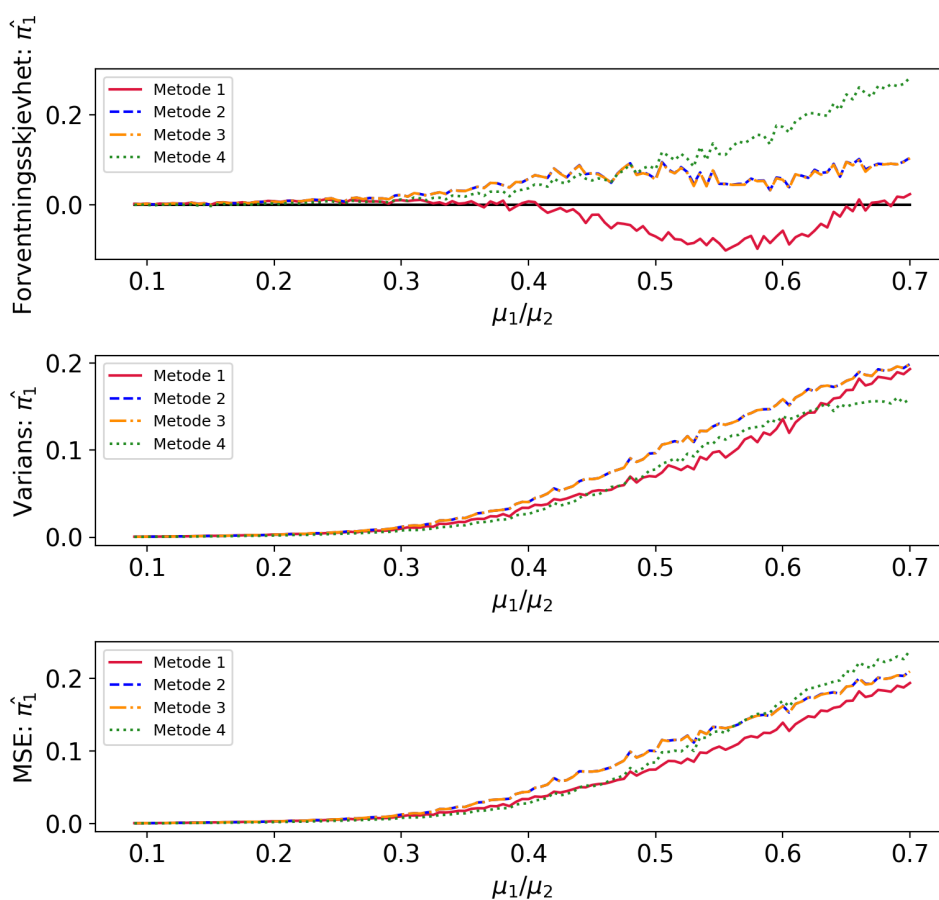
4.4 Varierende forhold μ_1/μ_2

For å kunne estimere π_1 og π_2 bør ikke μ_1 og μ_2 være for like. Hvis de likevel er det, vil man ikke kunne skille mellom fordelingene og man kan dermed ikke skille datasettet i to klasser. For å undersøke hva som skjer når μ_1 går mot μ_2 , kan man se på når forholdet nærmer seg 1. Tabell 4.5 viser hvilke verdier variablene er satt til i denne simuleringstudien.

L	π_1	n	μ_1/μ_2
3	0,4	$15^2 = 225$	μ_1 varierer fra 0,001 til 0,007 μ_2 er satt til 0.01

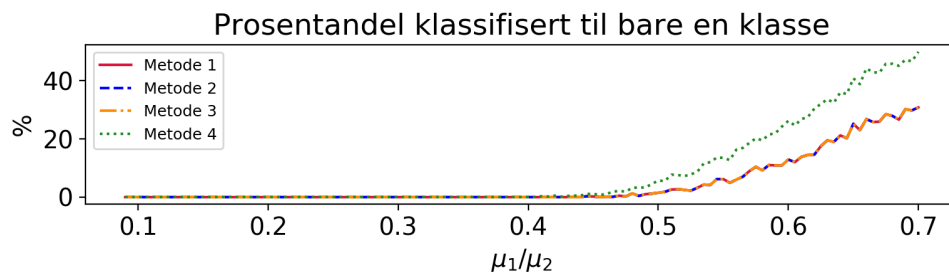
Tabell 4.5: Verdier som ble brukt i simuleringstudie med varierende μ_1/μ_2 .

I figur 4.15 ser det ikke ut til at det skjer så mye før forholdet mellom μ_1 og μ_2 er over 0,3. Etter dette begynner de forskjellige metodene å gi mer feil og ingen av metodene ser ut til å håndtere situasjonen godt. Når forholdet øker, øker også forventningsskjevheten til estimatet $\hat{\pi}_1$. I tillegg går variansen opp. Ut fra figuren for MSE, ser alle estimatene ut til å gi omtrent like dårlige estimater, det vil si at den midla kvadratiske feilen mellom sann og estimert verdi er omtrent like stor.



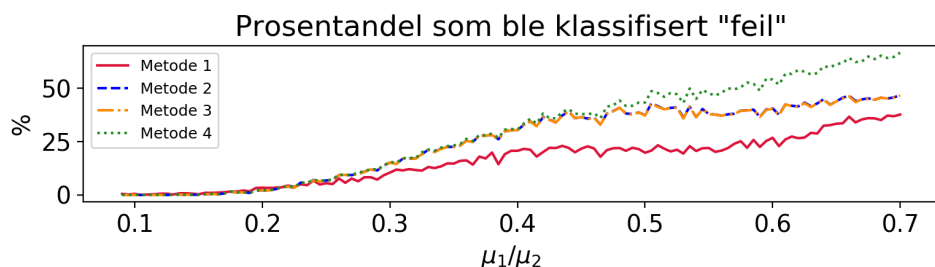
Figur 4.15: Egenskapene skjevhet, varians og MSE til π når forholdet mellom μ_1 og μ_2 endrer seg.

Det er to parametre i hver av klassene, og den ene parameteren L er allerede antatt å være lik. Bli de siste parameterne μ_1 og μ_2 også veldig like, er det ikke mulig å skille mellom de to klassene. Figur 4.16 viser når metodene begynner å klassifisere til bare én klasse. Det ser ut til at dette skjer når forholdet μ_1/μ_2 , er over 0,5. Metode 4 ser også ut til å gjøre dette i større grad enn metode 1 til 3.



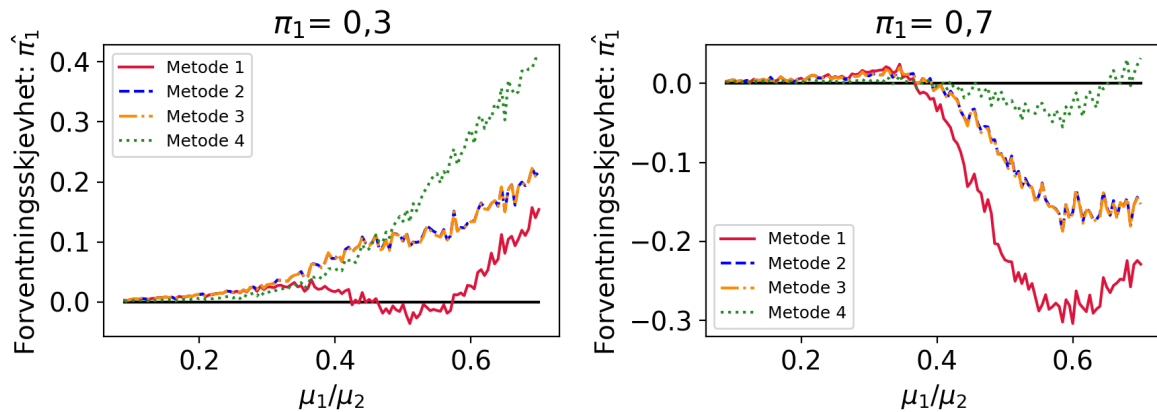
Figur 4.16: Figuren viser hvor stor prosentandel som er klassifisert til bare en klasse når forholdet mellom μ_1 og μ_2 endrer seg.

Det er også mulig å se på hvor ofte klassene er blitt byttet om, dette er sett på i figur 4.17. Denne figuren gir den samme konklusjon som over, det vil si at metodene ikke fungerer etter at forholdet mellom $\hat{\mu}_1$ og $\hat{\mu}_2$ er over $\mu_1/\mu_2 = 0,4$. Ser man på prosentandelen som ble klassifisert feil, kan det se ut til at det begynner å bli mye feil allerede ved $\mu_1/\mu_2 = 0,3$. Når forholdet blir høyere, stiger andelen som er feilplassert opp til rundt 40%.

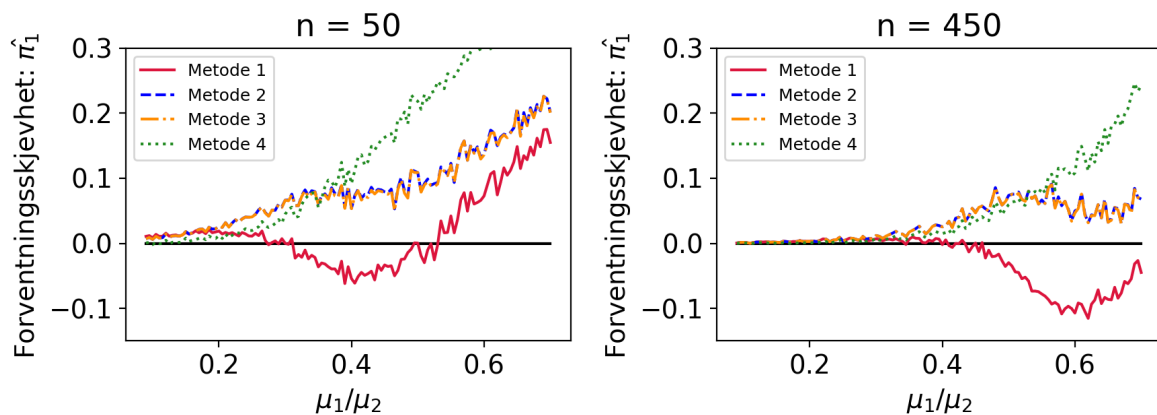


Figur 4.17: Figuren viser hvor mye som er gjort feil på grunn av ombytting av klasse 1 og 2

Jeg har også sett på andre simuleringer, der π_1 , n , μ_1 og μ_2 har vært forskjellige. Blant annet er det sett på når μ_1 og μ_2 var 100 ganger høyere enn i tabell 4.5. Også i dette tilfellet startet metodene å gjøre feil når forholdet mellom μ_1 og μ_2 var over 0,3. Figur 4.18 viser to simuleringer der π_1 er forskjellig. Figuren til venstre bruker $\pi_1 = 0,3$ og den til høyre bruker $\pi_1 = 0,7$. Formålet med simuleringene var å se hvordan π_1 påvirker estimatene. I disse to figurene ser det ut til at metodene begynner å gjøre feil ved samme forhold, noe som kan tyde på at π_1 ikke påvirker dette. Derimot ser det ut til at π_1 påvirker hvor mye feil som blir gjort når forholdet nærmer seg én. For området der μ_1/μ_2 er større enn 0,3 kan det se ut til ingen av metodene egentlig gir gode resultater for de variablene som sett på. Jeg har da sett på om dette kan avhenge av antall punkter som er brukt i estimeringene. Dette er vist i figur 4.19, der figuren til venstre bruker utvalg med $n = 50$ og den til høyre bruker $n = 450$. For simuleringene som brukte $n = 450$, kan det se ut til feilene begynner å bli stor rundt $\pi_1/\pi_2 = 0,4$. For den andre simuleringene der $n = 50$, ser forholdet ut til å ligge rundt 0,2. Det kan da se ut som om at dette avhenger av n . Når metodene skal brukes er dette effekter som man burde være klar over, siden dette kan påvirke hvor gode resultater man får.



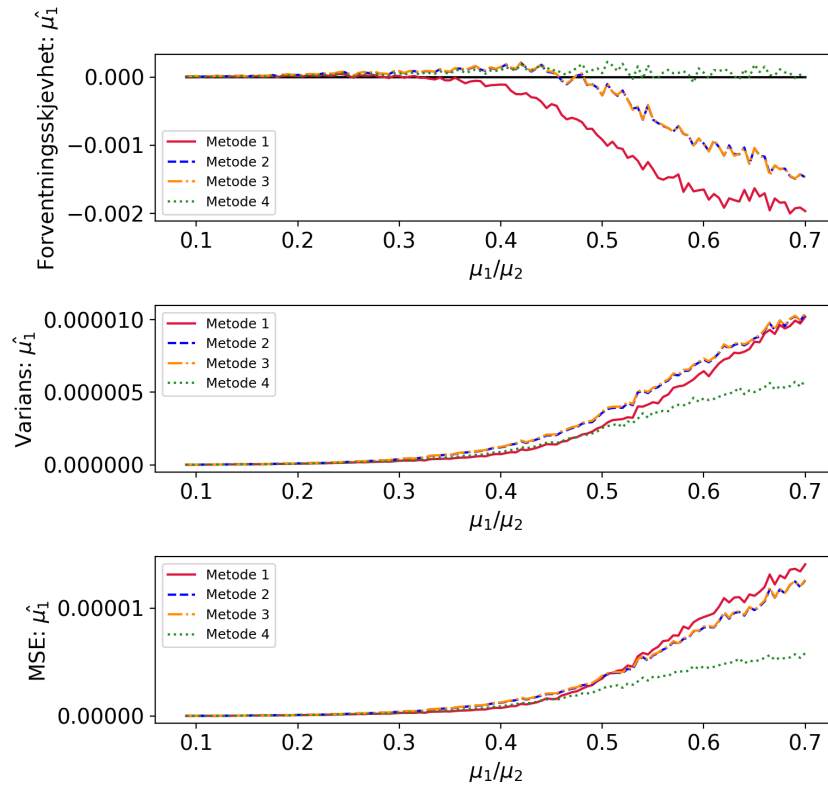
Figur 4.18: Ser hvordan metodene varierer for to tilfeller der π_1 endres til 0,3 og 0,7. Resten av variablene er lik som i tabell 4.5



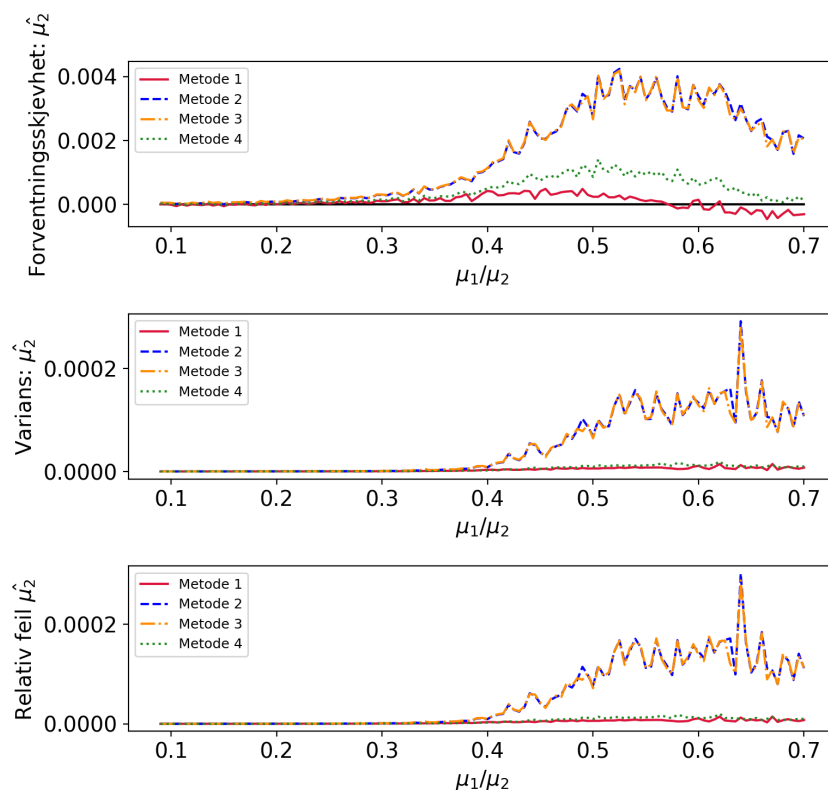
Figur 4.19: Ser hvordan metodene varierer for to tilfeller der størrelsen n til datasettene endres til 50 og 450. Resten av variablene er lik som i tabell 4.5

Når μ_1 og μ_2 begynner å bli like, blir det vanskeligere å skille mellom disse. Ser man på figur 4.20 og 4.21 øker variansen på estimatene til μ_1 og μ_2 . Dette gjør det nok enda vanskeligere å skille mellom klassene, og kan antagelig også påvirke den relative feilen i figur 4.15. I denne simuleringen er variansen til $\hat{\mu}_2$ relativt lav for metode 1 og 4 i forhold til metode 2 og 3. Den relative feilen er heller ikke så høy for disse to metodene. Ser man derimot på estimatene til μ_1 er det bare metode 4 som har gitt litt lavere varians og relativ feil. For metode 1 til 3 ser det ut til at $\hat{\mu}_1$ blir underestimert når forholdet mellom μ_1 og μ_2 er større enn 0,4. Etter $\mu_1/\mu_2 = 0,4$ øker i tillegg den relative feilen ganske fort og variansen går opp.

Simuleringsstudien viser at forskjellen mellom μ_1 og μ_2 ikke burde være for liten. Har man variabler som i tabell 4.5, burde man helst ha $\mu_1/\mu_2 < 0,3$. Har man mindre verdier som for eksempel $n = 50$, så det ut til at forholdet burde være enda mindre, og helst under $\mu_1/\mu_2 = 0,2$.



Figur 4.20: Egenskapene skjevhet, varians og MSE til μ_1 når forholdet mellom μ_1 og μ_2 endrer seg



Figur 4.21: Egenskapene skjevhet, varians og MSE til μ_2 når forholdet mellom μ_1 og μ_2 endrer seg

Kapittel 5

Resultat:

Simuleringsstudier med kovariansmatrise-data

I dette kapitlet vil jeg gå gjennom flere simuleringstudier der metode 5 og 6 blir sammenlignet. Dette er å gjort på samme måte som for metode 1 til 4. Den største forskjellen er at kovariansmatrisen har flere variabler. Dataene som skal testes i kapittel 6 er fra TerraSAR-X med dimensjon $d = 2$. Jeg har dermed valgt å begrense meg til kovariansmatriser med denne dimensjonen. Metodene er også testet på $d = 3$, men de er ikke sammenlignet. Til sammen vil jeg gjøre 5 simuleringer der jeg varierer: L , π_1 , n , forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$ og korrelasjon mellom kanalene i klasse 1 og klasse 2. I det todimensjonale tilfellet, der jeg har spredningskoeffisientene S_{hh} og S_{vv} , vil dette tilsvare å variere korrelasjonen til de kopolariserte målingene $E[S_{hh} \cdot E_{vv}^*]$ og $E[S_{vv} \cdot E_{hh}^*]$ (se ligning 2.4). Hvordan dette blir gjort går jeg videre inn på i delkapittel 5.5.

	metode 5	metode 6
Bruker :	$m_{Z,1}, m_{Q,2}$ og $m_{Q,3}$	$m_{Z,1}, m_{Q,1}$ og $m_{Q,2}$
Ekstra ligning ved tvetydighet:	$m_{Q,1}$	
Metode:	Momentmetoden	Momentmetoden og Newtons metode

Tabell 5.1: Oppsummering av metode 5 og 6

5.1 Varierende L

I denne simuleringstudien har jeg sammenlignet metode 5 og 6 når L er den eneste variabelen som ble variert. L vil i denne simuleringen være et heltall på grunn av måten jeg simulerer matriser trukket fra en kompleks Wishartfordeling. For hver L ble det simulert 1000 estimater for π_1 , μ_1 og μ_2 . Disse verdiene brukes videre til å regne ut forventningskjevhet og varians. Tabell 5.2 inneholder de gitte variablene som ble brukt i simuleringstudien. De to kovariansmatrisene som er brukt vises i ligning (5.1):

$$\Sigma_1 = \begin{pmatrix} 0.001 & 6 \cdot 10^{-7} e^{2\pi j} \\ 6 \cdot 10^{-7} e^{-2\pi j} & 0.003 \end{pmatrix}, \quad (5.1)$$

$$\Sigma_2 = \begin{pmatrix} 0.04 & 0.00023 e^{2\pi j} \\ 0.00023 e^{-2\pi j} & 0.03 \end{pmatrix}.$$

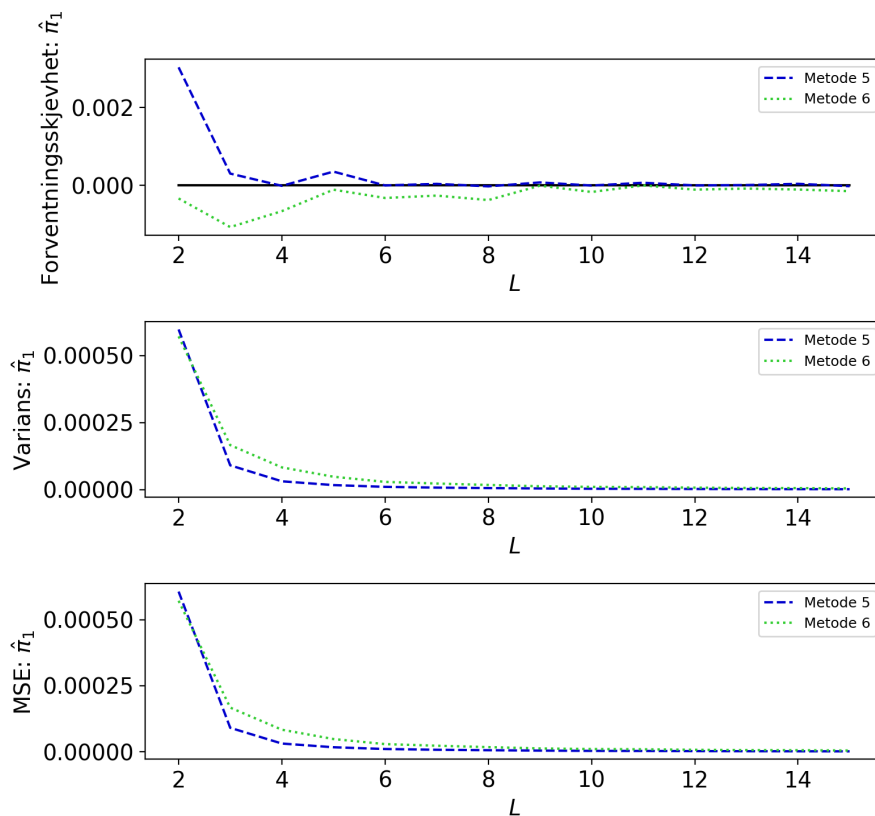
L	π_1	n	$ \Sigma_1 / \Sigma_2 $
Varierer fra 2 til 14	0,4	$15^2 = 225$	$\approx 0,0025$

Tabell 5.2: Verdier som er brukt i simuleringstudie med varierende L.

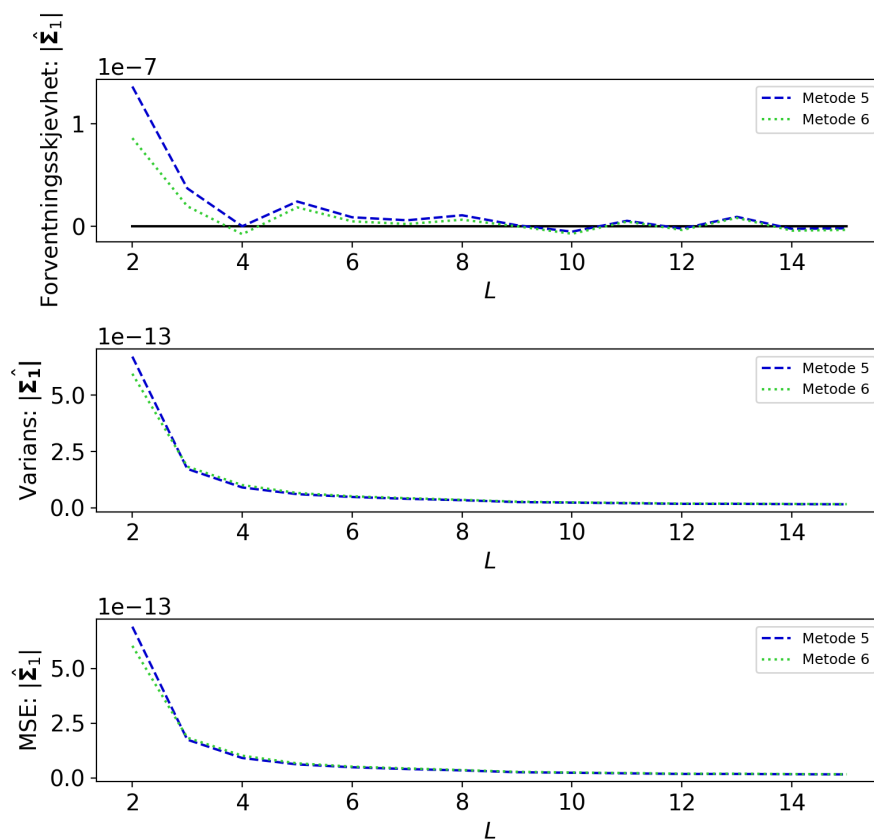
Figur 5.1 viser egenskapene til metode 5 og 6 når L er variert. Ser man på den øverste figuren ser det ut til at begge estimatorene er litt forventningsskjevne når L er lav, der metode 5 ser ut til å overestimere, og metode 6 ser ut til å underestimere. Allerede ved $L = 4$ ser det ut til at det er mindre feil. Variansen til metodene ser også ut til å være størst for lave verdier av L, der variansen avtar mye i starten. MSE, som ser på kvadratisk avstand mellom π_1 og $\hat{\pi}_1$, ser ut til å være dominert av variansen. Sammenlignes resultatene med den samme simuleringstudien gjort for intensiteten (se delkapittel 4.1), ser det ut til at metodene oppfører seg ganske likt. Metode 2, som bruker samme orden på momentene som metode 5, har også i denne simuleringstudien overestimert $\hat{\pi}_1$ for lave verdier av L. Det samme kan man se for metode 4, som underestimerer $\hat{\pi}_1$ (se figur 4.1). Brukes bare én dimensjon på matrisa, reduseres metode 5 til metode 2. Det samme gjelder også for metode 4 og 6.

For det flerdimensjonale tilfellet brukes $|\Sigma_1|$ og $|\Sigma_2|$ for å finne ut hvilken klasse som ses på. I det endimensjonale tilfellet vil $|\Sigma_1|$ og $|\Sigma_2|$ være lik estimatene til henholdsvis μ_1 og μ_2 . Det er også mulig å se på hvor mye feil metodene gjør, ved at sannsynligheten π_1 blir koblet til kovariansmatrisa $|\Sigma_2|$ og motsatt. Dette er gjort på samme måte som i kapittel 4, der $\hat{\pi}_1$ ble sammenlignet med både π_1 og π_2 . Er det gjort riktig vil avstanden mellom $\hat{\pi}_1$ og π_1 være minst. Under denne simuleringen forekom ikke slike bytter.

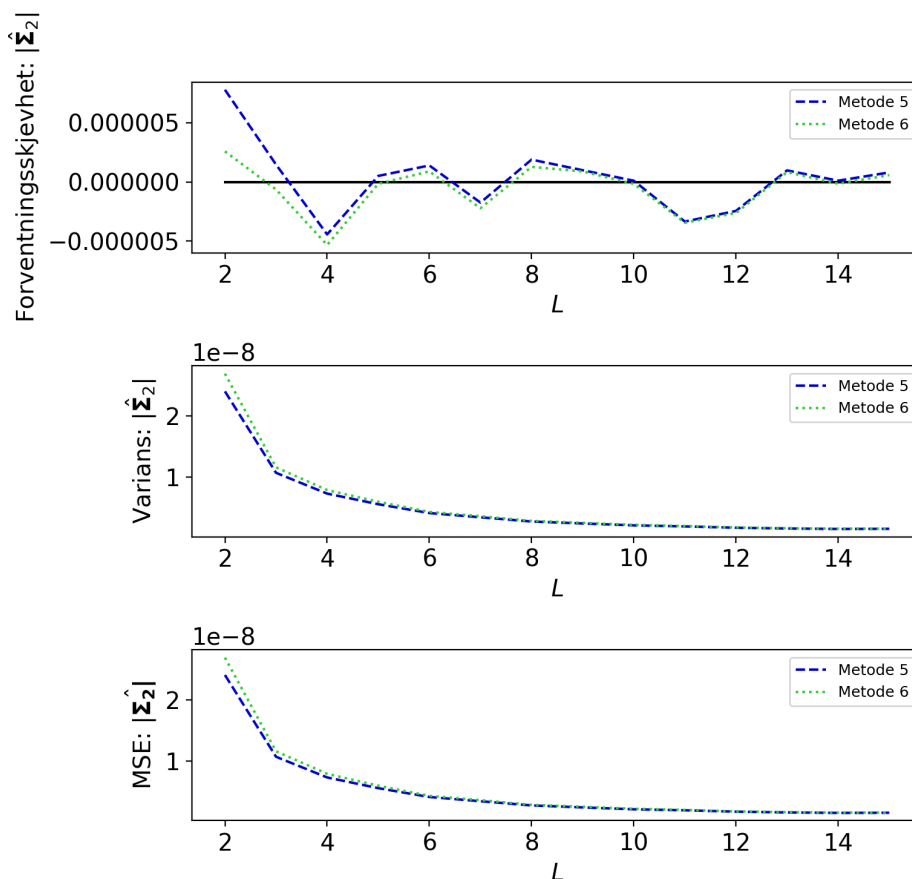
Estimatene til $|\Sigma_1|$ vises i figur 5.2. I denne figuren ser det ut til at estimatene blir litt bedre når L øker. Spesielt frem til $L = 4$ ser det ut til at estimatene forbedres. Her får estimatene både mindre forventningsskjevhet og varians. Variansen oppfører seg ganske likt som variansen til $\hat{\pi}_1$ (se figur 5.1), der variansen avtok raskt frem til $L = 4$. Etter $L = 4$ endrer variansen seg lite. Både metode 5 og 6 overestimerer $|\Sigma_1|$ frem til L er rundt 6. Estimatene til $|\Sigma_2|$ i figur 5.3 ser ut til å endre seg lite ved økende L. Ut fra forventningsskjevheten i figur 5.3, er det vanskelig å si at estimatene ser ut til å bli bedre med økende L. Ser man på variansen derimot, ser det ut til at $|\hat{\Sigma}_2|$ avtar med økende L. MSE, som her er dominert av variansen, tyder på at estimatene for $|\hat{\Sigma}_2|$ blir noe bedre ved økende L.



Figur 5.1: Egenskapene skjevhet, varians og MSE til \hat{r}_1 ved varierende L



Figur 5.2: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_1|$ ved varierende L



Figur 5.3: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_2|$ ved varierende L

5.2 Varierende π_1

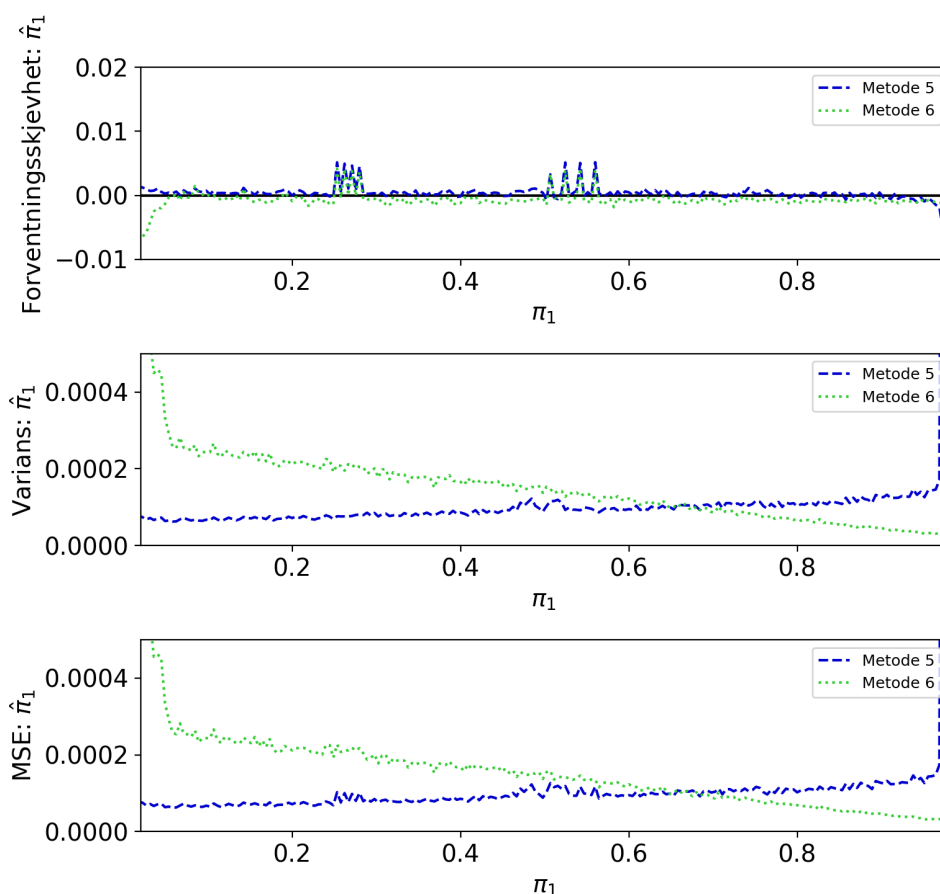
I denne simuleringstudien er π_1 variert, mens π_2 følger ligningen $\pi_2 = 1 - \pi_1$. Alle de andre variablene er holdt konstant. For hver verdi av π_1 ble det gjort 1000 estimater for π_1 , π_2 , $|\Sigma_1|$ og $|\Sigma_2|$. For disse estimatene ble gjennomsnittet og variansen regnet ut. Tabell 5.3 inneholder de forskjellige variablene som er brukt, i tillegg vil de samme kovariansmatrisene som i simuleringstudien med varierende L bli brukt i denne simuleringstudien (Se ligning 5.1).

L	π_1	n	$ \Sigma_1 / \Sigma_2 $
3	Variere fra 0 til 1	$15^2 = 225$	$\approx 0,0025$

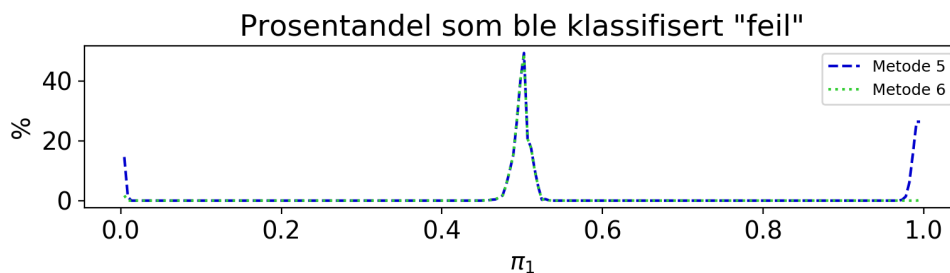
Tabell 5.3: Verdier som ble brukt i simuleringstudie med varierende π_1 .

Som i simuleringstudien med varierende π_1 for intensitetsdata, er det mest avvik nær endene til det gyldige intervallet til π_1 . Akkurat som i metode 1 til 4, så antar også metode 5 og 6 en blandingsmodell. For π_1 nær 0 og 1 er man i grenseland for denne antagelsen, og det er ikke overraskende at også metode 5 og 6 har problemer her. Jeg har valgt å ikke fokusere på disse to områdene, men heller når begge klassene har en viss størrelse. Resultatene vises i figur 5.4.

Ut fra figuren for forventningsjektivhet, ser det ut til at metodene 5 og 6 estimerer verdier for $\hat{\pi}_1$ som ligger rundt de sanne verdiene til π_1 . I denne figuren er det også to områder som skiller seg ut for begge metodene. Dette er nær $\pi_1 = 0,3$ og for π_1 litt over $0,5$. Siden dette ser ut til å gjelde begge metodene, er det antagelig ikke på grunn av den numeriske estimerte verdien for $\hat{\pi}_1$, og feilen denne metoden gjør. Disse feilene ville bare vært tilfelle for metode 6, men feilen ser ut til også å gjelde metode 5. Avviket ved $\pi_1 = 0,5$ kan skyldes bytting av klasser. Denne forklaringen er nok lite sannsynlig for området rundt $\pi_1 = 0,3$, siden klassene er lettere å skille i dette området. Figur 5.5 viser feilene som er gjort ved å klassifisere den estimerte verdien av π_1 til klasse 2 og omvendt. Det er vanskelig å trekke noen slutninger rundt $\pi_1 = 0,3$, men det ser ikke ut til at metodene har gjort feil rundt $\pi_1 = 0,3$. Feilen i dette området skyldes antagelig ikke klassebytter. Rundt $\pi_1 = 0,5$ ser det derimot ut til at det er feil på grunn av klassebytter, men det ser ut til at feilen i figur 5.4 er på høyre side av $\pi_1 = 0,5$. Variansen ser ut til å oppføre seg ganske likt som i simuleringstudien som så på intensitet, der variansen så ut til å variere med π_1 . Metode 6 er den som har dårligst egenskaper for lave verdier av π_1 , mens for høye verdier av π_1 , så det ut til at metode 5 hadde dårligere variansegenskaper. Også i denne simuleringstudien så det ut til at MSE var dominert av variansen.

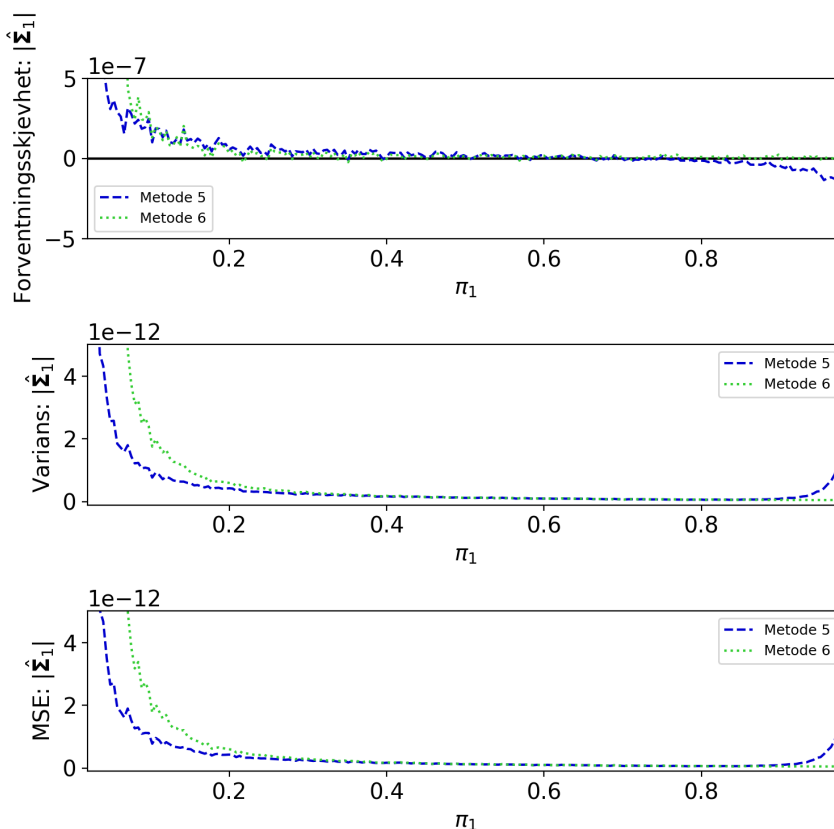


Figur 5.4: Ser på egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ i intervallet mellom 0,1 og 0,9

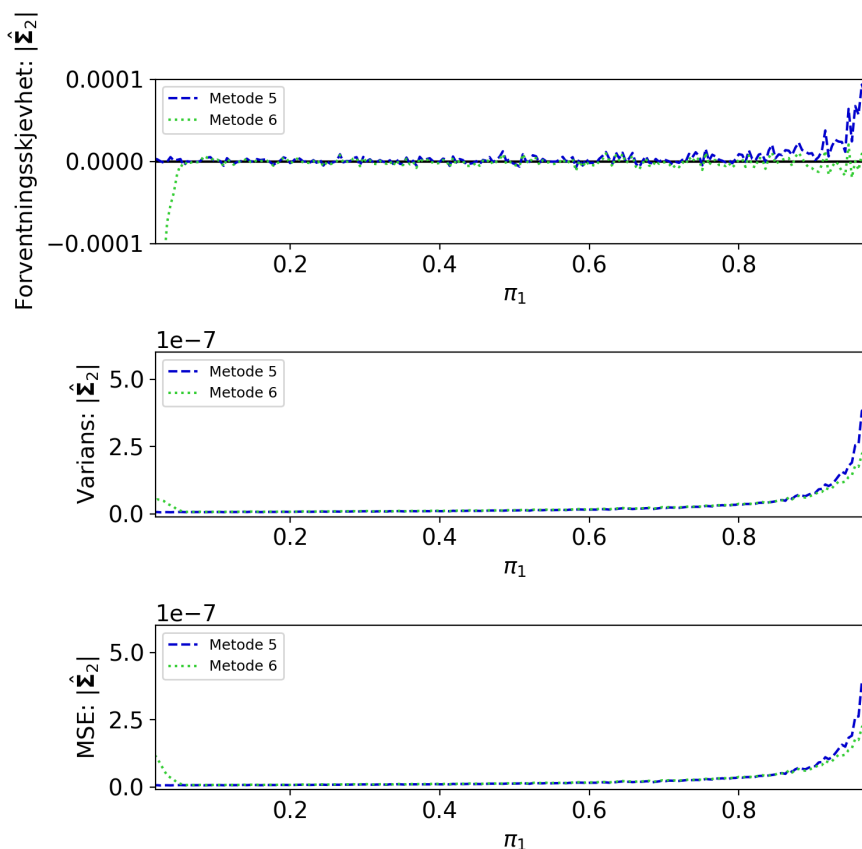


Figur 5.5: Figuren viser hvor mye feil som er gjort på grunn av ombytting av klasse 1 og 2

Estimatene til $|\Sigma_1|$ og $|\Sigma_2|$ er vist i henholdsvis figur 5.6 og 5.7. Her ser det ut til at $|\hat{\Sigma}_1|$ er overestimert ved for lave verdier av π_1 , en tendens som også ble observert for metoden 1 til 4. Dette kommer antagelig av at klasse 1 har mindre data, noe som gjør at $|\Sigma_1|$ blir vanskeligere å estimere og dermed mer feil. Variansen i dette området er også høyere, og er dominerende i MSE. Også for høye verdier av π_1 er det en del feil for $|\hat{\Sigma}_1|$ i metode 5. Her blir estimatet til $|\Sigma_1|$ underestimert. I figur 5.7 ser det derimot ut til at figurene er speilet. For høye verdier av π_1 er det mest feil og varians. Er sannsynligheten høy for klasse 1, vil den være lav for klasse 2. Det kan da se ut til at også estimatoren til $|\Sigma_2|$ blir dårligere når det er mindre data fra denne klassen.



Figur 5.6: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_1|$ ved varierende π_1



Figur 5.7: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_2|$ ved varierende π_1

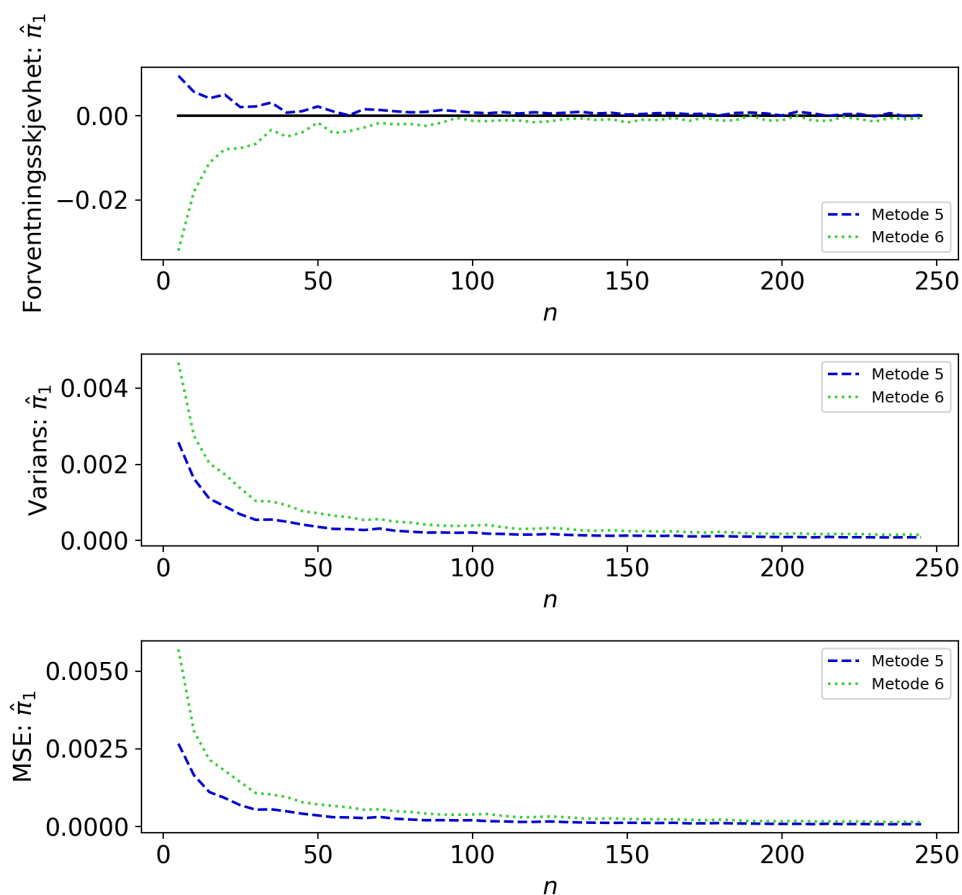
5.3 Varierende utvalg n

I denne simuleringstudien er n variert. Variabelen n er det totale antall datapunkter som er brukt for å estimere de ukjente parametrene. Jeg startet med $n = 5$ og økte med 5 for hver nye verdi n frem til $n = 250$. For hver nye verdi av n , ble det gjort 1000 estimater av hver ukjent parameter. Disse er videre bruk til å få et estimat for variansen og forventningsskjevheten. Variablene som er brukt vises i tabell 5.4. I tillegg er kovariansmatrisene de samme som i simuleringstudien med varierende L (Se ligning 5.1).

L	π_1	n	$ \Sigma_1 / \Sigma_2 $
3	0,4	Varies fra 5 til 250	$\approx 0,0025$

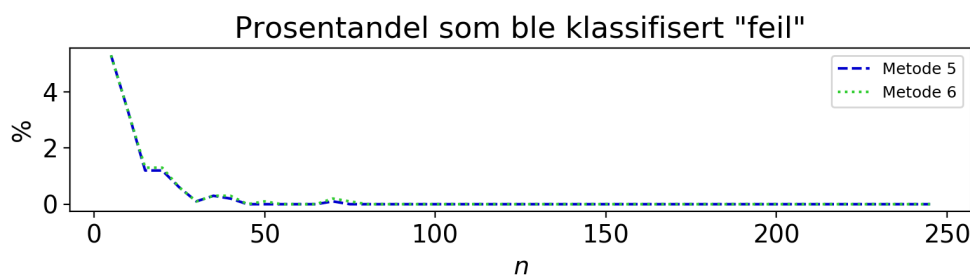
Tabell 5.4: Verdier som ble brukt i simuleringstudie, der utvalget n er variert.

I figur 5.8 kan det se ut som om variansen og forventningsskjevheten går mot null når utvalget n øker. Dette viser at estimatorene for $\hat{\pi}_1$ er en konsistent. Sammenlignes dette med metode 2 og 5, som bruker de samme type momentene, ser det ut til at begge disse metodene har en tendens til å overestimere for lave utvalg n . Det er også mulig å se likhet mellom metode 4 og 6, der metodene ser ut til å underestimere. Også i denne simuleringen er variansen dominerende i MSE.



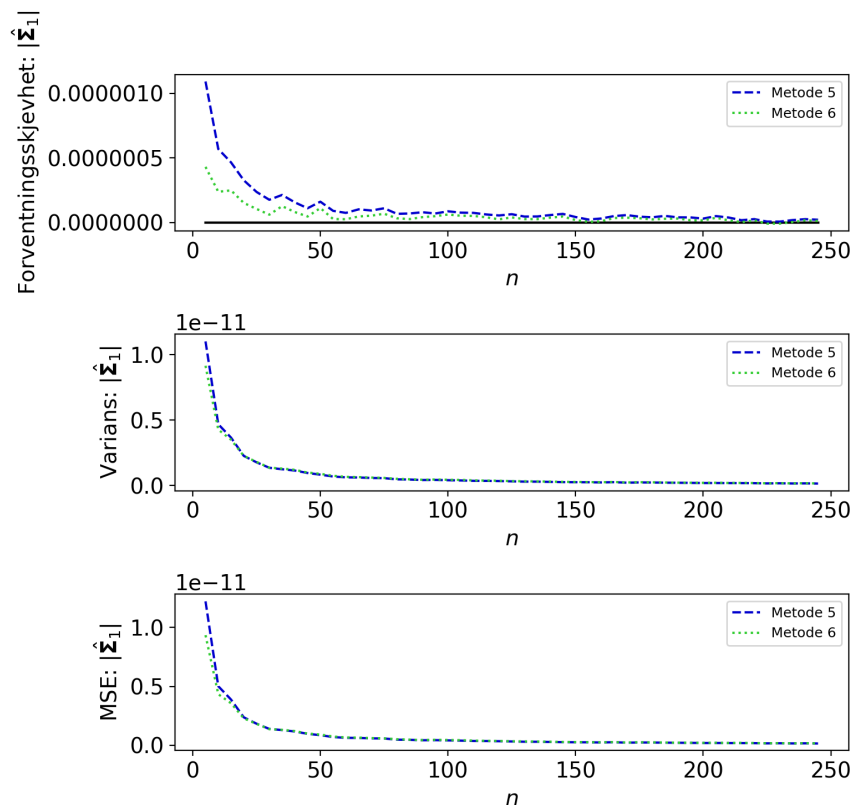
Figur 5.8: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ når antallet punker n er variert.

Figur 5.9 viser hvor mye feil metodene har gjort på grunn av at $\hat{\pi}_1$ er plassert i klasse 2 og omvendt. Dette gjøres i størst grad for små utvalg n , og dette kan også være med på å påvirke hvor godt estimatorene fungerer for små utvalg. Når utvalget øker til over $n = 50$, ser det ut til at denne typen feil er minimale. Dette vil også avhenge av de andre faktorene, som for eksempel hvor stor forskjell det er på $|\hat{\Sigma}_1|$ og $|\hat{\Sigma}_2|$. Dette vil jeg se næyere på i simuleringstudien der $|\hat{\Sigma}_1|/|\hat{\Sigma}_2|$ blir variert (se delkapittel 5.4).

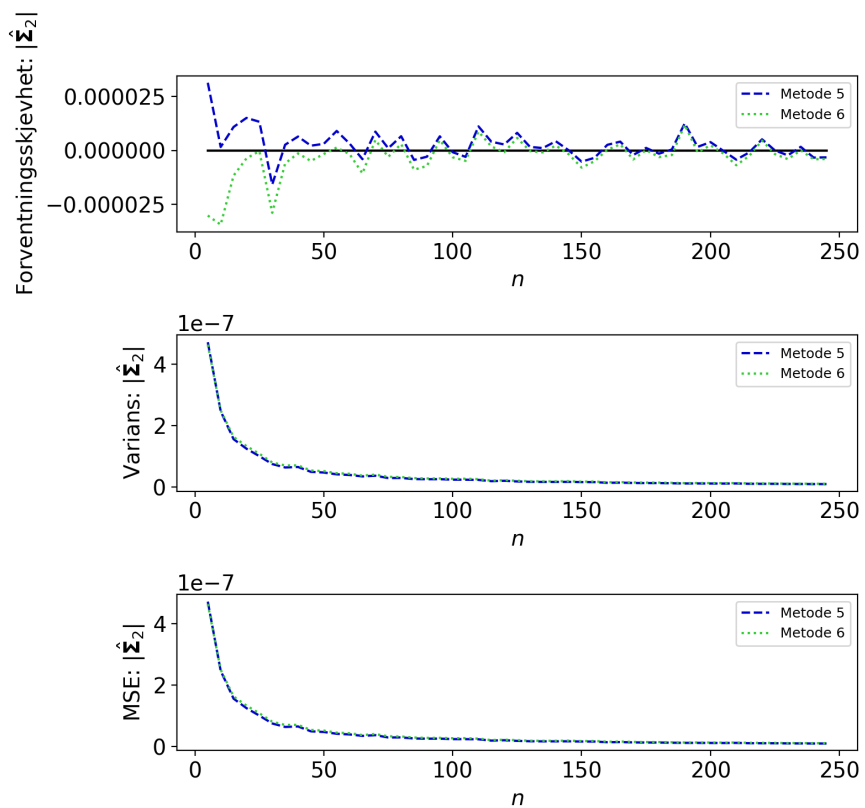


Figur 5.9: Figuren viser hvor mye feil som er gjort på grunn av ombytting av klasse 1 og 2

Ser man på figur 5.10, får estimatene $|\hat{\Sigma}_1|$ bedre egenskaper når utvalget n øker. Her går både variansen og forventningsskjevheten ned ved økende utvalg n , noe som viser at også denne estimatoren er konsistent. Det samme gjelder $|\hat{\Sigma}_2|$, men her er det mindre tydelig at forventningsskjevheten avtar. Estimatoren ser likevel ut til å få mindre MSE når utvalget n øker, noe som tyder på at også denne estimatoren blir bedre ved større utvalg.



Figur 5.10: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_1|$ når antallet punker n er variert.



Figur 5.11: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_2|$ når antallet punker n er variert.

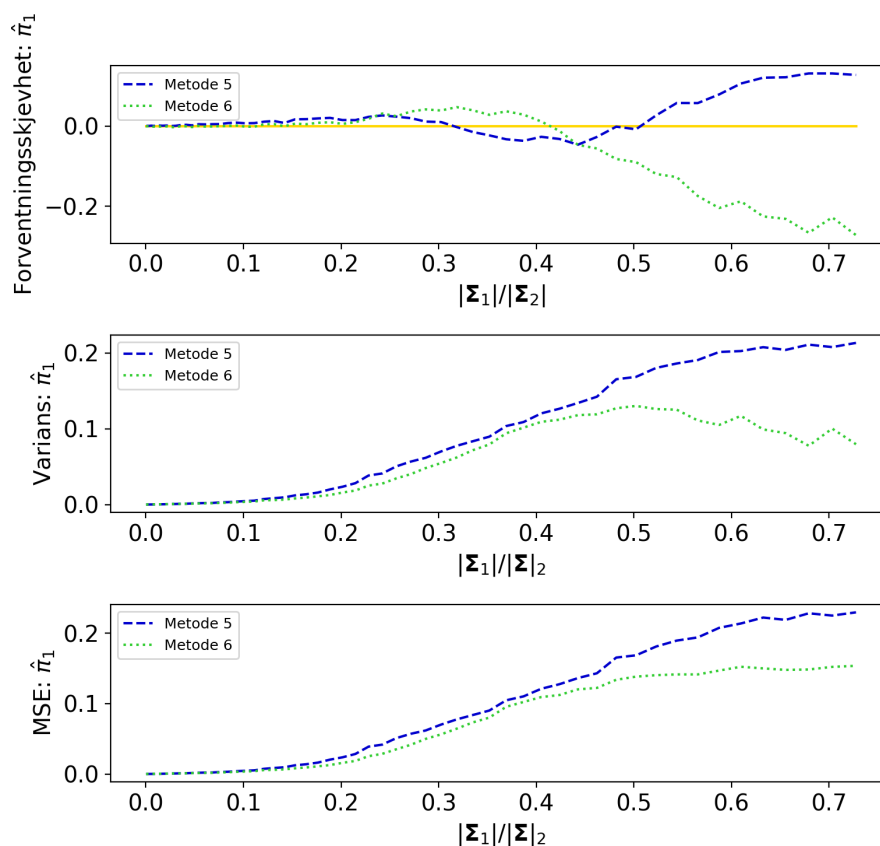
5.4 Varierende $|\Sigma_1|/|\Sigma_2|$

Når forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$ nærmer seg én vil det bli vanskeligere å skille mellom klasse 1 og klasse 2. I denne simuleringstudien vil jeg undersøke dette, der jeg vil holde $|\Sigma_2|$ konstant og variere $|\Sigma_1|$. Dette er gjort slik at forholdet $|\Sigma_1|/|\Sigma_2|$ vil øke mot 1. Variablene brukt i denne simuleringen er gitt i tabell 5.5. I tillegg vil $|\Sigma_2|$ være den samme som $|\Sigma_2|$ i ligning (5.1).

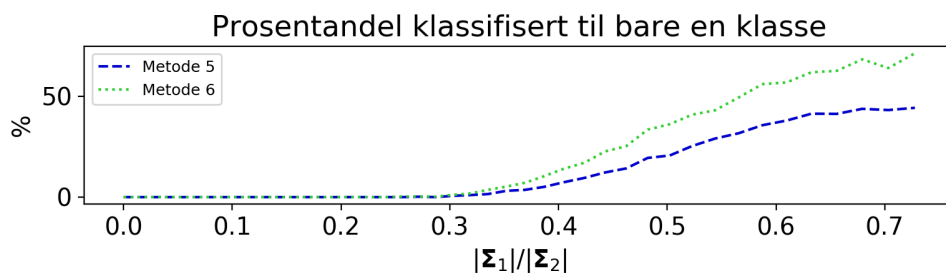
L	π_1	n	$ \Sigma_1 / \Sigma_2 $
3	0,4	$15^2 = 225$	varierer fra 0,1 til 0,7

Tabell 5.5: Verdier som er brukt i simuleringstudie, der $|\Sigma_1|/|\Sigma_2|$ er variert.

Figur 5.12 viser hvordan metode 5 og 6 håndterer at forskjellene mellom $|\Sigma_1|$ og $|\Sigma_2|$ blir mindre. I denne simuleringstudien ser det ut til at metodene håndterer situasjonen ganske bra frem til $|\Sigma_1|/|\Sigma_2| = 0,2$. For $|\Sigma_1|/|\Sigma_2|$ over 0,2 øker både variansen og forventningsskjevheten. Variansen til de to metodene ser ut til å være ganske lik frem til $|\Sigma_1|/|\Sigma_2| = 0,4$, men etter dette har metode 6 mindre varians. Dette betyr ikke at metode 6 er bedre, men dette kan ha sammenheng med hvor mange ganger metoden bare klassifiserer til én klasse. Hvor mye dette skjer er vist i figur 5.13. Dette ser ut til å skje når $|\Sigma_1|/|\Sigma_2| = 0,4$ og oppover, noe som samsvarer med når variansen til metodene skiller lag (se midterste figur i figur 5.12).

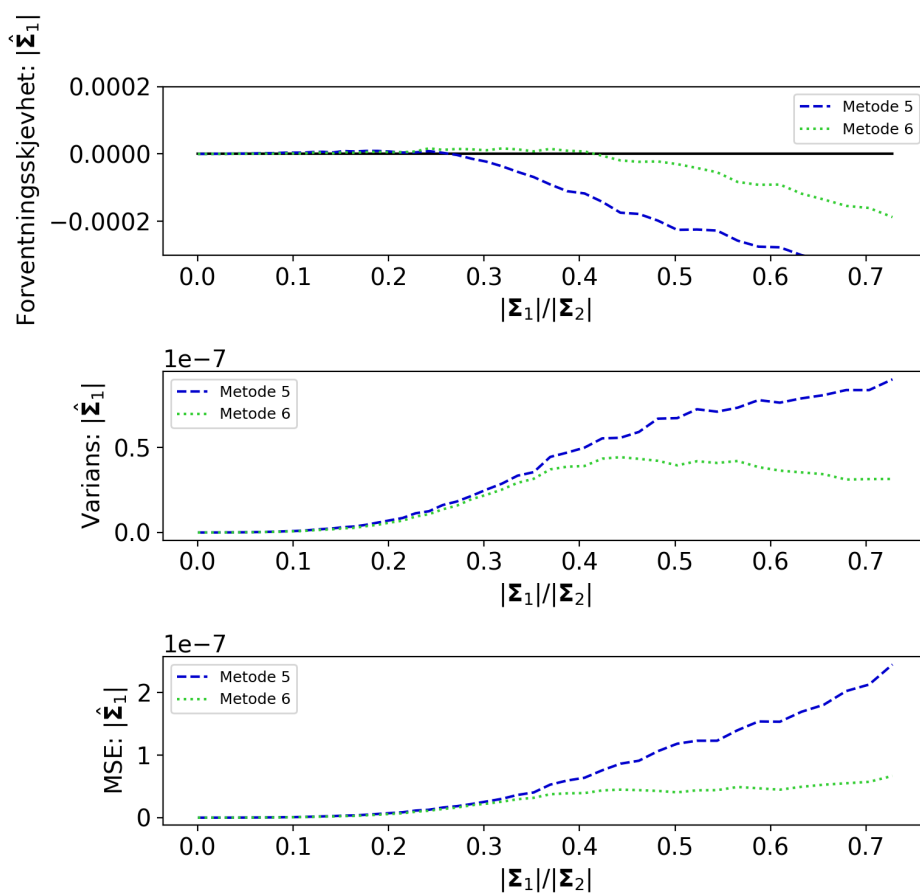


Figur 5.12: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ når forholdet mellom $|\Sigma_1|/|\Sigma_2|$ blir variert.

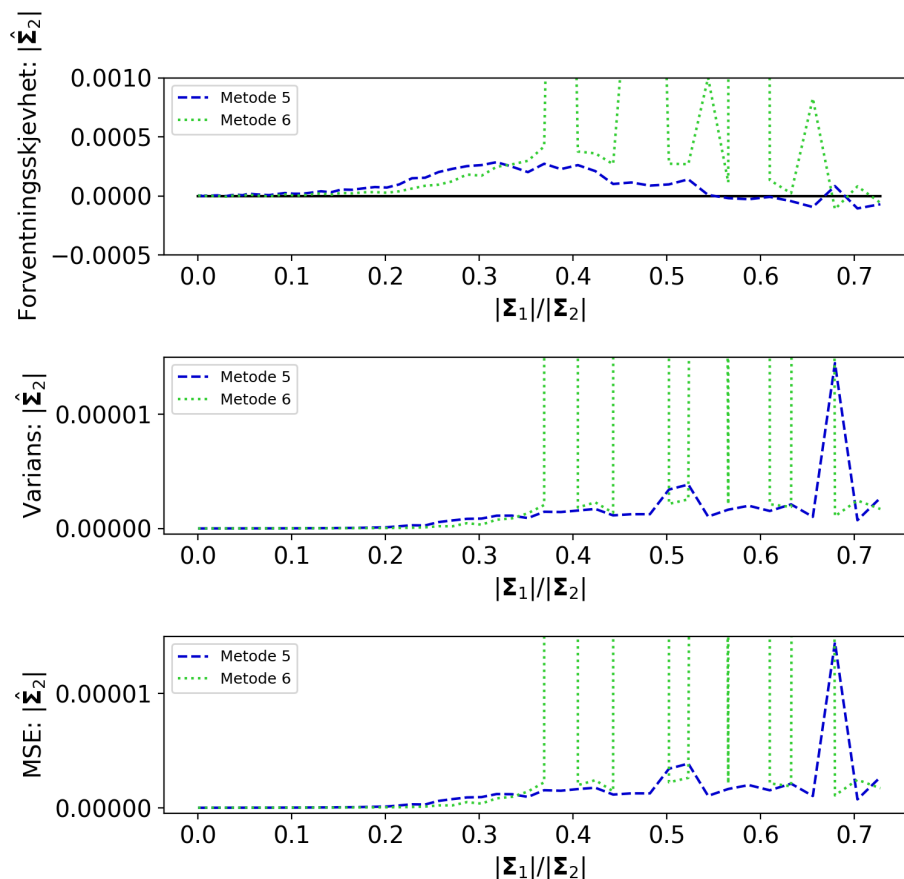


Figur 5.13: Figuren viser hvor stor prosentandel av blandingsutvalgene som er klassifisert som bare én klasse når $|\Sigma_1|/|\Sigma_2|$ varierer.

Estimatene til $|\Sigma_1|$ i figur 5.14, ser også ut til å få høy varians ved $|\Sigma_1|/|\Sigma_2| \geq 0,2$. Ser man på forventningsskjevheten, øker ikke denne drastisk før $|\Sigma_1|/|\Sigma_2| \geq 0,3$ for metode 5 og $|\Sigma_1|/|\Sigma_2| \geq 0,4$ for metode 6. For estimatene til $|\Sigma_2|$ i figur 5.14, ser det ut til at metode 6 har større problemer enn metode 5. Både variansen og forventningsskjevheten ser ut til å være betydelig når $|\Sigma_1|/|\Sigma_2| \geq 0,3$.

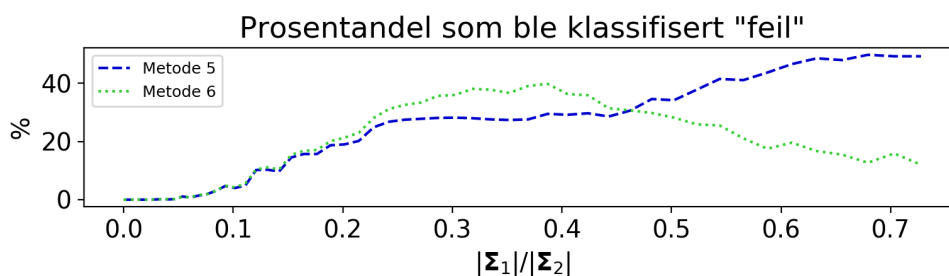


Figur 5.14: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_1|$ når forholdet mellom $|\Sigma_1|/|\Sigma_2|$ blir variert.



Figur 5.15: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_2|$ når forholdet mellom $|\Sigma_1|/|\Sigma_2|$ blir variert.

Det er også mulig å se hvor mye feil som kommer av at $\hat{\pi}_1$ er klassifisert til klasse 2 og motsatt. Dette er sett på i figur 5.16, og ut fra denne figuren ser det ut til at det blir mye feil allerede ved $|\Sigma_1|/|\Sigma_2| = 0,1$. Ved $|\Sigma_1|/|\Sigma_2| = 0,2$ er det allerede opp mot 20 % som er klassifisert feil.



Figur 5.16: Figuren viser hvor mye som er gjort feil på grunn av ombytting av klasse 1 og 2

Simuleringsstudien viser at metoden 5 og 6 hadde problemer for $|\Sigma_1|/|\Sigma_2| \geq 0,2$, men metoden har antagelig også litt problemer allerede ved $|\Sigma_1|/|\Sigma_2| = 0,1$. Dette gjelder for de variablene som ble brukt i denne simuleringsstudien og blant annet kan utvalget påvirke dette (se simuleringsstudien for varierende μ_1/μ_2 for intensitetsdata, metode 1 til 4).

5.5 Varierende korrelasjon mellom kopolariserte kanaler

I denne simuleringssstudien ble det sett på når korrelasjon mellom de kopolariserte kanaler varieres. Dette er gjort for både klasse 1 og klasse 2. For en todimensjonal kovariansmatrise med S_{hh} og S_{vv} , vil korrelasjonen mellom disse være $E[S_{hh} \cdot S_{vv}^*]$ og $E[S_{vv} \cdot S_{hh}^*]$, se ligning 2.4) Disse uttrykkene vil ha både en imaginær og en reell del, der $E[S_{hh} \cdot S_{vv}^*]$ er den komplekskonjugerte til $E[S_{vv} \cdot S_{hh}^*]$. Det er da mulig å uttrykke korrelasjonen mellom de kopolariserte kanalene på polarform:

$$E[S_{hh} \cdot S_{vv}^*] = \varrho e^{i\varsigma}$$

$$E[S_{vv} \cdot S_{hh}^*] = \varrho e^{-i\varsigma},$$

der ϱ er lengden til $E[S_{hh} \cdot S_{vv}^*]$, mens ς er retningen (Lindstrøm, 2006). Determinanten til en todimensjonal kovariansmatrise Σ bli da:

$$|\Sigma| = E[S_{hh} \cdot S_{hh}^*] \cdot E[S_{vv} \cdot S_{vv}^*] - \varrho^2.$$

Siden ς ikke vil endre på determinanten til kovariansmatrisa, har jeg valgt å definere:

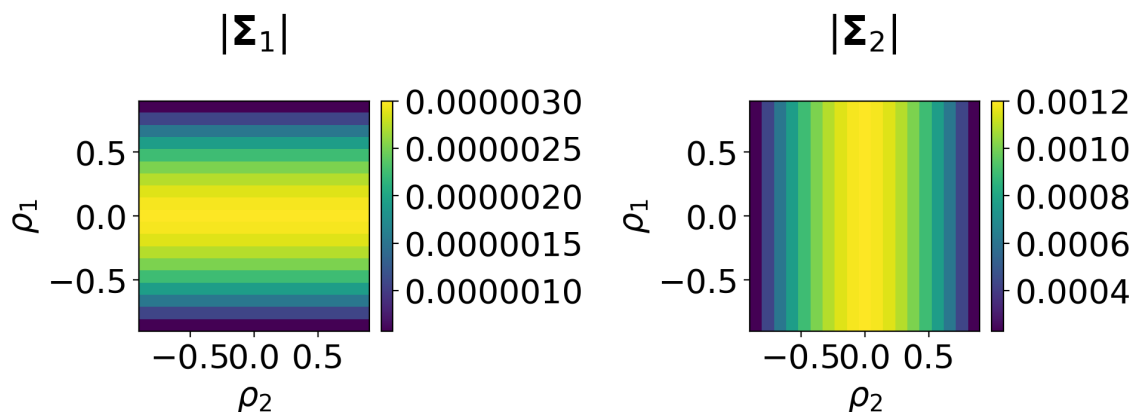
$$\rho \equiv \frac{\varrho}{\sqrt{E[S_{vv} \cdot S_{vv}^*] \cdot E[S_{hh} \cdot S_{hh}^*]}}. \quad (5.2)$$

Utrykket i ligning (5.2) kaller jeg korrelasjonskoeffisienten og den vil være normalisert mellom -1 og 1. Korrelasjonskoeffisienten til klasse 1 og klasse 2 definerer jeg som henholdsvis ρ_1 og ρ_2 . Ved å variere ρ_1 og ρ_2 vil korrelasjonen mellom S_{hh} og S_{vv} endre seg. I denne simuleringstudien har jeg valgt å variere både ρ_1 og ρ_2 mellom -1 og 1, der resultatene vil bli vist i bilder. For hver nye kombinasjon av ρ_1 og ρ_2 estimeres 1000 verdier, som så brukes til å finne et estimat for variansen og forventingskjevhet til estimatene π_1 , $|\Sigma_1|$ og $|\Sigma_2|$. Verdiene som ble brukt i simuleringene er vist i tabell 5.6.

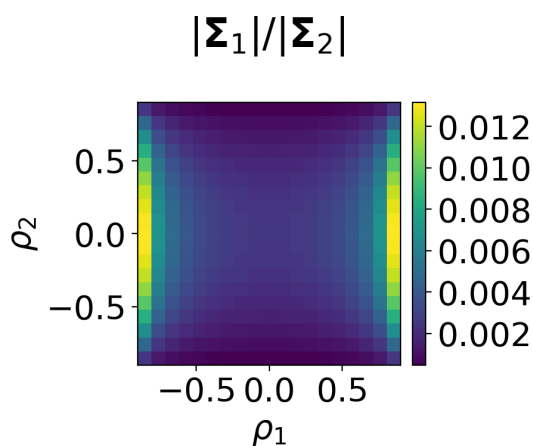
L	π_1	n	ρ_1	ρ_2	$ \Sigma_1 / \Sigma_2 $
3	0,4	$15^2 = 225$	$\in [-1, 1]$	$\in [-1, 1]$	varierer fra ca. 0,013 til 0,005

Tabell 5.6: Verdier som ble brukt i simuleringssstudien, der kovariansen mellom S_{hh} og S_{vv} er variert i de to klassene.

Figur 5.17 viser hvordan determinanten til de to matrisene $|\Sigma_1|$ og $|\Sigma_2|$ endrer seg når ρ_1 og ρ_2 endres. Siden $|\Sigma_1|$ og $|\Sigma_2|$ endrer seg, vil også forholdet $|\Sigma_1|/|\Sigma_2|$ endres. Dette er vist i figur 5.18. Diagonalen til matrisa endrer seg ikke, og er lik som i ligning (5.1).



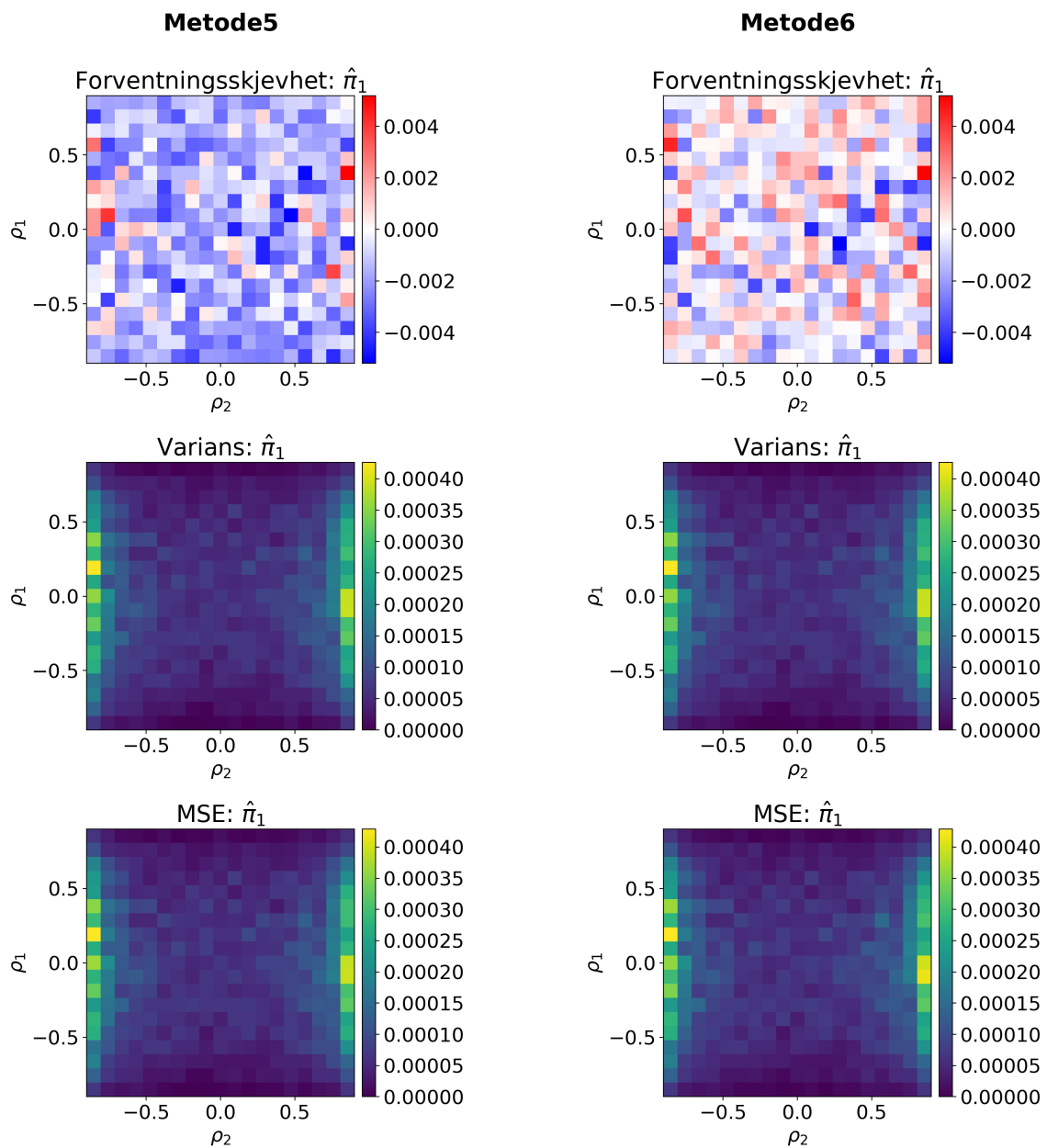
Figur 5.17: Endring av $|\Sigma_1|$ og $|\Sigma_2|$ når ρ_1 og ρ_2 varierer. ρ_1 varierer langs y-aksen, mens ρ_2 varierer langs x-aksen. Diagonalen til kovariansmatrisa er som i ligning (5.1).



Figur 5.18: Forholdet $|\Sigma_1|/|\Sigma_2|$ når ρ_1 og ρ_2 varierer langs henholdsvis y- og x-aksen.

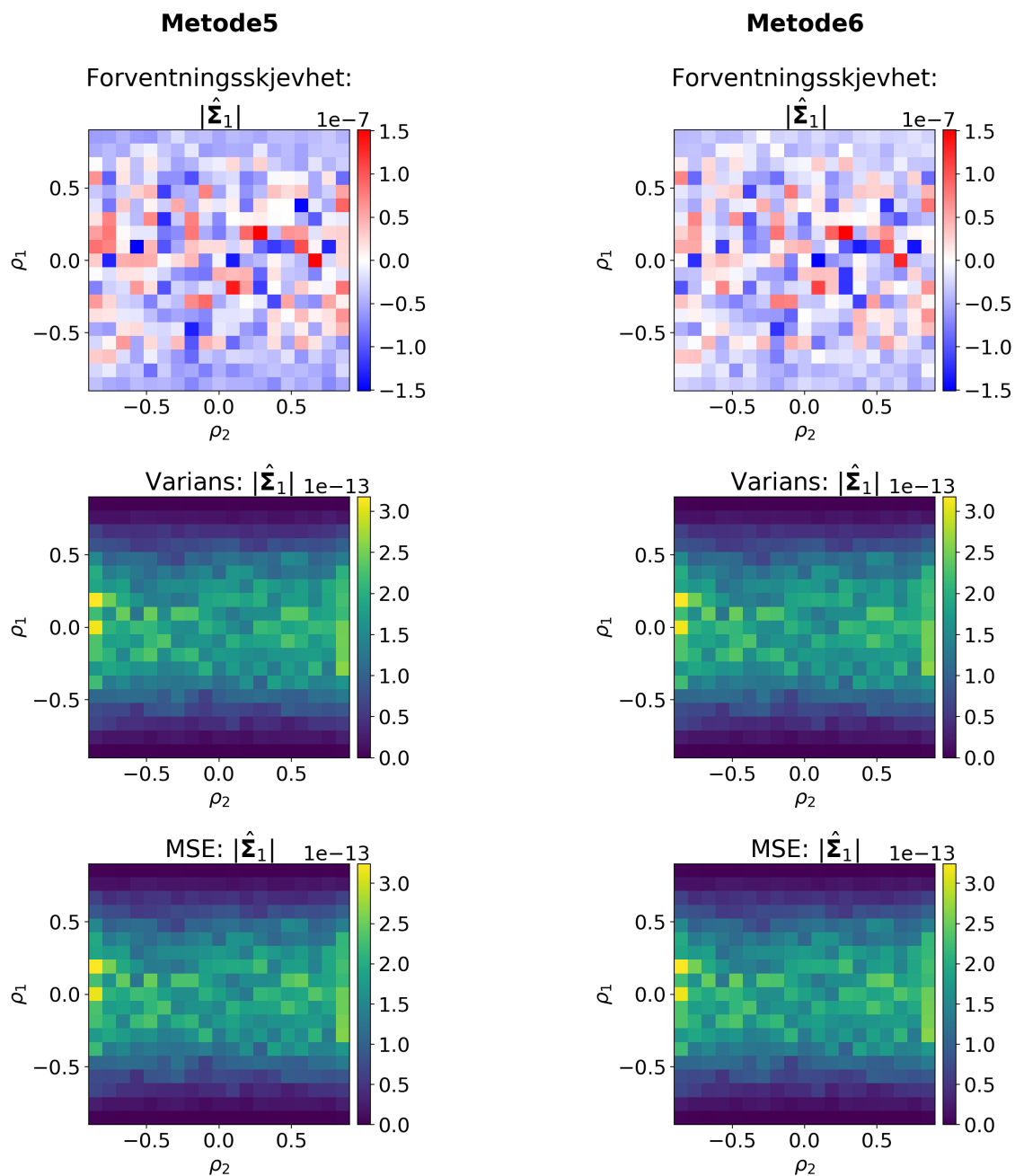
For hver egenskap i figur 5.19 til 5.21 er aksene holdt like, slik at metoden skal kunne sammenlignes. Figur 5.19, viser at metode 5 har en tendens til å underestimere $\hat{\pi}_1$, mens metode 6 ser ut til å være en blanding av over- og underestimer. Variansen ser ut til å være størst når $\rho_1 \approx 0$ og $\rho_2 \approx \pm 1$. Siden klasse 1 og 2 er definert som kovariansmatrisa med henholdsvis minst og størst determinant, vil man ved $\rho_1 \approx 0$ og $\rho_2 \approx \pm 1$ få den minste forskjellen mellom $|\Sigma_1|$ og $|\Sigma_2|$. Forholdet mellom $|\Sigma_1|$ og $|\Sigma_2|$ er vist i figur 5.18. Denne figuren har store likhetstrekk med variansen til $\hat{\pi}_1$. Siden $|\Sigma_1|$ og $|\Sigma_2|$ vil bli vanskeligere å skille når $|\Sigma_1|$ og $|\Sigma_2|$ nærmer seg hverandre, gir det mening at dette kan ha innvirkning på variansen.

Jeg har også undersøkt hvor mye feil som er gjort på grunn av at $\hat{\pi}_1$ er klassifiser til klasse 2 og motsatt. I denne simuleringstudien var det ingen slike feil.



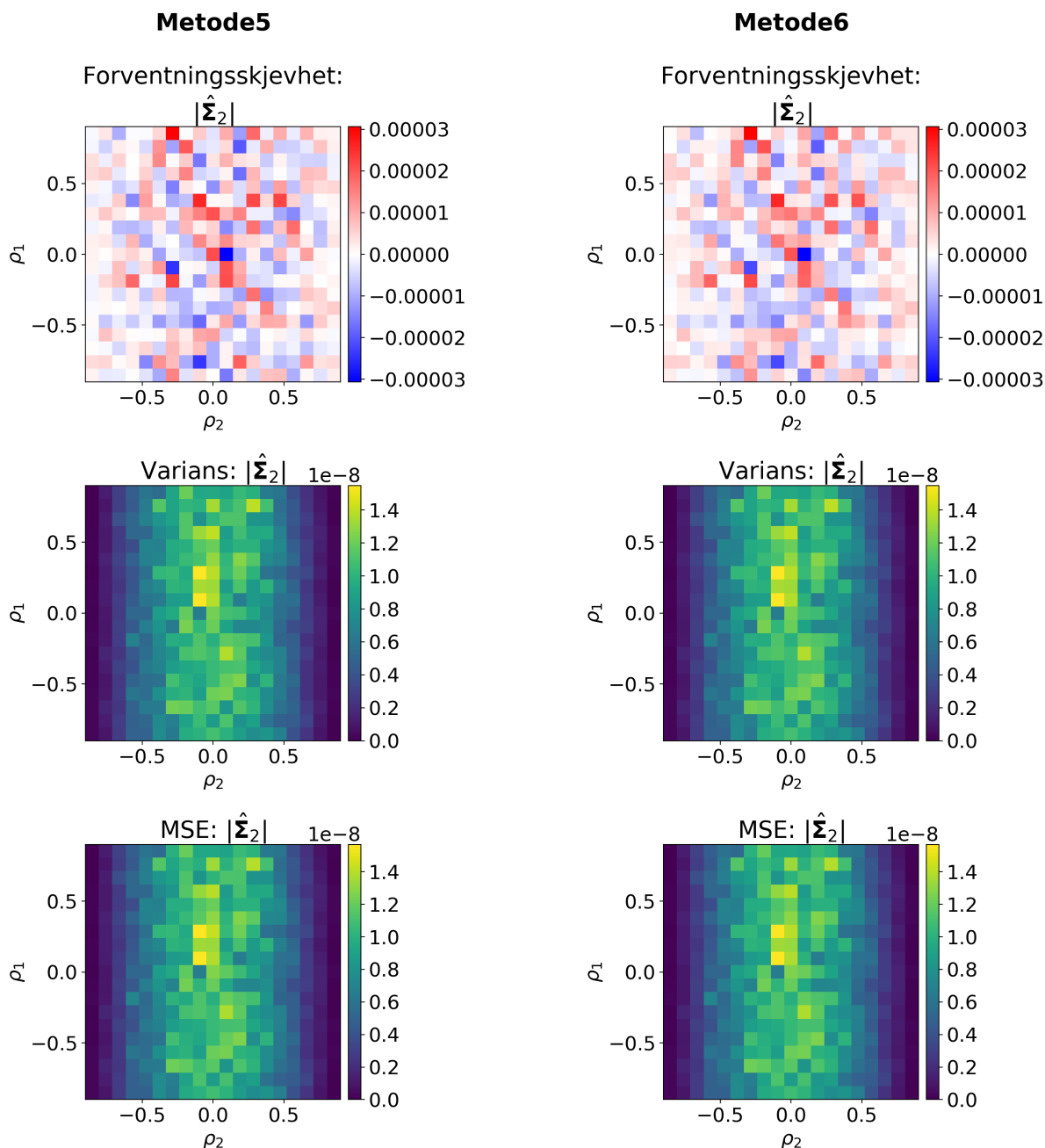
Figur 5.19: Egenskapene skjevhet, varians og MSE til $\hat{\pi}_1$ når både ρ_1 og ρ_2 blir variert

Jeg har også sett på estimatene $|\hat{\Sigma}_1|$ og $|\hat{\Sigma}_2|$, der figur 5.20 ser på estimatene til $|\Sigma_1|$. I denne figuren ser det ut til at $|\hat{\Sigma}_1|$ blir underestimert når ρ_1 er nær ± 1 . Når ρ_1 nærmer seg 0, blir forventningsskjevheten mindre og variansen ser ut til å øke. Denne økningen kommer antagelig av at $|\hat{\Sigma}_1|$ er høyere i dette området (Se figuren til venstre i figur 5.17). Dette er en tendens som også kan ses i andre simuleringstudier der $|\hat{\Sigma}_1|$ som er har lavest verdi, også har lavere varians. Det kan også se ut til at forventningsskjevheten og variansen ikke avhenger av hva $|\hat{\Sigma}_2|$ er.



Figur 5.20: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_1|$ når både ρ_1 og ρ_2 blir variert.

Ser man på egenskapene til $|\hat{\Sigma}_2|$, så ser det ut til at estimatene har mest forventningskjevhet når ρ_2 er nær 0. Det er også her variansen og $|\Sigma_2|$ er størst. $|\Sigma_2|$ er vist i figuren til høyre i figur 5.2. Med de variablene som er brukt i denne simuleringsstudien, ser det ikke ut til å være noen tendenser til å over- eller underestimere de to metodene 5 og 6. Det ser også ut til at $|\hat{\Sigma}_1|$ har liten påvirkning på variansen og forventningskjevheten til estimatet $|\hat{\Sigma}_2|$.



Figur 5.21: Egenskapene skjevhet, varians og MSE til $|\hat{\Sigma}_2|$ når både ρ_1 og ρ_2 blir variert.

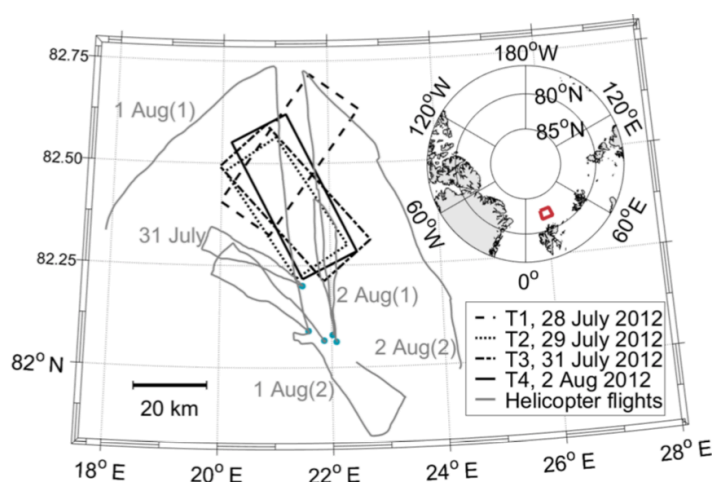
Kapittel 6

Reelle data

I dette kapitlet vil jeg teste metodene på et reellt datasett fra satellitten TerraSAR-X. Datasettet har også en bakkesannhet, og denne vil jeg bruke til å validere metodene. Resultatene fra metode 1 til 6 vil i tillegg bli sammenlignet med resultatene til Fors mfl. (2017), som bruker de samme datasettene.

6.1 Datasett

Datasettene som skal brukes består av satellittbilder og data fra helikopter. Området som datasettene dekker befinner seg litt nord for Svalbard (se figur 6.1), og besto av drivis som ikke var eldre enn ett år. Siden isflakene er i drift, kan de bli påvirket av flere faktorer som strøm, kollisjoner, tidevann og værforhold. Dataene ble samlet i forbindelse med kampanjen ICE2012, og består av data fra 28. juli til 2. august (Fors mfl., 2017). Det er fokusert på to av satellittbildene, der området er begrensa til ett spesifikt isflak. Dette flaket hadde også innsamlet bakkesannhet fra helikopter. I figur 6.1 er alle satellittbildene markert med rektangler over de områdene bildene dekker. De to bildene jeg har brukt er satellittbilde T3 og T4.



Figur 6.1: Figuren viser området satellittbildene dekker og helikopterrutene som samlet inn bakkesannheten. Satellittbildene er markert med rektangler, og har navnene T1 til T4 (Fors mfl., 2017). I venstre hjørne er det også lagt til et kart som viser Arktis, der utsnittet som er sett på er markert med rødt. (Gjengitt med velvillig tillatelse fra A. S. Fors)

6.1.1 Satellitdata

Radarbildene er fra satellitten TerraSAR-X. Denne satellitten kan sende med både vertikal (V) og horisontal (H) polarisasjon på den elektromagnetiske strålingen (EM-stråling). Det er også disse to polarisasjonene satellitten kan motta (Fritz, 2007). Dette gjør det mulig for satellitten å måle opp til fire polarimetrisk komponenter VV, VH, HH og HV. Den første bokstaven angir sendt polarisasjon, og den siste angir mottatt polarisasjon. Selv om det er mulig å måle alle, er det vanligvis bare VV og HH som tas opp av TerraSAR-X. Dette er også tilfellet i dette datasettet. Ellers sender radaren ut EM-stråling med en frekvens på 9,65 MHz (Fritz, 2007), noe som er innenfor frekvensområdet til X-båndet (Elachi og Zyl, 2006).

I oppgaven vil jeg bruke to bilder som er tatt på forskjellige dager. Disse to bildene har fått merkelappene T3 og T4, og dette er de samme navnene som er brukt i Fors mfl. (2017). Tabell 6.1 inneholder informasjon om disse to satellittbildene og værforholdene.

Navn	dato	tid	innfallsvinkel	Vindhastighet	Lufttemperatur
T3	31. Juli 2012	13:51	29,4°	6,2 m/s	-0,8°C
T4	2. August 2012	14:51	44,2°	0,6 m/s	0,8°C

Tabell 6.1: Oppsummering av satellittdata for T3 og T4(Fors mfl., 2017).

Radaren vil også motta termisk støy og under støygulvet vil man ikke kunne skille mellom denne støyen og signalet (Wackerman, 1992; Elachi og Zyl, 2006). Dette kan være problematisk, og dersom en andel av dataen under støygulvet, vil dette svekke resultatene. For satellittbildet T4 er ca. 25 % av det tilbakespredde signalet under støygulvet (Fors mfl., 2017). Satellittbildet T3 gir et bedre utgangspunkt, der mindre enn 10 % er under støygulvet (Fors mfl., 2017). De to satellittbildene har også forskjellig oppløsning, der T4 har en oppløsning på 2,4 m i asimutal retning (satellittens fartsretning) og 1,9 m vinkelrett på fartsretningen. For satellittbilde T3 er oppløsningen 3,0 m i asimutal retning og 1,3 m vinkelrett på fartsretningen. Ved estimering av kovariansmatrisa vil imidlertid oppløsninga gå ned, avhengig av hvor mye data som midles over i estimatene.

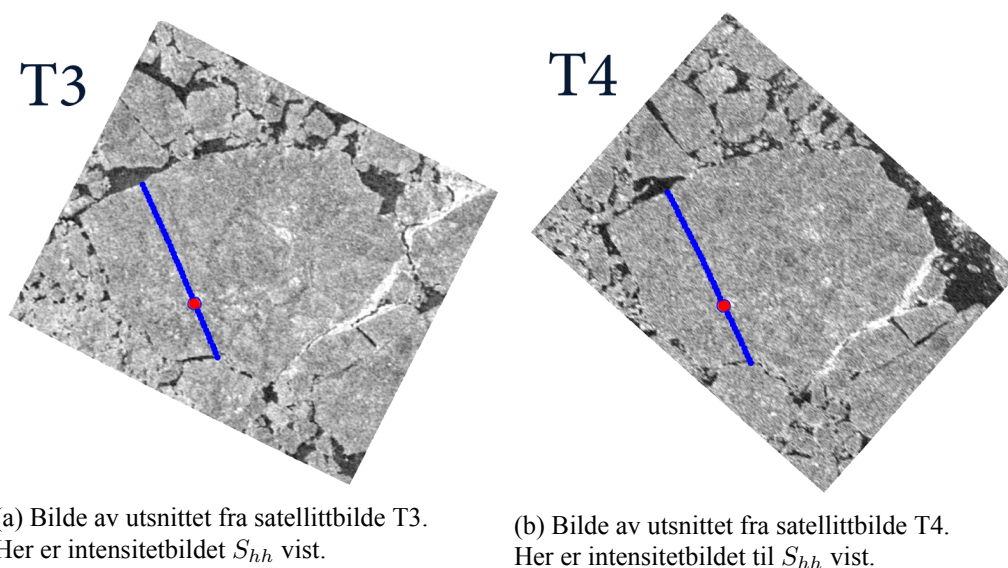
6.1.2 Helikopterdata

Bakkesannheten er laget ut fra data samlet inn med helikopter. Her er det tatt flere bilder ned mot bakken og ut fra disse bildene er andelen smeltevann og hav i hvert bilde regnet ut (Divine mfl., 2015). I denne oppgaven er det fokuset på smeltevannandelen og estimering av denne. Helikopteret fløy flere ganger for å samle inn data, men siden isen er i drift, er det ikke like lett å samsvare dataene fra helikopteret og satellitten. Området man ser på er begrenset til ett spesifikt isflak som var å se i både satellittbilde T3 og T4. For disse to bildene ble det laget en bakkesannhet med data fra helikoptertur nummer 2, 2. august 2012 (Divine mfl., 2017). Her er det litt tidsforskjell mellom helikopterturene og satellittbildene, og dette kan gi feil i bakkesannheten. Det vil si at forholdene på bakken kan ha endret seg mellom bildene ble tatt og bakkesannheten ble samlet inn. Dette gjelder spesielt T3, som er tatt ca. 2 dager før. (Fors mfl., 2017)

har manuelt justert helikopterdataene mot satellittbildene T3 og T4, der feilen på justeringa er anslått til maksimalt 7 m horisontalt og vertikalt. Denne korrigeringen vil også bli brukt i denne oppgaven. Divine mfl. (2015) gir mer informasjon om hvordan bakkesannheten er laget.

6.2 Satellittbildene T3 og T4

Figur 6.2a og 6.2b viser utsnittet med isflaket som det er sett på. I disse bildene har jeg filtrert med et relativt stort vindu for å vise satellittdataen og helikopterruta over isflaket. Jeg kommer ikke til å bruke så store vinduer når metode 1 til 6 undersøkes og testes. Grunnen til dette er at store vinduer under filtreringen vil gi mange blanda piksler og i verste fall kan smeltesammene bli filtrert samene med ispiksler. For å estimere kovariansmatrisa og intensiteten har jeg gjennomsnittsfiltrert de ufiltrerte dataene med to vinduer. Det ene filteret midler over 4 piksler totalt mens det andre midler over 9. Den estimerte verdien for L med disse vinduene er henholdsvis $L \approx 2$ og $L \approx 3$. Det er usikkerhet ved estimeringa av L , der estimatet ved bruk av 4 piksler varierte mellom 1,98 og 2. Det så ut til at estimatene for satellittbilde T3 og T4 ga ganske like resultater. Ved bruk av 9 piksler varierte L det litt mer, noe som kan komme av at det var færre piksler totalt i bildeutdragene siden bildene også er nedsamplet i filtreringen. L varierte her fra 2,66 til 3,03, der det så ut til at T4 hadde litt større estimater for L enn T3. Jeg vil videre bruke $L = 2$ for kovariansmatrisa som er estimert fra 4 piksler, og $L = 3$ for kovarians som er estimert fra 9 piksler. Det er gjort flere delestimeringer for små utvalg av satellittdataene og L er beregnet ut fra disse. Metoden er beskrevet i Anfinsen, Doulgeris og Eltoft (2009). Det er noen problemer ved bruk av denne metoden, og ett av dem oppstår når man ser på vinduer som inneholder flere klasser eller tekstur. I disse tilfellet vil L bli underestimert (Anfinsen, Doulgeris og Eltoft, 2009). Jeg antar at det over isflakene vil være to klasser i metode 1 til 6. Om denne antagelsen stemmer også for små utvalg, vil dette kunne påvirke estimatet av L . I figur 6.2a og 6.2b markerer de blå og røde prikkene hvor helikopterdataene er målt. Den røde markøren viser området jeg har undersøkt nærmere i delkapittel 6.3. Bildene er projisert ned på bakken ved bruk av stereografisk projeksjon.



Figur 6.2: De blå punktene markerer stedene med helikopterdata. Punktet som er markert med en rød prikk indikerer området som er undersøkt nærmere i delkapittel 6.3.

6.3 Histogram og estimert blandingsmodell

Ifølge helikoptermålingene er andelen smeltesdammer ved den røde markøren ca. 24 % (se figur 6.2a og 6.2b). Dette vil si at det er mer is enn det er smeltesdammer. Utvalgene for estimering av π_1 , π_2 , μ_1 og μ_2 er her variert, slik at det totale antallet piksler enten er 30×30 , 48×48 eller 60×60 . Alle disse utdragene vil tilsvare større områder enn det helikopterbildene dekker, som tilsvarer areal på rundt 21×21 piksler i de ufiltrerte satellittbildene. Dette vil variere fra bilde til bilde på grunn av forskjellig oppløsning på satellittbildene og flyvehøyde på helikopteret. For estimatene til kovariansmatrisa har jeg allerede midlet over 4 piksler (antar $L = 2$) eller 9 piksler (antar $L = 3$), og disse inngår som en del av det totale antallet piksler som er brukt. Dette er grunnen til at det er brukt færre piksler når jeg bruker den estimerte kovariansmatrisa der $L = 3$ i figur 6.4 og 6.5.

Figur 6.4 viser histogrammene rundt den røde markøren i figur 6.2a. Her er det sett på tilsammen 6 utvalg fra samme områder og for hver av disse er det estimert en blandingsmodell med metode 1 til 4. Den gule heltrukkede linjen er et estimat når det ikke antas en blandingsmodell. For figurene der det totale antallet brukte piksler er 48^2 og 60^2 , ser blandingsmodellen ut til å forklare dataene på en bedre måte enn om man ikke hadde antatt en blandingsmodell. For de to minste utvalgene med totalt 30×30 piksler kan det se ut til at estimatene ikke forklarer dataene bedre enn uten en blandingsmodell. I tillegg ser det ut til at metode 1 har gjort feil når $L = 2$.

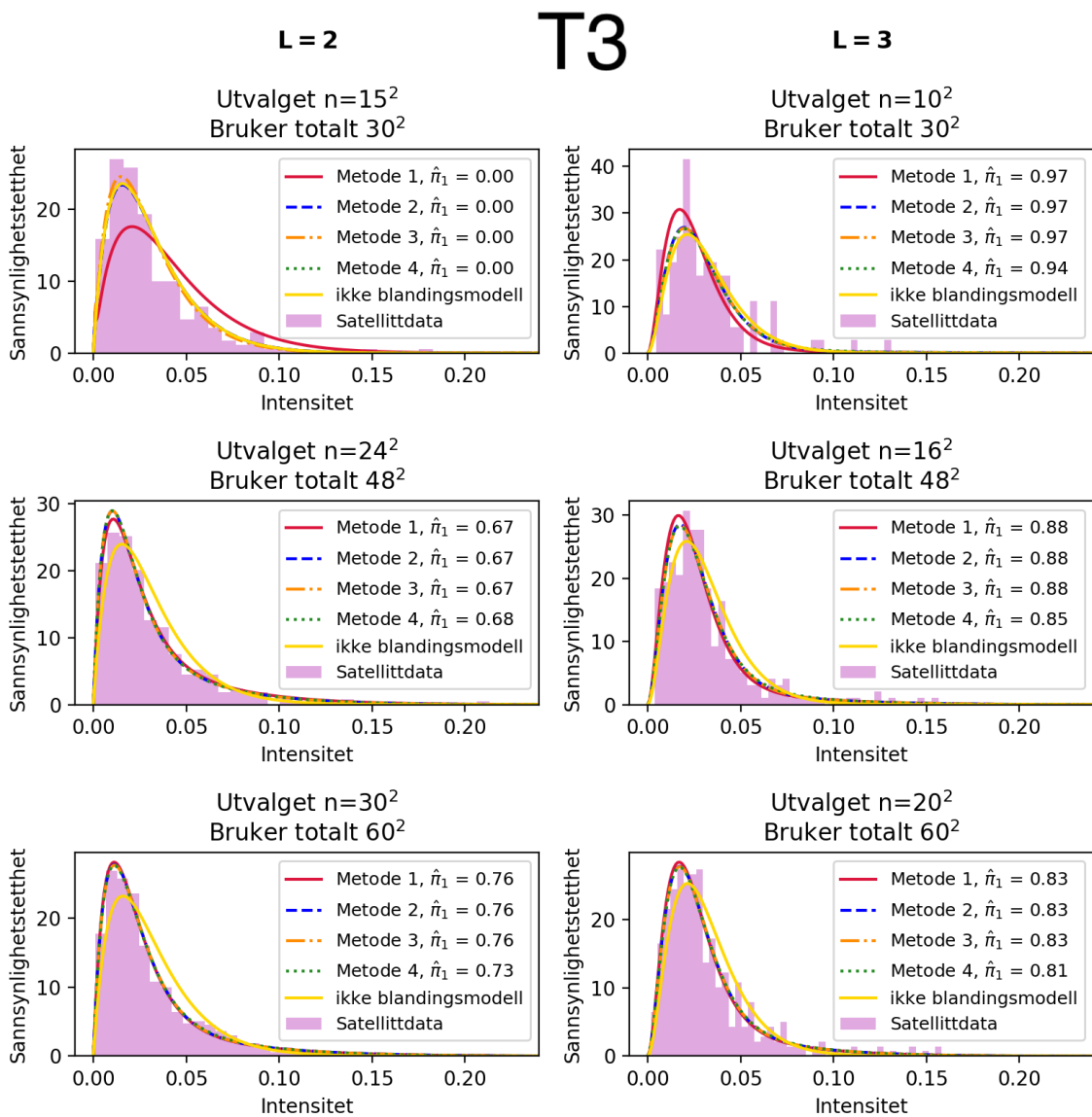
Smeltevannet er antatt å ha mindre tilbakespredning enn is, noe som gir at $\hat{\pi}_1$ er estimatet for smeltevannandelen. Denne antagelsen er basert på at vann ofte har mye overflatespredning, noe som vil føre til at store deler av den sendte strålingen vil bli reflektert bort fra satellitten. Ser man på bildet av isflaket i figur 6.2a og 6.2b, har isen høyere intensitet enn havet. I de fleste estimatene gir metode 1 til 4 veldig like estimater for $\hat{\pi}_1$, der andelen smeltesdammer varierer fra ca. 67 % til 97 % avhengig av størrelsen på området som er estimert fra og hvilken kovariansmatrise som er brukt. Unntaket vises i figuren øverst til venstre der andelen smeltesdammer er estimert til 0 %. Sammenlignes estimatene til smeltevannandelene $\hat{\pi}_1$, så er disse estimatene mye større enn helikoptermålingene der smeltevannandelen var funnet til ca. 24 %. Området helikopterbildene dekker er mindre enn områdene som $\hat{\pi}_1$ er estimert fra. Dersom man antar at metoden stemmer vil de ukjente pikslene som helikopterbildene ikke dekker bestå av omtrent bare vann. Det er usannsynlig og det kan tyde på at noen av antagelsene for metoden er feil.

En mulighet er at det ikke er sett på en blanding mellom rene piksler av is og smeltesdammer, men blanda piksler. I dette tilfellet vil pikslene bestå av både vann og is. Under forutsetning om at vann har lavere intensitet, vil intensiteten til en blanda piksel fra klasse 1 være høyere enn om denne ikke var blanda. Dette gjør at $\hat{\pi}_1$ ikke representerer smeltevannandelen. Klasse 2 kan også ha det samme problemet, men i dette tilfellet vil intensiteten bli lavere. Et resultat av dette vil være at det blir vanskeligere å skille i to klasser siden μ_1 og μ_2 vil være nærmere hverandre. Dette er illustrert i figur 6.3, og er et resultat av ligning 2.8. Blanda piksler kan også føre til at man får flere klasser (Campbell og Wynne, 2011, kap. 10). Metodene er ikke utviklet for flere enn to klasser og dette vil gi feil siden den antatte modellen ikke er gyldig.



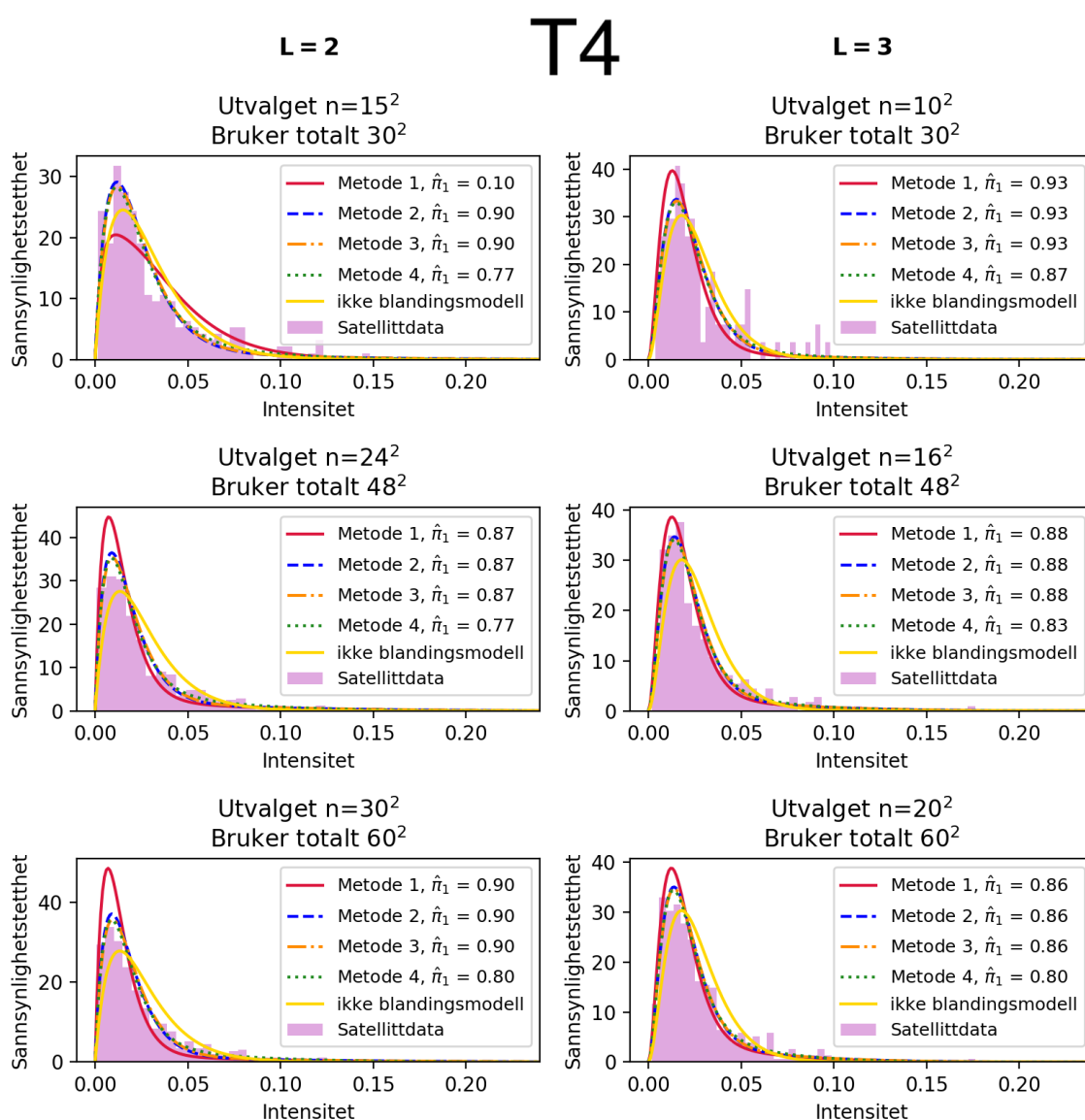
Figur 6.3: Illustrasjonen viser hva som skjer med den observerte μ_1^* og μ_2^* når disse består av en blanding mellom μ_1 og μ_2 .

Estimatene kan også bli påvirket av strukturer på isen, estimatet for L og tidsforskjell mellom bakkesannheten og satellittbildet. Forskjellene mellom estimatet og observasjonene er så stor at tidsforskjellen ikke er en god forklaring alene. Da jeg varierte L innenfor usikkerhetsområdet, så det ikke ut til å ha veldig stor påvirkning for den estimerte andelen smeltevann. Dette er ikke undersøkt i detalj, og det kan ha større innvirkning enn det så ut til.



Figur 6.4: Histogrammene for smeltevannandelen er vist for satellittbildet T3, der jeg ha sett på to forskjellige estimater for kovariansmatrisa. Figurene til venstre bruker fire datapunkter til å estimere kovariansmatrisa, mens de til høyre bruker 9.

Jeg har også sett på histogrammet til T4. Dette er vist i figur 6.5 og er over det samme området som i figur 6.4 (rød markør i bilde 6.2b). Dette datasettet er målt bare noen timer etter helikoptermålingene, noe som gjør at helikopterdataene antagelig reflekterer forholdene bedre for satellittbilde T4 enn for satellittbildet T3. Jeg har brukt de samme størrelsene for å estimere blandingsmodellen med metode 1 til 4, og her varierte $\hat{\pi}_1$ fra 77 % til 90 %. Ut fra dette ser det ikke ut til at forholdet på flaket har endret seg så mye og jeg tviler på at dette er forklaringen for det store avviket fra helikoptermålingene. I figur 6.5 ser det ut til at metodene 2 til 4 gir en blandingsmodell som forklarer satellittobservasjonene ganske godt i forhold til om man bare antar én klasse.



Figur 6.5: Histogrammene for smeltevannandelen er vist for satellittbildet T3, der jeg ha sett på to forskjellige estimater for kovariansmatrisa. Figurene til venstre bruker fire datapunkter til å estimere kovariansmatrisa, mens de til høyre bruker 9.

6.4 Resultater sammenlignet med bakkesannhet

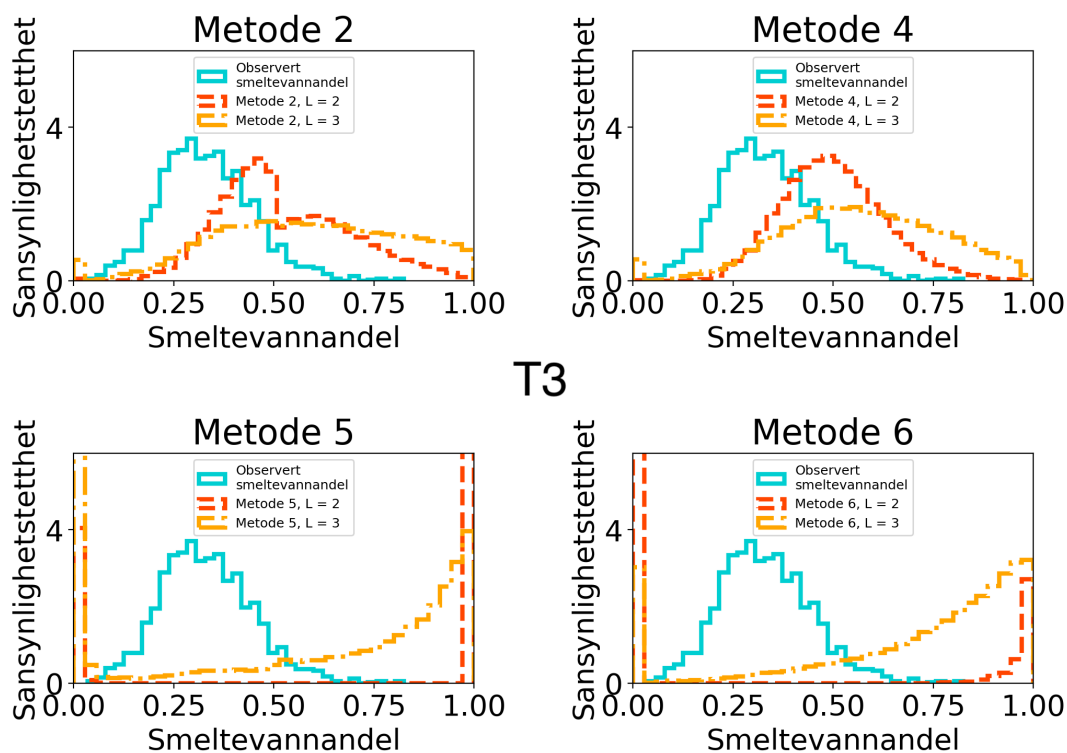
I dette delkapitlet vil jeg gjøre to sammenligninger med bakkesannheten. I den første sammenlignes histogrammet til de observerte smeltevannandelene med histogrammene til estimert smeltevannandel. I den siste sammenligningen vil estimatene til smeltevannandelen sammenlignes direkte med observasjonene. I dette tilfellet dekker helikopterbildene og satellittdataen brukt i metodene omtrent de samme område.

6.4.1 Histogram til metode 2, 4, 5, 6 og bakkesannhet

For de to satellittbildene av isflaket, har jeg for hvert bilde brukt metode 2, 4, 5 og 6 til å estimere andelen smeltedammer. Metode 1 har jeg unngått å bruke, siden denne metoden gjorde mer feil i simuleringstudiene (se kap. 4). Siden metode 2 og 3 ga ganske like resultater i simuleringene, har jeg valgt å bare bruke metode 2. For å se hvordan metodene fungerer har jeg sammenlignet histogrammene over de estimerte smeltevannandelen på samme måte som i Fors mfl. (2017). I denne artikkelen er det sett på områder på 21×21 piksler og 51×51 , der den sistnevnte ga best resultater (Fors mfl., 2017). Jeg har valgt å ikke bruke de samme størrelsene i mine estimater, fordi jeg ønsket å sammenligne metodene ved bruk av to forskjellige estimater for kovariansmatrisa. Det antallet som ligger nærmest 51×51 , er for estimater som totalt bruker 48×48 piksler. Dette gir ca 11,5% færre piksler enn for området på 51×51 piksler. Jeg har valgt å se på området av denne størrelsen videre, siden dette er mer sammenlignbart med det som ble gjort i Fors mfl. (2017).

For å regne ut histogrammet, er det valgt ut to områder godt plassert innpå isflaket. Hvilket område man ser på avhenger av satellittbildet, men for et bestemt bilde vil området være like stort uavhengig av kovariansmatrisa og metoden som er brukt.

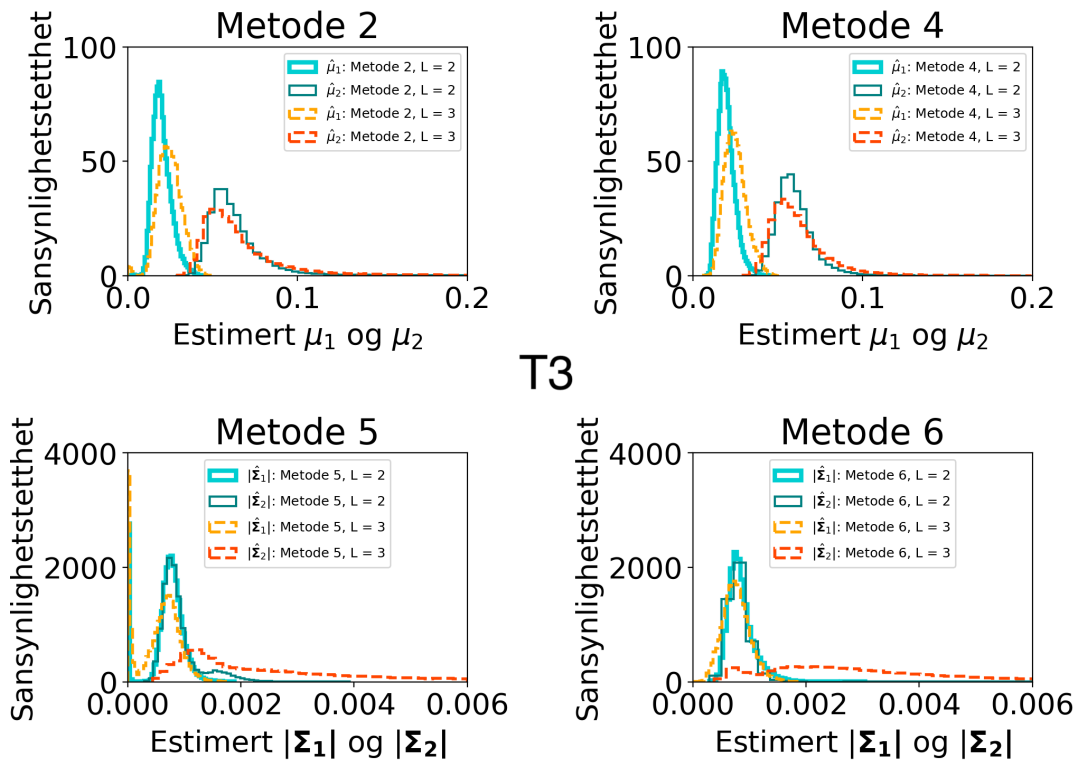
For histogrammene til den estimerte smeltevannandelen $\hat{\pi}_1$ i figur 6.6, ser det ut til at histogrammene er forskjøvet til høyre i forhold til de observerte dataene (bakkesannheten). Denne tendensen ser ut til å gjelde for de fleste estimatene. Sammenlignes metode 2 og 4 med metodene som bruker hele kovariansmatrisa, metode 5 og 6, ser det ut til at metode 2 og 4 gir estimater der histogrammene er nærmere histogrammet til de observerte dataene. Dette gjelder spesielt metode 4 der kovariansmatrisa med $L = 2$ er brukt (se figur 6.6). I dette tilfellet er histogrammet til estimatene ganske lik observasjonene, men er et godt stykke til høyre. For de andre tre estimatene som bruker intensitetsdata ser histogrammene bredere ut, og estimatene som bruker kovariansmatrisa med $L = 3$ er bredest. I disse tilfellene er det flere høye estimater for $\hat{\pi}_1$. Dette gjør at gjennomsnittet til $\hat{\pi}_1$ med $L = 3$ er høyere enn for $L = 2$. Andelen blanda piksler øker ofte når oppløsningen minker og om klasse 1 er blanda piksler, kan dette være med på å forklare den høyere andelen. Ved bruk av metode 5 og 6, der $L = 2$, gir de fleste estimatene $\hat{\pi}_1 = 0$ eller 1. Dette betyr at dataene modelleres fra én klasse. I disse tilfellene er klassene valgt på grunnlag av estimater der blandingsmodellen så ut til å fungere med metode 2 og 4. For metode 5 og 6 er grensen satt til 0.0005, og for metode 1 til 4 er den satt til 0.01650. Dette vil påvirke hvilken klasse estimatoren klassifiserer til. For estimatorene som er sett på, brukes områder på 48×48 piksler, noe som tilsvarer områder på nesten 9000 m^2 i satellittbildet T3. I Tucker mfl. (1999) var det ikke observert dammer på denne størrelsen, noe som tyder på at estimatene nær 1 er for høyt. Bakkesannheten tyder også på at det ikke er så mye smeltedammer.



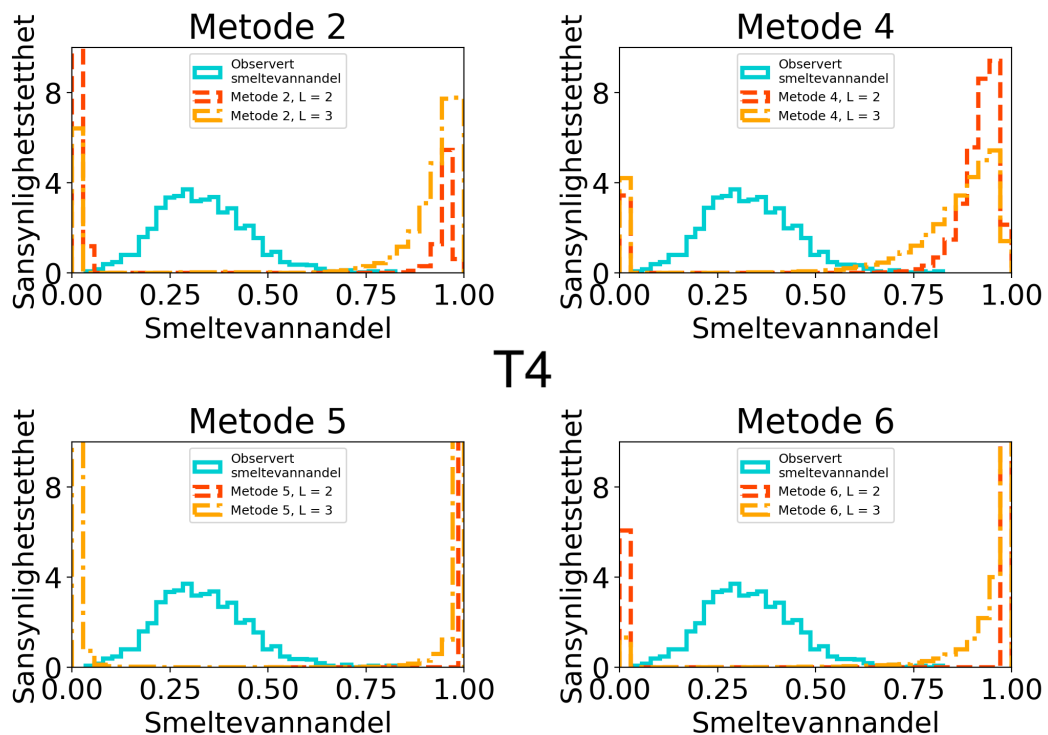
Figur 6.6: Histogram for satellittbilde T3, der histogrammene til estimert smeltevannandel er vist sammen med histogrammet til observert smeltevannandel. Estimaterne er fra metodene 2, 4, 5 og 6, der det er sett på både $L = 2$ og $L = 3$, for utvalg som bruker $n = 48^2$ piksler i estimatene.

For det samme området som er sett på i figur 6.6, kan man også se på hvordan estimatene til μ_1 og μ_2 fordeler seg. Dette er gjort i figur 6.7. For metode 2 og 4 ser det ut til at man får to adskilte grupper, der $\hat{\mu}_1$ er i gruppen til venstre, mens $\hat{\mu}_2$ er i gruppen til høyre. For estimatene som baserer seg på kovariansmatrisa ligger gruppene delvis eller helt oppå hverandre. Dette kan være en grunn til at estimatene av $\hat{\pi}_1$ ser ut til å bli feil. I disse tilfellene er antagelig flere estimater havnet i feil klasse. Estimaterne som bruker kovariansmatrisa det $L = 2$, ser ut til å ha minimal forskjell på estimatene til $\hat{\Sigma}_1$ og $\hat{\Sigma}_2$. Metoden vil i dette tilfellet sette begge estimatene lik hverandre og dataen antas å være fra én kanal. Dette kan være et resultat av at den termiske støyen påvirker dataen i større grad, siden denne vil være tilstede i begge kanalene.

Figur 6.8 viser smeltevannandelen $\hat{\pi}_1$ for satellittbilde T4. I dette tilfellet har alle metodene estimert $\hat{\pi}_1$ nær 0 og 1. Dette vil si at området domineres av den ene klassen og at metoden eventuelt bare ser én klasse. For satellittbilde T4 er en stor andel av den målte tilbakespredningen under støygolvet, noe som gjelder tilbakespredningen fra både is og vann (Fors mfl., 2017). Dette kan forklare hvorfor metodene ser ut til å klassifisere til én klasse, siden støyen gjør at det blir vanskeligere å se forskjell mellom klassene (Wackerman, 1992). Den termiske støyen er heller ikke tatt hensyn til i metoden. For metode 5 og 6 vil antagelig estimatene bli påvirket i større grad på grunn av mye støy i begge kanalene. I figur 6.8 kan der se ut til at metodene 5 og 6 i de fleste tilfellene ikke klassifiserer i to klasser. For satellittbilde T4 er antagelig støynivået så høyt at dette gjør det vanskeligere å bruke metodene som er utviklet.

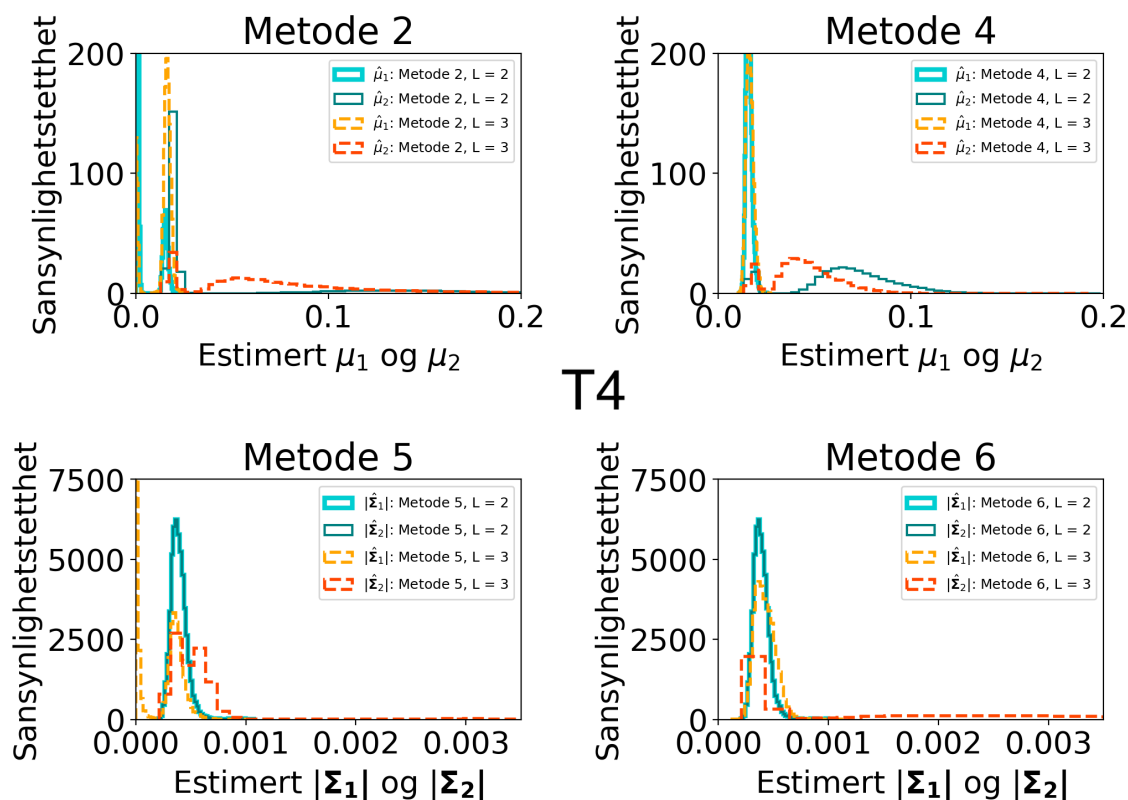


Figur 6.7: Histogram for satellittbilde T3, der histogrammene til estimert $\hat{\mu}_1$ og $\hat{\mu}_2$ er vist. Estimatenes er fra metode 2, 4, 5 og 6, der det er sett på både L = 2 og L = 3, for utvalg som bruker $n = 48^2$ piksler i estimatene. Utvalget er det samme som i figur 6.6.



Figur 6.8: Histogram for satellittbilde T4, der histogrammene til estimert smeltevannandel er vist sammen med histogrammet til observert smeltevannandel. Estimatenes er fra metodene 2, 4, 5 og 6, der det er sett på både L = 2 og L = 3, for utvalg som bruker $n = 48^2$ piksler i estimatene.

Figur 6.9 ser på estimatene til μ_1 og μ_2 . I denne figuren ser det ut til at estimatene μ_1 og μ_2 overlapper en del. Dette er også tilfellet for estimatene til metode 2 og 4, som for satellittbilde T3 så ut til å fungere bedre. Disse to bildene har ikke veldig stor forskjell på oppløsninga, noe som kan tyde på at feilen kommer av at store deler av dataen er under støygolvet. For metode 5 og 6 ser det ut til at metodene har enda mer overlapp. Spesielt ved bruk av kovariansmatrisa med $L=2$ ser histogrammene til $|\hat{\Sigma}_1|$ og $|\hat{\Sigma}_2|$ ut til å være lik.

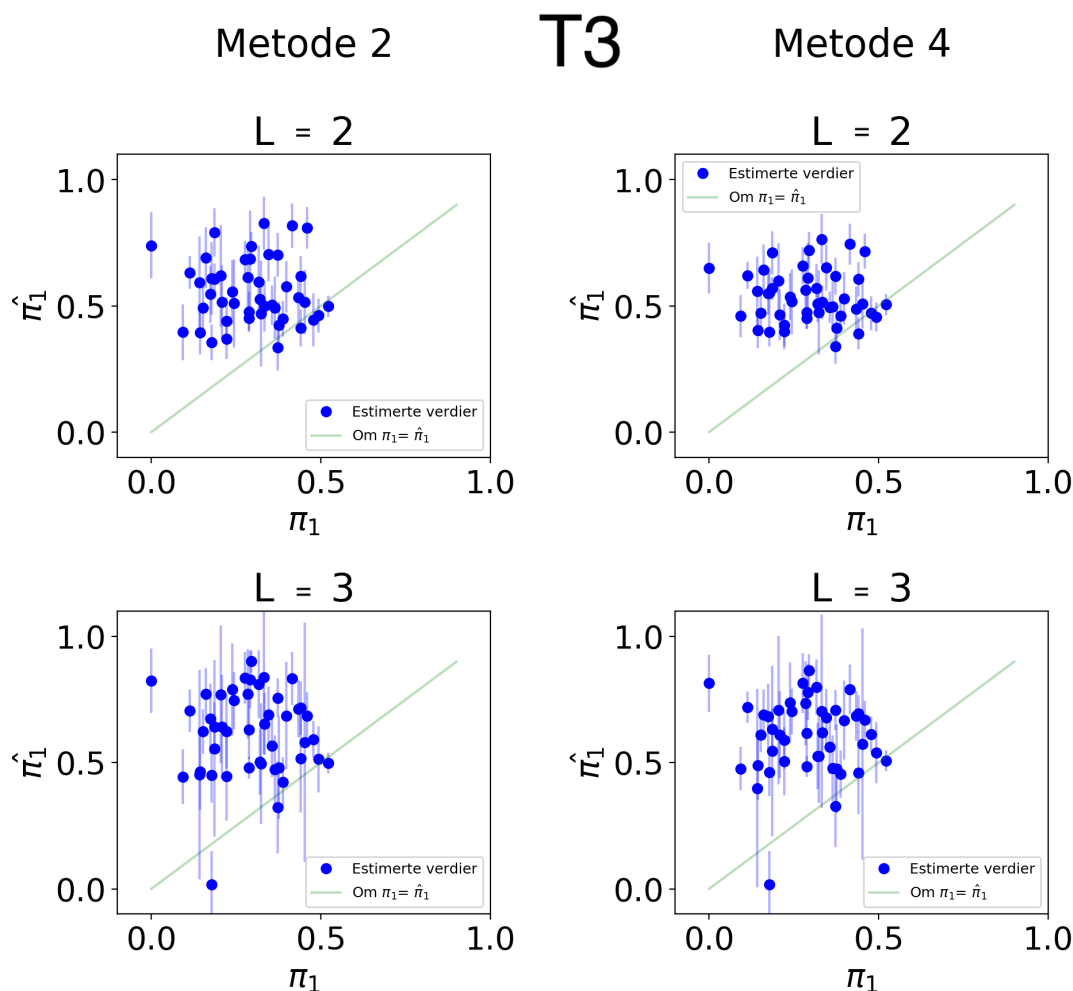


Figur 6.9: Histogram for satellittbilde T4, der histogrammene til estimert $\hat{\mu}_1$ og $\hat{\mu}_2$ er vist. Estimaterne er fra de fire metodene 2, 4, 5 og 6, der det er sett på både $L=2$ og $L=3$, for utvalg som bruker $n=48^2$ piksler i estimatene. Utvalget er det samme som i figur 6.8.

6.4.2 Sammenligner estimater til metode 2 og 4 med bakkesannheten

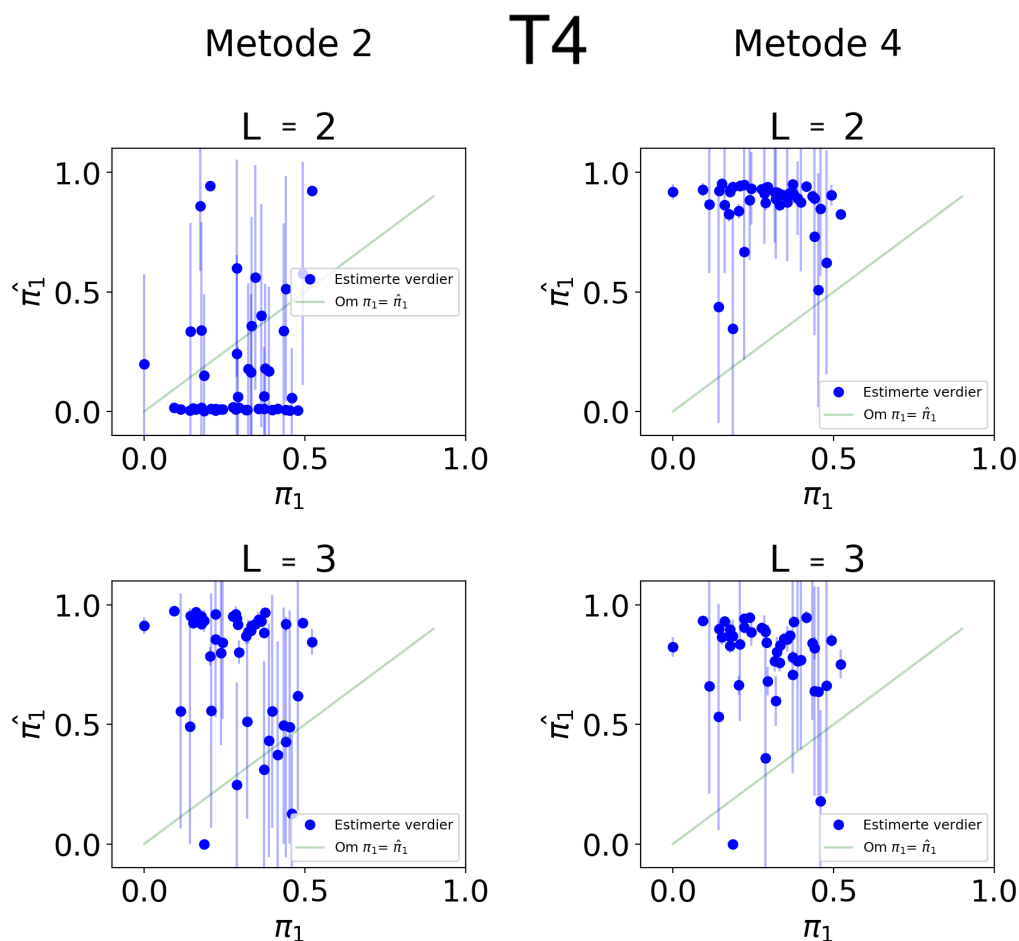
Jeg har videre valgt å fokusere på metode 2 og 4, siden disse metodene så ut til å gi bedre estimater. I figuren 6.10 og 6.11 er metode 2 plassert til venstre og metode 4 er til høyre. Områdene som bakkesannheten er regnet ut fra, dekker flere estimerte verdier for smeltevannandelen $\hat{\pi}_1$. Fra estimatene er gjennomsnittet og standardavviket regnet ut. Dette er vist i figur 6.10 og 6.11. I disse figurene er smeltevannandelen $\hat{\pi}_1$ sammenlignet med den observerte smeltevannandelen π_1 , der linjene viser \pm et standardavvik.

I figur 6.10, som ser på satellittbilde T3, er bakkesannheten langs x-aksen og estimatene er langs y-aksen. Ideelt sett ville estimatene ligget langs den grønne linjen. I dette tilfellet ser det derimot ut til at det ikke er noen sammenheng mellom observasjonene og de estimerte verdiene. Dette gjelder også for metode 4 som brukte estimatet for kovariansmatrisa der $L = 2$. Ut fra figur 6.6 hadde denne metoden ganske likt histogram som de observerte verdiene, men det ser likevel ikke ut til at metoden fungerer bedre. Generelt ser det ut til at estimatene er overestimert, og ingen av metoden i figur 6.10 skiller seg positivt ut.



Figur 6.10: Estimatene til $\hat{\pi}_1$ med metode 2 og 4 mot observert bakkesannhet π_1 . Estimatene bruker kovariansmatrisa med $L = 2$ og $L = 3$, og områder på totalt 48×48 piksler. Verdien for $\hat{\pi}_1$ er gjennomsnittet av de pikslene som bakkesannheten er laget ut fra. De horisontale strekene viser \pm et standardavvik.

For satellittbildet T4 er store deler av de estimerte verdiene rundt 0 og 1 (se figur 6.11). De få verdiene som ligger mellom 0 og 1 har veldig stort standardavvik, noe som tyder på at gjennomsnittet er regnet ut fra estimer som både ligger nær 0 og 1. Heller ikke for satellittbilde T4 ser det ut til å være en sammenheng mellom de estimerte verdiene og observasjonene. I de fleste tilfellene gir estimatene en én klasse modell. Dette var det samme som ble observert i figur 6.6.



Figur 6.11: Estimaten til $\hat{\pi}_1$ med metode 2 og 4 mot observert bakkesannhet π_1 . Estimaten bruker kovariansmatrisa med $L = 2$ og $L = 3$, og områder på totalt 48×48 piksler. Verdien for $\hat{\pi}_1$ er gjennomsnittet av de pikslene som bakkesannheten er laget ut fra. De horisontale strekene viser \pm et standardavvik.

6.5 Sammenligning med Fors mfl. (2017)

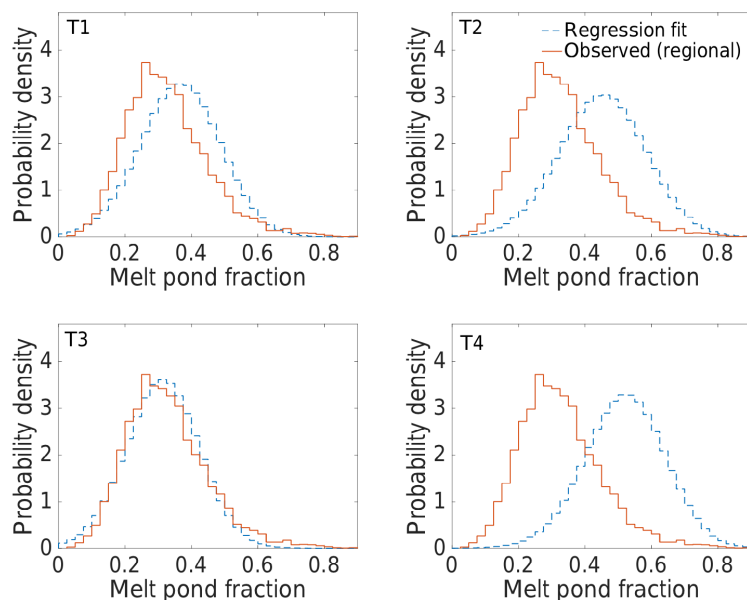
Figurene 6.12 og 6.13 viser resultatene i Fors mfl. (2017) ved bruk av to regresjonsmodeller. Den første metoden er laget med utgangspunkt i satellittbildet T3. Her er vindforholdene slik at estimaten basert på den polarimetrisk egenskapen:

$$R_{VV/HH} = \frac{\langle |S_{VV}|^2 \rangle}{\langle |S_{HH}|^2 \rangle},$$

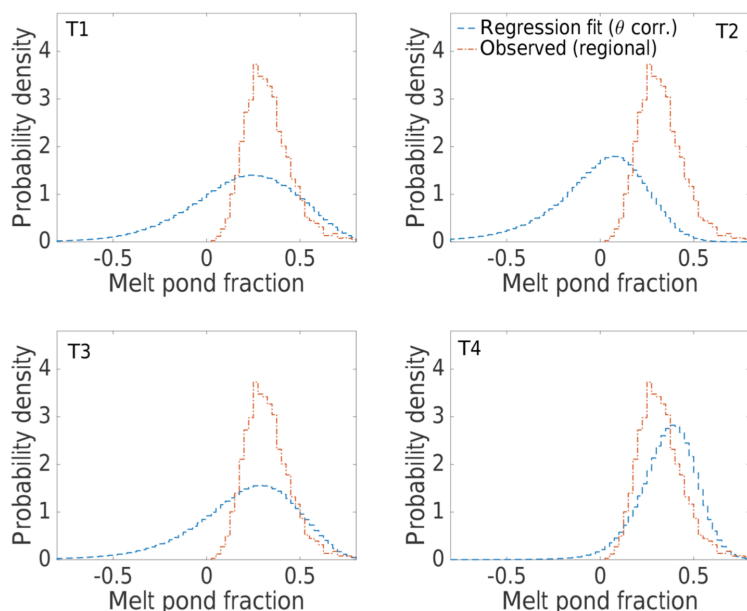
ga høyest korrelasjon med smeltevannandelen (Fors mfl., 2017). For satellittbildet T3 var derimot vinden så svak at den polarimetrisk egenskapen til $\langle |S_{VV}|^2 \rangle$ var bedre egnet.

Resultatene av regresjonsmodellene er vist i figur 6.12 og 6.13. Her er også de to siste satellittbildene T1 og T2 tatt med. Smeltevannandelen er estimert ved bruk av et filteringsvindu på 51×51 piksler. I figurene er histogrammet til estimert smeltevannandel for de fire satellittbildene vist med en blå kurve og histogrammet til observasjonene fra alle de 5 helikopterturene er rød. For satellittbildet T3 i figur 6.12, overlapper histogrammene i veldig stor grad. Det samme gjelder i stor grad satellittbildet T4 i figur 6.13. De andre bildene har mer eller mindre overlapp og grunnen til avvikene for er i følge Fors mfl. (2017)

forskjell på innfallsvinkel og vindforhold. Resultatene for T3 i figur 6.12 og T4 i figur 6.13 ser ut til å samsvare bedre med observasjonene enn metode 2, 4, 5 og 6. For de forholdene som regresjonsmodellen ikke bygger på ser derimot resultatene ut til å bli dårligere. Dette problemet var motivasjonen til å lage en ny metode, men de utvikla metodene løser ikke dette problemet. De nye metodene ser generelt ut til å ha større feil.



Figur 6.12: Resultater fra Fors mfl. (2017). Regresjonsmodellen som er brukt til å estimere smeltevannandelen bruker R_{VV}/HH og satellittbildet T3. Her er alle de 4 satellittbildene i ICE2012 brukt. Estimaten er gjort med et filtreringsvindu på 51×51 . I tillegg vises histogrammet til den observerte smeltevannandelen for alle helikopterturene. (Gjengitt med velvillig tillatelse fra A. S. Fors)



Figur 6.13: Resultater fra Fors mfl. (2017). Regresjonsmodellen som er brukt til å estimere smeltevannandelen bruker $\langle |S_{VV}|^2 \rangle$ og satellittbildet T4. Her er alle de 4 satellittbildene i ICE2012 brukt. Estimaten er gjort med et filtreringsvindu på 51×51 . I tillegg vises histogrammet til den observerte smeltevannandelen for alle helikopterturene. (Gjengitt med velvillig tillatelse fra A. S. Fors)

Kapittel 7

Konklusjon

I denne oppgaven er det utviklet seks metoder som bruker statistiske egenskaper til SAR-data, momentmetoden og blandingsmodell. Metodene er utviklet for å estimere smeltevannandelen i satellittbilder så ut til å fungerer godt i flere simuleringsstudier. Disse simuleringsstudiene er det sett på hvordan estimatene avhenger av flere ulike parametere som blant annet utvalget, L og π_1 . For hvert av estimatene ble det sett på de statistiske egenskapene varians, forventningsskjevhet og den midlede kvadratiske feilen (MSE). Ut fra disse studiene så det ut til at estimatorene var konsistent og fungere godt. Men det oppstod problemer når utvalget var for lite, klassenes var for små og forholdet mellom parametrene μ_1 og μ_2 (metode 1 til 4) eller $|\Sigma_1|$ og $|\Sigma_2|$ (metode 5 og 6) var nær 1. Grensen for når μ_1/μ_2 ble for nær 1 så ut til å avhenge av utvalget, der det for små utvalg oppstod problemene tidligere. For variablene $L = 3$, $\pi_1 = 0,4$ og utvalget $n = 50$ så dette ut til å bli problemer ved $\mu_1/\mu_2 = 0,2$. Endrer man utvalget til $n = 450$ så problemet ut til å starte ved $\mu_1/\mu_2 = 0,4$. Forholdet mellom μ_1 og μ_2 så ut til å ha mer å si enn selve verdiene på disse variablene. Ved økende L så det ut til at alle metodene ga bedre resultater, men at denne forbedringen var størst for L mindre enn 4. For metodene som så på intensitetdata, så skilte metode 1 seg negativt ut. Blant annet måtte utvalget være mye større for å unngå bytter mellom klassene og det var mer feil rundt $\pi_1 = 0,5$ i simuleringen der π_1 ble variert mellom 0 og 1. De andre metodene så ut til å fungere bra der metode 2 og 3 oppførte seg veldig likt.

Ved bruk av reelle data fra satellitten TerraSAR-X så det derimot ikke ut til at metodene fungerte. Her var estimatene generelt for høy. For eksempel estimerte metodene med et utvalg på totalt 24^2 piksler fra satellittbildet T3 over 67 % smeltevannandel for det samme området som bakkesannheten hadde en andel på ca. 24 %. I tillegg var alle histogrammene til den estimerte smeltevannandelen forskjøvet mot en høyere andel i forhold til bakkesannheten. Flere av estimatene som brukte kovariansmatrisa skilte heller ikke dataen i to klasser. Dette var mest utbredt på satellittbildet T4, der nesten alle estimatene var 1 eller 0. Dette er også det bildet med mest støy, der 25 % av støyen er under støygolvet (Fors mfl., 2017). Estimaten fra samme sted som bakkesannheten er også undersøkt nærmere ved sammenligning. I denne sammenligningen ser det ikke ut til at bakkesannheten avhenger av estimatene. Noe av problemet med metodene kan da ha sammenheng med at det er mye støy. Et annet problem med metoden kan være oppløsningen. Er denne for lav, vil det bli en god del blanda piksler, spesielt ved situasjoner der smeltesammene ikke er så store. Dette kan føre til at man får flere klasser og at klassene som blir sett på ikke er rene. Det tyder på at metoden ikke fungerer så godt til å estimere smeltevannandelen.

7.1 Videre arbeid

Siden metodene så ut til å ha problemer på reelle data, kan det utvikles en ny metode. Et forslag kan være å utvikle en metode som estimerer smeltevannandelen direkte på den målte spredningskoeffisienten S . Disse dataene antas å være komplekst binormalfordelt og vil ha høyest mulig romlig oppløsning. Dette vil forhåpentligvis redusere problemet med blanda piksler og effekten dette har på resultatene. Det er også mulig å utvikle en ny modell som bruker data med blanda piksler. I en slik modell vil også π_1 og π_2 følge en fordeling. Denne fordelingen kan for eksempel være betafordelingen, siden denne fordelingen er definert for det samme området som π_1 og π_2 .

Problemet med metoden kan også bestå i at det er flere klasser, både på grunn av blanda piksler og at området som ses på har flere naturlige klasser. For å undersøke denne situasjonen, og for å se hvordan dette påvirker resultatene, kan man gå gjennom noen flere simuleringsstudier. I disse studiene kan man teste metodene på data der man for eksempel har en fordeling som består av tre klasser.

I simuleringsstudiene er L antatt å være kjent. Det er også mulig å se på hvordan estimatene påvirkes når fordelingen har en L og metodene bruker en annen.

Selv om metodene ikke så ut til å fungere for smeltevann på is, kan metodene brukes i andre situasjoner der datasettet består av to klasser. Et eksempel på kan være satellittbildet over skog og gress. Metodene vil kunne fungere på SAR-data, så lenge det er sett på områder med to klasser.

Bibliografi

- Anfinsen, S. N. (2016). *Statistical unmixing of SAR images*. Tekn. rapp. Tromsø, Norway: Munin Open Research Archive, University of Tromsø -- The Arctic University of Norway, s. 1–4. URL: <http://hdl.handle.net/10037/8425>.
- Anfinsen, S. N., A. P. Doulgeris og T. Eltoft (2009). «Estimation of the equivalent number of looks in polarimetric synthetic aperture radar imagery». I: *IEEE Trans. Geosci. Remote Sens* 47.11, s. 3795–3809.
- Anfinsen, S. N. og T. Eltoft (2011). «Application of the matrix-variate Mellin transform to analysis of polarimetric radar images». I: *IEEE Transactions on Geoscience and Remote Sensing* 49.6, s. 2281–2295.
- Anton, H. og C. Rorres (2011). *Elementary Linear Algebra*. 10. utgave. uten sted: John wiley & Sons. ISBN: 978-0-470-56157-7.
- Beckers, J., A. Renner, G. Spreen, S. Gerland og C. Haas (2015). «Sea-ice surface roughness estimates from airborne laser scanner and laser altimeter observations in Fram Strait and north of Svalbard». I: 56, s. 235–244.
- Campbell, J. B. og R. H. Wynne (2011). *Introduction to Remote Sensing*. 5. utgave. New York, NY: The Guilford Press. ISBN: 978-1-60918-176-5.
- Casella, G og R. L. Berger (2002). *Statistical Inference*. 2. utgave. Pacific Grove, CA: Duxbury- Thomson Learning. ISBN: 978-0-495-39187-6.
- Comiso, J. C. (2012). «Large decadal decline of the Arctic multiyear ice cover». I: *Journal of Climate* 25.4, s. 1176–1193.
- Divine, D. V., C. A. Pedersen, T. I. Karlsen, M. A. Granskog, S. R. Hudson, S. Gerland, S. A. Divina og A. H. H. Renner (2015). «Regional melt-pond fraction and albedo of thin Arctic first-year drift ice in late summer». I: *The Cryosphere* 9.1, s. 255–268.
- Divine, D. V, C. A. Pedersen, T. I. Karlsen, H. F Aas, M. A. Granskog, S. R. Hudson, A. H. H. Renner, S. Divina og S. Gerland (2017). *Regional melt-pond and open water fractions of thin Arctic first-year drift ice during ICE12 drift experiment 26.07-03.08.2012 north of Svalbard*. URL: <https://doi.org/10.21334/npolar.2017.5de6b1e4>.

- Elachi, C. og J. Zyl (2006). *Introduction to the physics and techniques of remote sensing*. Hoboken, NJ: Wiley-Interscience. ISBN: 9780471475699.
- Fetterer, F. og N. Untersteiner (1998). «Observations of melt ponds on Arctic sea ice». I: *Journal of geophysical research* 103.C11, s. 24821–24835.
- Fors, A. S., D. V. Divine, A.P. Doulgeris, A. H. H. Renner og S. Gerland (2017). «Signature of Arctic first-year ice melt pond fraction in X-band SAR imagery». I: *The Cryosphere* 11.2.
- Fritz, T. (2007). *TerraSAR-X Ground Segment Level 1b Product Format Specification, Cluster Applied Remote Sensing*. Tekn. rapp. TX-GS-DD-3307, Issue: 1.3. DLR.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer Science & Business Media.
- Goodman, J. W. (1984). Statistical Properties of Laser Speckle Patterns. I: *Laser Speckle and Related Phenomena*. Red. av J. C. Dainty. 2. utgave. Bd. 9. Topics in Applied Physics. Berlin: Springer-Verlag. Kap. 2, s. 9–75.
- Grenfell, T. C. og D. K. Perovich (1984). «Spectral albedos of sea ice and incident solar irradiance in the southern Beaufort Sea». I: *Journal of Geophysical Research: Oceans* 89.C3, s. 3573–3580.
- Holt, B. og S. A. Digby (1985). «Processes and Imagery of First-Year Fast Sea Ice During the Melt Season». I: *Journal of geophysical research* 90.C3, s. 5045–5062.
- JAXA (2016). *PALSAR-2 Level 1.1/2.1/1.5/3.1 CEOS SAR Product Format Description*. Tekn. rapp. Revisjon F.
- Kim, T. H. og H. White (2004). «On more robust estimation of skewness and kurtosis». I: *Finance Research Letters* 1.1, s. 56–73.
- Langleben, M. P. (1971). «Albedo of melting sea ice in the southern Beaufort sea». I: *Journal of Glaciology* 10.58.
- Lindstrøm, T. (2006). *Kalkulus*. 3. utgave. Oslo: Universitetsforlaget. ISBN: 978-82-15-00977-3.
- Mathai, A. M. (1997). *Jacobians of Matrix Transformations and Functions of Matrix Argument*. Singapore: World Scientific Publishing. ISBN: 9810230958.
- MDA (2016). *RADARSAT-2 product description*. Tekn. rapp. RN-SP-52-1238, Issue: 1/13. MDA.
- Oliver, C. og S. Quegan (2004). *Understanding synthetic aperture radar images*. Raleigh, NC: SciTech Publishing. ISBN: 1-891121-31-6.
- Parkinson, C. L. og W. W. Kellogg (1979). «Arctic sea ice decay simulated for a CO₂-induced temperature rise». I: *Climatic Change* 2.2, s. 149–162. DOI: [10.1007/BF00133221](https://doi.org/10.1007/BF00133221).

- Perovich, D. K., W. B. Tucker og K. A. Ligett (2002). «Aerial observations of the evolution of ice surface conditions during summer». I: *Journal of Geophysical Research: Oceans* 107.C10.
- Scharien, R. K., J. J. Yackel, D. G. Barber, M. Asplin og D. Gupta M. Isleifson (2012). «Geophysical controls on C band polarimetric backscatter from melt pond covered Arctic first-year sea ice: Assessment using high-resolution scatterometry». I: *Journal of Geophysical Research: Oceans* 117.C9.
- Stroeve, J. C., T. Markus, L. Boisvert, J. Miller og A. Barrett (2014). «Changes in Arctic melt season and implications for sea ice loss». I: *Geophysical Research Letters* 41.4, s. 1216–1225. ISSN: 1944-8007. DOI: [10.1002/2013GL058951](https://doi.org/10.1002/2013GL058951).
- Tucker, W. B., A. J. Glow, Meese D. A. og H. W. Bosworth (1999). «Physical characteristics of summer sea ice across the Arctic Ocean». I: *Journal of geophysical reserch* 104.C1, s. 1489–1504.
- Upton, G. og I. Cook (2014). *A Dictionary of Statistics*. 3. utgave. Oxford: Oxford University Press. ISBN: 9780191758317.
- Wackerman, C. C. (1992). Digital SAR Image Formation. I: *Microwave Remote Sensing of Sea Ice*. Red. av F. D. Carsey. Washington, DC: The American Geophysical Union. Kap. 6, s. 105–110.
- Walploe, R. E., R. H. Myers, S. L. Myers og K. Ye (2014). *Probability & Statistics for Engineers & Scientists*. 9. utgave. Harlow: Pearson. ISBN: 978-1-29202-392-2.
- Yackel, J. J. og D. G. Barber (2000). «Melt ponds on sea ice in the Canadian Archipelago 2. On the use of RADARSAT-1 synthetic aperture radar for geophysical inversion». I: *Journal of geophysical reserch* 105.C9, s. 22061–22070.

Forkortelser

EM-ståling: Elektromagnetisk stråling

MSE: Mean squared error eller midla kvadratisk feil

SAR: syntetisk aparturradar

Figurer

2.1	W_a måler variasjon innad i egen klasse, mens B_a måler variasjon mellom klassene. . . .	19
3.1	Funksjonen $f(\pi_2)$, for intensitetsdata	28
3.2	Funksjonen $f(\pi_2)$, for kovariansmatriser med dimensjonene $d = 2$ og $d = 3$	32
4.1	Egenskaper til $\hat{\pi}_1$ ved varierende L	34
4.2	Feil på grunn av bytter når L varierer	35
4.3	Egenskaper til $\hat{\mu}_1$ ved varierende L	36
4.4	Egenskaper til $\hat{\mu}_2$ ved varierende L	36
4.5	Egenskaper til $\hat{\pi}_1$ ved varierende π_1	37
4.6	Egenskaper til $\hat{\pi}_1$ ved varierende π_1	38
4.7	Feil på grunn av bytter når π_1 varierer	39
4.8	Egenskaper til $\hat{\mu}_1$ ved varierende π_1	40
4.9	Egenskaper til $\hat{\mu}_2$ ved varierende π_1	40
4.10	Egenskaper til $\hat{\pi}_1$ ved varierende n	41
4.11	Sammenligning av histogrammene til metode 1 og 3	42
4.12	Feil på grunn av bytter når n varierer	42
4.13	Egenskaper til $\hat{\mu}_1$ ved varierende n	43
4.14	Egenskaper til $\hat{\mu}_2$ ved varierende n	43
4.15	Egenskaper til $\hat{\pi}_1$ ved varierende μ_1 og μ_2	44
4.16	Prosentandel klassifisert til én klasse når μ_1/μ_2 varierer	45
4.17	Feil på grunn av bytter når μ_1 og μ_2 varierer	45
4.18	To tildeller av π_1 når μ_1 og μ_2 variere	46
4.19	To tildeller av n når μ_1 og μ_2 variere	46
4.20	Egenskaper til $\hat{\mu}_1$ ved varierende μ_1 og μ_2	47
4.21	Egenskaper til $\hat{\mu}_2$ ved varierende μ_1 og μ_2	47
5.1	Egenskaper til $\hat{\pi}_1$ ved varierende L	51
5.2	Egenskaper til $ \hat{\Sigma}_1 $ ved varierende L	51
5.3	Egenskaper til $ \hat{\Sigma}_2 $ ved varierende L	52
5.4	Egenskaper til $\hat{\pi}_1$ i intervlalet $0,1 < \pi_1 < 0,9$	53
5.5	Feil på grunn av bytter når π_1 varierer	54
5.6	Egenskaper til $ \hat{\Sigma}_1 $ i intervlalet $0,1 < \pi_1 < 0,9$	54
5.7	Egenskaper til $ \hat{\Sigma}_2 $ i intervlalet $0,1 < \pi_1 < 0,9$	55
5.8	Egenskaper til $\hat{\pi}_1$ ved varierende n	56
5.9	Feil på grunn av bytter når n varierer	56
5.10	Egenskaper til $ \hat{\Sigma}_1 $ ved varierende n	57

5.11	Egenskaper til $ \hat{\Sigma}_2 $ ved varierende n	57
5.12	Egenskaper til $\hat{\pi}_1$ ved varierende $ \Sigma_1 / \Sigma_2 $	58
5.13	Prosentandel av utvalgene som klassifisert til én klasse når $ \Sigma_1 / \Sigma_2 $ varierer	59
5.14	Egenskaper til $ \hat{\Sigma}_1 $ ved varierende $ \Sigma_1 / \Sigma_2 $	59
5.15	Egenskaper til $ \hat{\Sigma}_2 $ ved varierende $ \Sigma_1 / \Sigma_2 $	60
5.16	Feil på grunn av bytter når $ \Sigma_1 / \Sigma_2 $ varierer	60
5.17	Endring av $ \Sigma_1 $ og $ \Sigma_2 $ når ρ_1 og ρ_2 varierer	62
5.18	$ \Sigma_1 / \Sigma_2 $ når ρ_1 og ρ_2 varierer	62
5.19	Egenskaper til $\hat{\pi}_1$ ved varierende korrelasjon	63
5.20	Egenskaper til $ \hat{\Sigma}_1 $ ved varierende korrelasjon	64
5.21	Egenskaper til $ \hat{\Sigma}_2 $ ved varierende korrelasjon	65
6.1	Overrsikt over datasettet fra TerraSAR-X og helikopter	67
6.2	Satellittbildene T3 og T4	69
6.3	Virkningen på μ_i av å se på blanda klasser i stedet for rene	71
6.4	Sammenligning mellom estimert histogram og data fra satellittbilde T3	71
6.5	Sammenligning mellom estimert histogram og data fra satellittbilde T4	72
6.6	Hilistogram til estimert smeltedamandel i satellittbilde T3	74
6.7	Estimat $\hat{\mu}_1$ og $\hat{\mu}_2$ for satellittbildet T3	75
6.8	Hilistogram til estimert smeltedamandel i satellittbilde T4	75
6.9	Estimat $\hat{\mu}_1$ og $\hat{\mu}_2$ for satellittbildet T4	76
6.10	sammenligner resultater for T3 med observasjoner	77
6.11	sammenligner resultater for T4 med observasjoner	78
6.12	Resultater fra Fors mfl. (2017), regresjonsmodell basert på $R_{VV/HH}$	79
6.13	Resultater fra Fors mfl. (2017), regresjonsmodell basert på $\langle S_{VV} ^2 \rangle$	79

Tabeller

2.1	Utregnede verdier av μ , σ^2 , γ og κ for gammafunksjonen	17
2.2	Utregnede verdier av μ , σ^2 , γ og κ for kompleks Wishart-fordeling	18
4.1	Oppsummering av metode 1 til 4	33
4.2	Variabler i simmuleringstudien: varierende L	33
4.3	Variabler i simmuleringstudien: varierende π_1	37
4.4	Variabler i simmuleringstudien: varierende n	41
4.5	Variabler i simmuleringstudien: varierende μ_1/μ_2	44
5.1	Oppsummering av metode 5 og 6	49
5.2	Variabler i simuleringsstudien: varierende L	50
5.3	Variabler i simuleringsstudien: varierende π_1	52
5.4	Variabler i simuleringsstudien: varierende utvalg n	55
5.5	Variabler i simuleringsstudien: varierende $ \Sigma_1 / \Sigma_2 $	58
5.6	Variabler i simuleringsstudien: varierende kovarians	61
6.1	Oppsummering av satellittdata for T3 og T4(Fors mfl., 2017).	68