Norges miljø- og
biovitenskapelige
universitet

Masteroppgave 2016 30 stp.
Institutt for matematiske realfag og teknologi

Eksplorativ analyse av EPR-spektre av alanin og Gorilla® Glass

Explorative Analysis of EPR Spectra of Alanine and Gorilla® Glass

Eirik Ogner Jåstad
Miljøfysikk og fornybar energi

Forord

Denne masteroppgaven markerer avslutningen av mitt studie i Miljøfysikk og fornybar energi ved Norges miljø- og biovitenskapelige universitet. Oppgaven har et omfang på 30 studiepoeng og har blitt skrevet våren 2016.

Jeg vil takke alle som har hjulpet meg i arbeidet med denne oppgaven og gjennom studietiden.

Takk til Cecilia Marie Futsæther som har vært min hovedveileder, hun har gitt solid støtte og hjelp med arbeidet som ligger bak denne oppgaven. Hun har gjort en fantastisk innsats som veileder og har vært veldig flink til å komme med innspill i arbeidet, og bidratt til korrektur av oppgaven.

Takk til Knut Kvaal og Turid K Gjerstad Torheim for mange gode råd og hjelp på veien.

Takk til Eirik Malinen, Einar Sagstuen og Eli Olaug Hole ved Universitetet i Oslo for å skaffe og bearbeide datasettene, og for å ha introdusert meg for EPR-spektroskopi.

Takk til Kristin for støtte og tålmodighet med meg denne våren.

Ås, 12.05.16

_____

Eirik Ogner Jåstad

Sammendrag

Elektron Paramagnetisk Resonans (EPR) spektroskopi er en måleteknikk som tar opp spektre som kan brukes til å estimere absorbert stråledose, såkalt EPR-dosimetri. EPR-dosimetri måler mengden frie radikaler i et materiale, som er proporsjonalt med den absorberte dosen. De fleste frie radikaler er kortlivede ved romtemperatur, dette er en utfordring ved EPR-dosimetri. Et materiale som ofte brukes i dosimetri er aminosyren L-α-alanin. Alanin egner seg for planlagte eksponeringer, men ikke nødvendigvis ved uforutsette eksponeringer. Et materiale som kan egne seg til dette er Gorilla® Glass, som brukes i dagens berøringsskjermer.

I denne oppgaven studeres EPR-spektre av bestrålt alanin og Gorilla® Glass. Alanin datasettet inneholder EPR-spektre av 19 bestrålte alaninprøver utsatt for varierende temperatur og oppvarmingstid. Det antas at minst tre radikaler (R1, R2 og R3) dannes og at disse påvirkes ulikt av oppvarming. Hensikten med denne oppgaven har vært å identifisere EPR-spektrene til disse tre radikalene ved hjelp av flere statistiske metoder for dekomponering av EPR-spektrene i ladninger og skårer. Ladningene fra de ulike metodene har blitt analysert med mål å finne ut om de ligner på de teoretiske radikalspektrene som er funnet fra målinger og kvantemekaniske simuleringer.

Metodene flervariabel kurveoppløsning (MCR), faktoranalyse (MLCFA) og selvmodellerings miksturanalyse (SMA) gir estimater av radikalspektre som til en viss grad ligner på de teoretiske spektrene og kan dermed egne seg til å estimere formen til radikalspektrene. Derimot klarer ikke metodene prinsipalkomponent analyse (PCA) og uavhengig komponentanalyse (ICA) å identifisere alle tre radikalspektrene. EPR-spektrene til alanin har blitt preprosessert med flere teknikker uten at det har gitt stor påvirkning på de identifiserte komponentene.

Det er vist at andelen til R1-radikalet avtar, mens andelen R2- og R3-radikaler øker ved lengre oppvarmingstid. De estimerte prosentandelene stemmer ikke med tidligere studier.

Gorilla® Glass datasettet består av EPR-spektre av ni prøver av Gorilla® Glass bestrålt med ulike doser fra 0-10 Gy. Hensikten med denne oppgaven har vært å lage regresjonsmodeller basert på EPR-spektre for å estimere doser gitt til berøringsskjermer, samt å lage klassifiseringsmodeller for å dele prøvene inn i grupper basert på doser.

Prinsipalkomponentregresjon (PCR) og delvis minste kvadraters metode (PLS) gir nesten identiske regresjonsmodeller og begge fungerer delvis til å estimere doser ved utelatt en kryss-

validering (LOOCV). Minste kvadraters metode overtilpasser prøvene og bør derfor ikke brukes.

Variabelseleksjon utført ved lasso, intervall PLS (IPLS) og variansanalyse for å velge variablene som gir best doseestimeringer, førte til feilestimering av dosene.

Ved kryssvalidering klarte lineær diskriminant analyse (LDA) med Mahalanobis distanse å estimere prøvenes klassetilhørighet i en lav- eller høydosegruppe med en nøyaktighet på 100%, mens PLS diskriminant analyse (PLSDA) og K-gjennomsnitt (K-means) kun klarte å estimere klassene med en nøyaktighet rundt 50 %.

Abstract

Electron Paramagnetic Resonance (EPR) is a spectroscopic technique, which records spectra suited to estimate radiation doses, so-called EPR dosimetry. EPR dosimetry measures the relative number of free radicals in a material, which is proportional to the absorbed dose. Most of the free radicals are short-lived at room temperature. This is a challenge for EPR dosimetry. A material often used in dosimetry is the amino acid L-α-alanine. Alanine is suitable for planned exposures, but not necessarily for accidental exposures. One materials, which might be suited as a dosimeter in accidents is Gorilla® Glass, which is used in touch screen.

This Master thesis investigates EPR spectra of irradiated alanine and Gorilla® Glass samples. The alanine data set contains EPR spectra of 19 irradiated alanine samples exposed to different temperatures and heating time. It has been assumed that at least three radicals (R1, R2 and R3) occur in irradiated alanine and that they react differently to heating. The purpose of this thesis is to identify the EPR spectra of the radicals, using several statistical methods for decomposing alanine spectra to scores and loadings. The loadings from the methods have been analysed with the goal of identifying and comparing the loadings to the theoretical radical derived from measurements and quantum-mechanical simulations.

The methods Multivariate Curve Resolution (MCR), Maximum Likelihood Common Factor Analysis (MLCFA) and Self-modelling Mixture Analysis (SMA) gave estimates of the radicals' spectra similar to the theoretical spectra, and may therefore be suitable methods for estimating the R1, R2 and R3 spectra. On the other hand, the methods Principal Component Analysis (PCA) and Independent Component Analysis (ICA) did not identify the component spectra. The EPR spectra of alanine were preprocessed with several techniques without giving any differences in the identified components.

The fraction of the R1 radical was found to decrease and the fraction of radicals R2 and R3 were found to increase with increasing heating time, but the estimated percentages did not agree with the literature.

The Gorilla® Glass data set contained EPR spectra from nine samples of Gorilla® Glass irradiated with doses in the interval 0-10 Gy. The aim in this thesis was to make regression models based on EPR spectra with goal of predicting doses given to touch screens and to make classifications models to separate samples into groups based on absorbed doses.

Principal Component Regression (PCR) and Partial Least Squares (PLS) gave almost identical regression model. Both are therefore partially suited to estimate doses with leave-one-out cross-validation. The least squares method overfitted the samples and should therefore not be used.

The variable selection methods lasso, interval PLS (IPLS) and variable variance were tested, to find the best variables to predict doses. However, the doses were often incorrectly predicted.

For cross-validation, Linear Discriminant Analysis (LDA) using the Mahalanobis distance classified the samples into high dose and low dose groups with 100 % accuracy. Partial Least Squares for Discrimination Analysis (PLSDA) and the K-means algorithm had a classification accuracy of around 50 %.

Forkortelser

ALS	Alternating Least Squares	Alternerende minste kvadraters metode
CV	Cross-Validation	Kryssvalidering
EFA	Evolving Factor Analysis	Utviklende faktoranalyse
EMSC	Extended Multiplicative Scatter Correction	Utvidet multiplikativ signal korreksjon
EPR	Electron Paramagnetic Resonance	Elektron paramagnetisk resonans
ESR	Electron Spin Resonance	Elektron spinn resonans
ICA	Independent Component Analysis	Uavhengig komponent analyse
IPLS	Interval Partial Least Squares	Intervall delvis minste kvadraters metode
LASSO	Least Absolute Shrinkage and Selection Operator	Minste absolutte krympende og selekterende operator
LD	Lethal Dose	Dødelig dose
LDA	Linear Discriminant Analysis	Lineær diskriminant analyse
LOF	Lack Off Fit	Manglende tilpasning
LOOCV	Leave-One-Out Cross-Validation	Utelat en kryssvalidering
MCR	Multivariate Curve Resolution	Flervariabel kurveoppløsning
MLCFA	Maximum Likelihood Common Factor Analysis	Maksimum sannsynlighet felles faktoranalyse
MSC	Multiplicative Scatter Correction	Multiplikativ signal korreksjon
NIPLS	Nonlinear Iterative Partial Least Squares	Ikke lineær iterativ delvis minste kvadraters metode
NMR	Nuclear Magnetic Resonance	Nukleær magnetisk resonans
PC	Principal Component	Prinsipalkomponent
PCA	Principal Component Analysis	Prinsipalkomponent analyse
PCR	Principal Component Regression	Prinsipalkomponentregresjon

PLS	Partial Least Squares	Delvis minste kvadraters metode
PLSDA	Partial Least Squares for Discrimination Analysis	Delvis minste kvadraters metode for diskriminant analyse
RMSE	Root Mean Square Error	Kvadratroten av den gjennomsnittlige feilen
RMSEC	Root-Mean-Square Error of Calibration	Kvadratroten av den gjennomsnittlige feilen til kalibrering
RMSECV	Root-Mean-Square Error of Cross Validation	Kvadratroten av den gjennomsnittlige feilen til kryssvalidering
RMSEP	Root-Mean-Square Error of Prediction	Kvadratroten av den gjennomsnittlige feilen til estimering
SMA	Self-modelling Mixture Analysis	Selvmodellerings miksturanalyse
SVD	Singular Value Decomposition	Singulær verdi dekomponering

Innholdsfortegnelse

Forord 1
Sammendrag 3
Abstract 5
Forkortelser 7
Innholdsfortegnelse 9
1 Innledning 13
2 Teori 15
2.1 Elektron paramagnetisk resonans 15
2.2 Ioniserende stråling og stråledoser 17
2.2.1 Strålingstyper 17
2.2.2 Stråledoser 18
2.3 Frie radikaler 19
2.4 Aminosyren alanin 20
2.4.1 Radikaldannelse i alanin 20
2.4.2 Temperaturavhengighet til alanin radikaler 22
2.5 Dosimeter for uforutsette hendelser 23
3 Materialer og metoder 25
3.1 Alanin datasettet 25
3.2 Gorilla® Glass datasettet 27
3.3 Notasjon 28
3.4 Preprosesseringstekniker 28
3.5 Prinsipalkomponent analyse (PCA) 30
3.5.1 Ulike typer figurer 32
3.5.2 Fordeler med prinsipalkomponent analyse 34
3.5.3 Prinsipalkomponent regresjon (PCR) 34
3.6 Dekomponering av spektre 35
3.6.1 Multivariat kurveoppløsning (MCR) 35
3.6.2 Selvmodellerings miksturanalyse (SMA) 40
3.6.3 Faktoranalyse (MLCFA) 42
3.6.4 Uavhengig komponent analyse (ICA) 43
3.7 Regresjon 45
3.7.1 Minste kvadraters metode 45
3.7.2 Lasso 46
3.7.3 Delvis minste kvadraters metode (PLS) 47

3.8 Klassifisering 48
3.8.1 Avstandsmål 49
3.8.2 Lineær diskriminant analyse (LDA) 49
3.8.3 Delvis minste kvadraters metode for diskriminant analyse (PLSDA) 50
3.8.4 K-gjennomsnitt (K-means) 50
3.9 Metodevalidering 51
3.10 Programvare 56
4 Resultater 57
4.1 Analyser av alanin datasettet 57
4.1.1 Minste kvadraters tilpasning av alanin spektrene 57
4.1.2 Prinsipalkomponent analyse (PCA) 60
4.1.3 Faktoranalyse (MLCFA) 67
4.1.4 Flervariabel kurveoppløsning (MCR) 70
4.1.5 Selvmodellerings miksturanalyse (SMA) 82
4.1.6 Uavhengig komponent analyse (ICA) 85
4.1.7 Andeler av de ulike komponentene 86
4.1.8 Estimater av R3. 88
4.1.9 Korrelasjon mellom de teoretiske og estimerte spektrene 90
4.1.10 Egenverdier 91
4.1.11 Sammenligning mellom målte og teoretiske EPR-spektre 92
4.2 Analyser av Gorilla® Glass datasettene 93
4.2.1 Prinsipalkomponent analyse (PCA) 93
4.2.2 Regresjonsmodeller for dosebestemmelse 94
4.2.3 Delvis minste kvadraters metode (PLS) 99
4.2.4 Regresjon med IPLS-variabler 101
4.2.5 Variabelreduksjon 106
4.2.6 Klassifiseringsmodeller 112
5 Diskusjon 117
5.1 Formål 117
5.2 Validering 117
5.3 Alanin datasettet 118
5.3.1 Om datasettet 118
5.3.2 Antall radikalkomponenter 118
5.3.3 Estimerte radikalspektre 119
5.3.4 Mengder og andeler av radikalkomponentene 123

5.4 Gorilla® Glass datasettene 124
5.4.1 Om datasettene 124
5.4.2 Regresjonsmodeller for doseestimering 125
5.4.3 Klassifisering ut fra lav eller høy absorbert dose 129
5.5 Videre arbeid 130
6 Konklusjon 131
7 Referanser 133
8 Vedlegg 137
8.1 Vedlegg 1: Eksempler på preprosesseringer 137
8.2 Vedlegg 2: Andeler av R1, R2 og R3* fra residualanalysene 138
8.3 Vedlegg 3: MCR betingelser 140
8.4 Vedlegg 4: Ladninger, skårer og residualer fra MCR 142
8.5 Vedlegg 5: Andeler av R1*, R2* og R3* i målespektrene 145
8.6 Vedlegg 6: Glatting av R3* 149
8.7 Vedlegg 7: PCA på Gorilla® Glass datasettene 151
8.8 Vedlegg 8: PCR analyser 153
8.9 Vedlegg 9: PLS analyser 154
8.10 Vedlegg 10: K-gjennomsnittsanalyse 155

1 Innledning

Bruk av ioniserende strålingskilder har blitt mer og mer vanlig innenfor en rekke områder, som for eksempel, bestråling av mat for å drepe bakterier, sterilisering av medisinsk utsyr og kreftbehandling [1]. Siden det er mange kilder for ioniserende stråling er det viktig å kunne bestemme doser som blir levert, både til objektet som blir bestrålt og personalet ansvarlig for bestrålingen. Ioniserende stråling kan, for eksempel, øke risikoen for å utvikle kreft [2]. Det er vanlig å dele eksponeringen av ioniserende stråling inn i to grupper. Den ene gruppen er kontrollert eksponering som er planlagt, for eksempel, bestråling av en kreftsvulst. Den andre grupper er uforutsett eksponering som er all ikke planlagt eksponering. Siden det finnes mange bruksområder for ioniserende stråling, er det en vis fare for at det kan skje en ulykke eller at radioaktivt materiale skal komme på avveie, et senario er at radioaktivt materiale kan i fremtiden kommer i hendende på terrorister, og bli brukt i en kjernefysisk eksplosjon eller i en «dirty bomb» [3]. Hvis dette skulle skje vil mange mennesker kunne bli utsatt for radioaktiv stråling i varierende dose. Ved kontrollert eksponering er det god kontroll på absorbert dose, mens det er vanskelig å beregne absorbert dose ved uforutsett eksponering og følgelig vanskelig å vite om lindrende behandling mot stråleskader er nødvendig eller ikke.

Elektron Paramagnetisk Resonans (EPR) spektroskopi er en måleteknikk som tar opp spektre som kan brukes til å estimere absorberte doser, såkalt EPR-dosimetri. EPR-dosimetri måler antallet frie radikaler i et materiale [4]. For at EPR skal kunne brukes til doseberegninger må det gjøres en antagelse om at antallet frie radikaler i en prøve øker proporsjonalt med dosen [5]. De fleste frie radikaler er kortlivede ved romtemperatur. Dette er et problem ved EPR-dosimetri [6]. Derfor er det siden 1960-tallet blitt forsket på aminosyren L-α-alanin (heretter kun alanin) som et materiale som egner seg til å estimere doser absorbert av biologisk materiale [7]. Det er kjent at det dannes flere stabile frie radikaler i bestrålt alanin [8]. Disse radikalene kan detekteres ved hjelp av EPR-spektroskopi [6]. Ioniserende stråling induserer minst tre typer frie radikaler i alanin som er stabile nok til at de kan måles ved hjelp av EPR [9].

Alanin egner seg til dosimetri ved kontrollert eksponering og eksponering på plasser med høy risiko for eksponering, sånn som for personalet som arbeider med stråling. Det er vanskeligere å beregne doser ved uforutsette eksponeringer, siden det ikke er vanlig å gå med dosimetre. For å bestemme dosen og eventuelt hvem som kan dra nytte av lindrende behandling ved en uforutsett eksponering, kan Gorilla® Glass være et materiale som kan brukes som dosimetre [10]. Gorilla® Glass brukes i moderne smarttelefoner og er dermed en av vår tids mest allestedsnærværende gjenstander. Siden smarttelefonen ofte befinner seg i lommen eller

vesken, vil den motta omtrent samme mengde stråling som eieren. Dosimetri basert på Gorilla® Glass er derfor en mulighet, siden det dannes stabile frie radikaler i glasset som i ettertid kan måles i et EPR-spektrometer [10].

Hovedformålet har vært å bestemme bidragene fra de tre antatt stabile radikalene (R1, R2 og R3) til det observerte EPR-spekteret til bestrålt alanin. I denne oppgaven, studeres EPR-spektre av bestrålt alanin. Disse radikalspektrene overlapper EPR [11], og det er derfor ikke mulig å måle spektrene til R1, R2 og R3 direkte og hver for seg. Siden det er størst usikkerhet knyttet til EPR-spektrene av denne tredje komposten har et av målene i oppgaven vært å bestemme denne. Spektrene til R1, R2 og R3 kan bestemmes enten ved simuleringer basert på kvantefysikk [9] eller fra statistiske metoder som brukt i dette arbeidet. Flere forskjellige statistiske modeller har blitt testet basert på dekomponering av spektre, med sikte på bestemme om de egner seg til bruk på EPR-målinger av alanin. I tillegg, er EPR-spektre av bestrålt Gorilla® Glass undersøkt for å bestemme om det er mulig å beregne dosen mobilglasset har blitt utsatt for ved hjelp av EPR-målinger. Videre studeres det om det er mulig å klassifisere prøvene inn i klasser basert på høy eller lavdose, avhengig om dosen er så høy at lindrende behandling er nødvendig eller ikke.

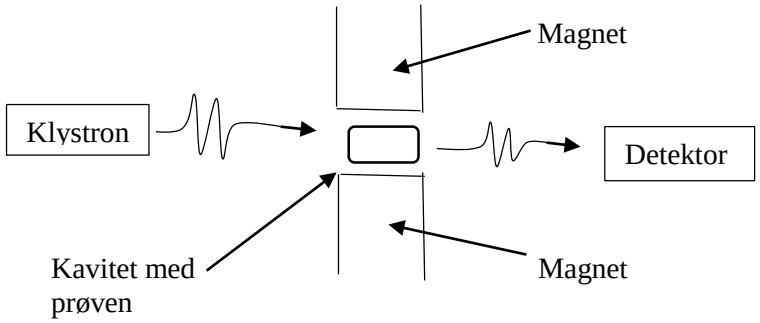
Denne oppgaven er bygget opp ved at den bakenforliggende teorien til EPR, alanin og Gorilla® Glass er forklart i kapittel 2. De statistiske metodene som er benyttet for å dekomponere spektrene og finne regresjonsmodeller er forklart i kapittel 3. Kapittel 4 tar for seg resultatene og kapittel 4.2.7 diskuterer resultatene, før det oppsummeres og konkluderes i kapittel 6.

2 Teori

2.1 Elektron paramagnetisk resonans

Elektron paramagnetisk resonans (EPR) også kalt Elektron spinn resonans (ESR) er en målemetode for å studere radikaler i materie [5]. EPR ligner mye på Nuclear Magnetic Resonanece (NMR) [12]. Forskjellen mellom NMR og EPR er at signalene som EPR registrer kommer fra det magnetiske momentet til elektronene [12], mens i NMR er det magnetiske momentene til kjernene av betydning.

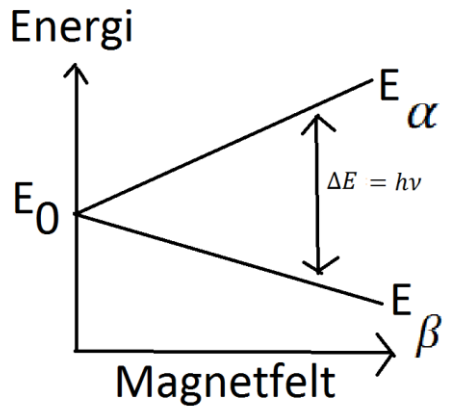
Figur 2-1 viser oppbygningen av et EPR-spektrometer. En prøve av et paramagnetisk substans blir plassert inn i hulrommet (kavitet). Et paramagnetisk substans er et stoff som har minst et uparet elektron [5]. Elektromagneten som er plassert rundt kaviteten setter opp et eksternt magnetfelt [13]. Klystronen sender ut mikrobølger som blir sendt igjennom kaviteten [13]. Noen av disse mikrobølgene blir absorbert i prøven, resten går igjennom og treffer detektoren. Spinnet til hvert elektron i prøven gir opphav til et magnetisk dipolmoment. Spin oppstår som regel i par med motsatt retning, slik at netto magnetisk dipolmoment er null [12]. For at en EPR-måling skal være mulig må det være minst et uparet elektron til stede [12], og er dermed netto magnetisk dipolmoment ulikt null.



Figur 2-1, forenklet skisse av et EPR-spektrometer, magnetene setter opp et magnetfelt og klystronen setter opp mikrobølger, som sendes inn mot prøven

Et elektron kan ha, enten spinn +1/2 (α) eller -1/2 (β) [12]. Dette utnytter EPR. I utgangspunktet er retningene til spinnene tilfeldig orientert. Når prøven derimot settes i et eksternt magnetfelt vil spinnvektorene orientere seg i enten samme eller motsatt retning av det eksterne magnetfeltet [14]. Ved å variere det påtrykte magnetiske feltet vil energien og spinnet variere [12]. Resonans absorbsjon skjer hvis magnetfeltet og mikrobølgefrekvens oppfyller $\Delta E = E_{+1/2} - E_{-1/2} = h\nu = g_e\beta_e B$, hvor E er energien til α og β spinnet, h er Plancks konstant

($h = 6{,}6261 * 10^{-34}$ Js), ν mikrobølgefrekvensen, β_e er Bohr magneton og g_e er fritt elektron g-faktor også kalt Zeeman faktor og B er det eksterne magnetfeltet [12]. Figur 2-2 viser oppsplittingen av energinivåer [12], E_0 er grunntilstanden, som er spinnuavhengig og lik for α og β-spinn når det eksterne magnetfelt B er lik null.



Figur 2-2, viser energinivåene til α og β spinnet, som en funksjon av magnetfeltet. Resonansfrekvensen er når det påtrykket magnetfeltet er akkurat stort nok til at overgangen i energi tilsvarer energien imellom energinivåene.

Den første deriverte av mikrobølge resonansfrekvensen som funksjon av det eksterne magnetfeltet blir kalt for EPR-signalet [15]. Tolkningen av et EPR-spekter har tradisjonelt vært å se på høydeforskjellen mellom topp og bunnpunktet til signalet. Denne høyden er et mål for hvor mye spinn det er i prøven [15]. Alle frie radikaler har sitt eget unike EPR-spekter [15]. I litteraturen er det ikke vanlig å ha med y-verdier på EPR-spekter [16]. Det finnes to måter å ta opp et EPR-spekter på. Den ene er å holde magnetfeltet konstant, mens det tilhørende mikrobølgefrekvensen blir variert til absorbsjonsfrekvensen blir funnet. Den andre måten er å holde mikrobølgefrekvensen konstant mens magnetfeltet varieres [12]. I denne oppgaven er det brukt EPR-spektre som er tatt opp med konstant mikrobølgefrekvens.

Hvor stort det totale spinnet er i en prøve, er ikke lett å finne ut av, derfor er det vanlig at resultatet fra EPR-målinger kun angir relativ mengde spinn i prøven [5]. Dette betyr at når EPR bruktes til å fastsette hvor mye spinn det er i en prøve, kan resultatet bare sammenlignes mot tilsvarende målinger med nøyaktig samme opptaksparametere og samme prøvemateriale. Til gjengjeld kan EPR måle det relative spinnet i prøvene ganske nøyaktig, ±1% av nøyaktig verdi [5].

2.2 Ioniserende stråling og stråledoser

Ioniserende stråling kan enten være høyenergiske fotoner (γ-stråling), elektroner (β-stråling) eller små kjerner (α-stråling) [5]. Ioniserende stråling kan ha tilstrekkelig energi til å slå løs et eller flere valenselektron til en kjerne [5]. Hvis dette skjer i et biologisk materiale, fører tapet av et valenselektron til at kjernen eller molekylet blir til et positivt fritt radikal [5]. Valenselektronet som har blitt slått løst kan binde seg til et nytt atom eller molekyl og bli til et negativt fritt radikal [5].

2.2.1 Strålingstyper

α-stråling er tyngre ladde partikler, ofte heliumkjerner, med høy energi [2]. α-stråler mister energien ved støt med elektroner i materialet som absorberer α-partiklene [2]. Ioniseringstettheten til α-partiklene er høy og skjer i en tilnærmet rett linje. Dette fører til nesten alle atomene som ligger i veien for α-strålingen vil miste et elektron. α-partiklene har en rekkevidde på 4 cm i luft og trenger ikke igjennom et papirark [17]. På grunn av sterke Coulomb krefter vil α-stråling nesten aldri vekselvirke med kjernene i atomene [2].

β-stråling er elektroner som har høy energi. β-stråling mister energi til elektroner bundet til atomene i materialet som absorberer β-partiklene [2], på samme måte som α-stråling. Siden elektroner er betydelig lettere enn α-strålinger, er ikke energiavsetningen til β-stråling rettlinjet og ioniseringstettheten er lavere. β-stråler har lengre rekkevidde i vev enn α-stråler. Rekkevidden for β-stråler i luft er opp til 3 meter og trenger ikke i gjennom aluminiumsfolie [17].

γ-stråling er fotoner og vekselvirker med materie igjennom tre hovedprosesser. Disse er: Fotoelektrisk effekt, Comptonspredning og pardannelse [2]. Fotoelektrisk effekt dominerer ved lavenergisk γ-strålingen, Comptonspredning ved midlere og pardannelse ved høyere energinivåer >1,022 MeV [2]. γ-stråling har langt rekkevidde og har en halveringslengde i luft på ca. 150 meter [18]. Det trengs et lag med bly for å stoppe γ-stråling. Fotoelektrisk effekt er når et elektron i et av de indre skallene slås løs [2]. Fotonenergien brukes til å bryte bindingen mellom elektronet og kjernen [2], og til kinetisk energi til det løsrevete elektronet. Et elektron fra de ytre skallene vil falle ned og ta den ledige plassen i det indre skallet, og det sender ut et nytt foton som igjen kan forårsake eksitasjoner [2]. I Comptonspredning vekselvirker et foton med et løstbudent elektron, og overfører noe energi til dette elektronet, mens fotonet fortsetter i en annen vinkel med forminsket energi [2]. I parproduksjon går fotonenergien med til å danne

et positron-elektron par inni i feltet til et atom [2]. For at pardannelse skal skje må energien til fotonet være minimum 1,022 MeV [2]. Positronet, som er et positivt ladet elektron vil annihilere med et elektron og danne to ny fotoner med retning 180^o med hverandre [2].

2.2.2 Stråledoser

Skader fra ioniserende stråling kan grupperes inn i to kategorier. Den ene er direkte effekter, som er når energien til den ioniserende strålingen blir levert direkte til molekylet, som for eksempel, et valenselektron slås løs og molekylet blir til et fritt radikal. Den andre kategorien er indirekte effekter, som er at absorpsjonen av stråling skjer i et annet molekyl enn der hvor skaden oppstår [5]. De fleste strålingsinduserte frie radikaler er kortlivede [5]. Det finnes noen unntak, for eksempel, for flere av radikaltypene som oppstår i aminosyren alanin [9]. Temperatur er en viktig faktor for hvor lenge frie radikaler lever [19]. Noen frie radikaler er stabile ved romtemperatur og trenger ingen spesiell behandling under lagring før de kan måles. Felles for de fleste radikalene er at de blir borte hvis prøvene blir utsatt for høy temperatur over lengre tid.

Stråledoser måles i Gray (Gy), hvor en gray er definert som 1 J/kg [2], det vil si at stråledoser måles i hvor mye energi som deponeres pr. kilogram materiale. For å estimere og sammenligne påvirkningen av ioniserende stråling på mennesker brukes enheten Sivert (Sv), som har samme benevning som Gray. Ulike typer stråling skader levende organismer ulikt, derfor blir en strålingsvektfaktor - ω_R introdusert for å vekte effekten av strålingstype. α-stråler skader vev mer enn β- og γ-stråler [2]. Derfor vektes α- stråler med $\omega_R = 20$ og β og γ-stråler vektes med $\omega_R = 1$. I tillegg introduseres en vevsvektfaktor - ω_T, for å ta hensyn til strålefølsomhet til forskjellige vevstyper [2]. For eksempel blir kjønnskjertlene vektet med $\omega_T = 0{,}20$ og hud blir vektet med $\omega_T = 0{,}01$ [2]. Den effektive dosen (E) blir regnet ut fra den absorberte dosen (A) ved: $E = \omega_T \omega_R A$.

Uttrykket dødelig dose (Lethal Dose, LD) blir brukt for å beskrive hvor stor sannsynlighet det er for at organismen dør, som regel innen 30 dager etter bestråling [20]. LD50 for mennesker er 4-6 Gy, det vil si at i gjennomsnitt estimeres det at 50 % av alle personer som mottar en dose over 4-6 Gy vil dø i løpet av 30 dager etter bestråling [20]. Et røntgenbilde av tennene gir en dose på 5-10 µSv [21] og gjennomsnittlig bakgrunnsstråling på jorden er 2,4 mSv/år [21]. Stråledoser i denne størrelsesorden blir antatt å være ufarlige for mennesker. Normalt blir det

antatt doser <200 mGy har liten eller ingen innvirkning på mennesker [22]. Doser <200 mGy blir i litteraturen som regel kalt små doser og doser over >2 Gy blir definert som store doser.

2.3 Frie radikaler

Definisjonen på et fritt radikal er et atom eller molekyl som har minst et uparet elektron [2]. Frie radikaler er alltid elektrisk nøytrale [2]. Det elektronet som er fritt vil lett la seg binde til andre radikaler [2]. Frie radikaler er gode oksidasjons- og reduksjonsstoffer [2] siden de både mangler et elektron og lett tar opp elektroner. Frie radikaler i levende celler kan være skadelig, siden radikalene kan reagere med og ødelegge eller forandre arvematerialet i en celle [2]. Fri radikaldannelse i levende organismer har størst sannsynlighet for å skje i vann, siden en organisme primært består av vann [2]. Når ioniserende hν stråling treffer et vannmolekyl skjer følgende [2]:

$$H_2O + h\nu \rightarrow H_2O^+ + e^-$$

$$e^- + H_2O \rightarrow H_2O^-$$

H_2O^- og H_2O^+ er ustabile og vil reagere videre til:

$$H_2O^- \rightarrow H^* + OH^-$$

$$H_2O^+ \rightarrow H^+ + OH^*$$

Hvor $*$ betyr fritt radikal, H^* og OH^* kan reagere videre med et organisk molekyl R:

$$OH^* + RH \rightarrow R^* + H_2O$$

$$H^* + RH \rightarrow R^* + H_2$$

Hvor R^* nå er et fritt organisk radikal. Et organisk molekyl kan også omdannes til et fritt radikal ved å interagere direkte med stråling. Dette er tilfellet i rene alanin tabletter. For eksempel:

$$CH_3C(NH_2)HCOOH + h\nu \rightarrow CH_3C^*(NH_3)COOH$$

Hvor et av karbonatomene har blitt til et fritt radikal.

2.4 Aminosyren alanin

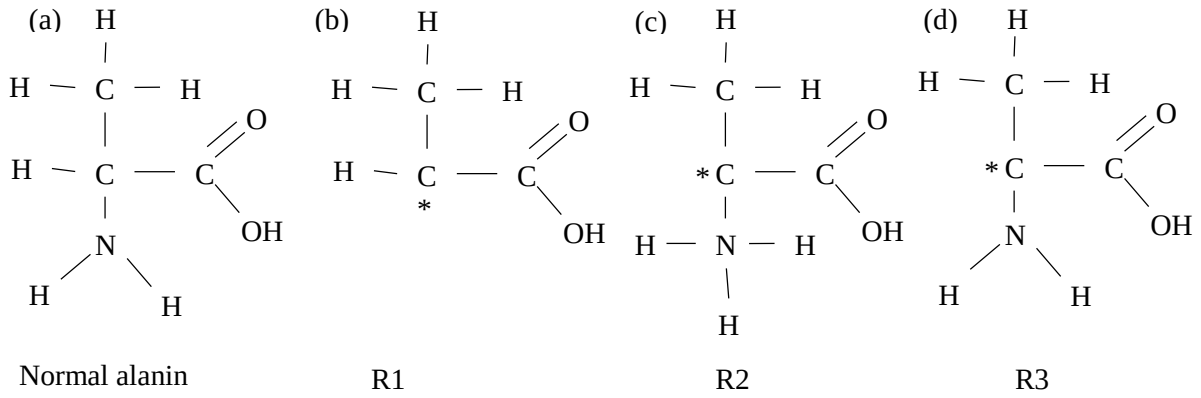
Alanin er en av 20 aminosyrer. Aminosyrer er byggeklossene i proteiner [23]. Derfor finnes det naturlig alanin i alle levende celler [23]. I 1962 foreslo Bradshaw et al. [7] å bruke alanin som det aktive materialet i dosimetre, siden alanin har stabile frie radikaler, daner frie radikaler proporsjonalt med den absorberte dosen og har lignende egenskapene som biologisk vev [7] i forhold til absorpsjon av energi fra ioniserende stråling [5]. Alanin har en lineær sammenheng mellom dose og mengden induserte frie radikaler i doseområdet 2 Gy til 10 kGy. Strålingstype, doseraten samt energien til strålingen har liten betydning for målt dose i alanin, forutsatt at energien er >150 keV [5]. Alanin tabletter er standard til bruk ved doseberegninger for høyere doser (2 Gy – 10 kGy) [5]. En vanlig måte å estimere dosene som alanin har mottatt er ved hjelp av EPR-målinger [5].

2.4.1 Radikaldannelse i alanin

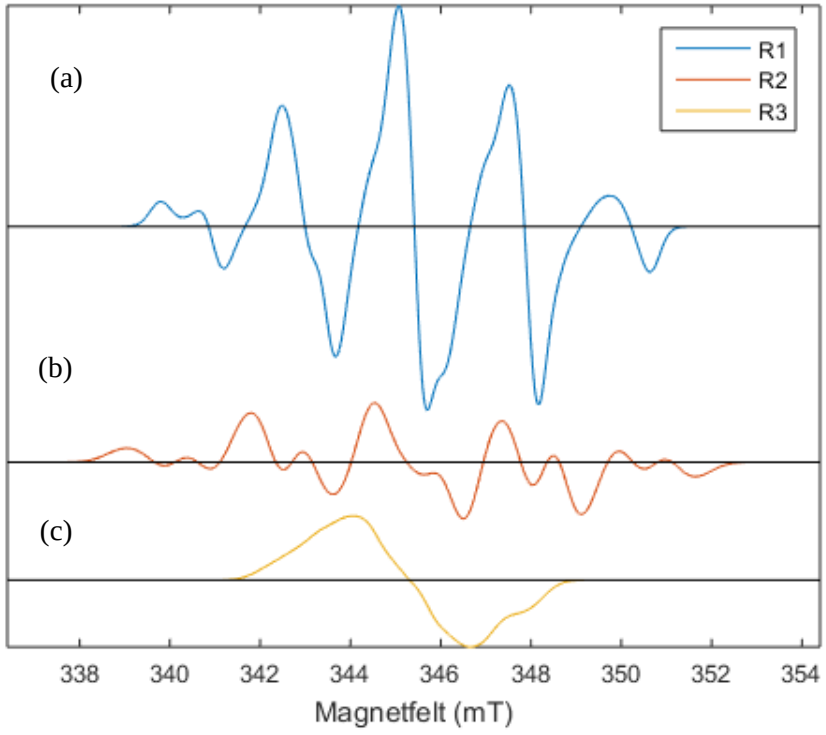
For at EPR-målinger skal kunne estimere dosen som alanin har mottatt nøyaktig, er det viktig å vite hvilke frie radikaler som bidrar til spektrene. Det har lenge vært kjent at det finnes to stabile strålingsinduserte radikaler i alanin. Spektrene til disse er godt kartlagt [9]. Det har også blitt identifisert et tredje radikal [24]. Sagstuen et al. [8] fant at dersom det antas at det finnes kun to frie radikaler (R1 og R2) ved romtemperatur, er disse fordelt i 60 % R1 og 40 % R2. Denne likevekten forskyver seg til 11 % R1 og 89 % R2 ved oppvarming til 480 K [9]. Vanhaelewyn et al. [19] visste at alanin spektrene forandre seg dersom prøvene blir utsatt for temperaturer over 463 K. Vanhaelewyn et al. [19] fant også at ved høyere temperaturer var det mulig å isolere R2-spekteret fra R1-spekteret. Dette stemmer bare dersom antagelsen at det ikke finnes et bidrag fra en mulig R3-radikal gjelder [9]. Heydari et al. [9] fant at R1-radikalet forsvinner fortere enn R2- og R3-radikalene ved varmebehandling. R1/R2/R3 går fra å ha forholdet 55/35/10 % ved romtemperatur til 6/51/43 % ved oppvarming til 480 K, men det er stor usikkerhet knyttet til andelen R3.

De tre strukturformelene for alanin radikalene R1, R2 og R3 er vist sammen med normal alanin i figur 2-3 [5]. R1 er et stabil alanin radikal, som oppstår når den ioniserende strålingen bryter bindingen til ammonium gruppen i alanin (se figur 2-3). R1 blir også kalt stabil alanin radikalet (SAR), siden R1 lenge var antatt å være det eneste stabile radikalet til alanin ved romtemperatur [9]. R2 oppstår når den ioniserende strålingen forårsaker at et hydrogenatomer forsvinner fra karbon-2 og flytter seg til nitrogenatomet (se figur 2-3), mens R3 oppstår når hydrogenatomet

fra karbon-2 forsvinner (se figur 2-3). Alanin kan lagres i opptil 90 dager ved romtemperatur uten at mere enn 5 % av R1-radikalene blir borte [7].



*Figur 2-3, alanin med sine tre stabile former av frie radikaler, R1, R2 og R3 dannet ved bestråling. Eksitasjonen er markert med *. (a) Normalt alanin ubestrålt og er ikke et radikal, (b) R1 den vanligste radikaltypen til alanin og oppstår når bindingen mellom karbonatomet og nitrogenatomet brytes. (c) R2 oppstår når et hydrogenatom flytter seg fra det midterste karbonatomet og over til nitrogenatomet. (c) R3 oppstår når bindingen mellom hydrogenatomet og den midterste karbonatomet brytes.*



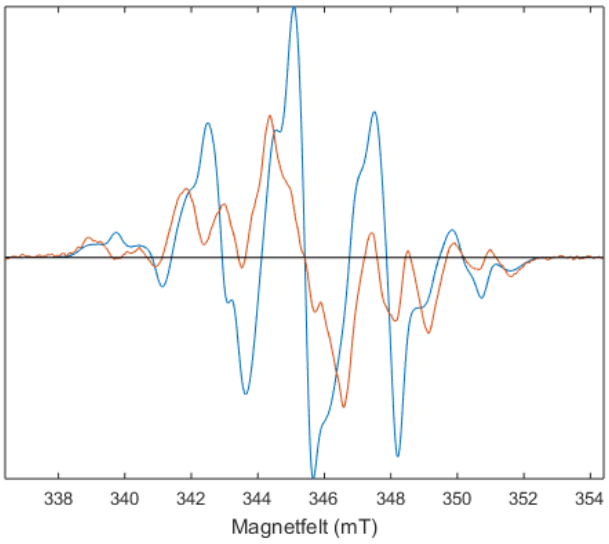
Figur 2-4, de teoretiske basisspektrene til alanin radikalene R1 (a), R2 (b) og R3 (c). Figurene er skalert riktig i forhold til hverandre og vist fra hverandre. Spektrene er hentet fra Heydari et al. [9] og Malinen [16] og er basert på kvantemekaniske simuleringer.

R1-, R2- og R3-radikalene bidrar til det totale EPR-spektret for alanin. Figur 2-4 viser de teoretiske basisspektrene for radikalene, hvor R1- og R2-spekteret er hentet fra i Heydari et al. [9], og R3-spekteret fra Malinen [16]. R3-spekteret beskrevet av Heydari et al. [9] er litt

annerledes enn R3-spekteret som er framstilt i figur 2-4. Forskjellen er at R3 vist her er litt mer glattet enn den som er beskrevet i Heydari et al. [9]. Dette skyldes parameterne benyttet for de kvantemekaniske simuleringene som ligger til grunn for R3-spekteret [16]. Disse radikalspektrene (figur 2-4) er resultater av kvantemekaniske simuleringer. Det teoretiske spekteret til R3 er forbundet med stor usikkerhet grunnet usikkerheter i simuleringene [16].

2.4.2 Temperaturavhengighet til alanin radikaler

Radikalene R1, R2 og R3 er stabile ved romtemperatur, og opptrer som regel med en fordeling på: 55 %, 35 % og 5-10 % [9]. Når alanin utsettes for varme over lengre tid oppfører radikalene seg forskjellig. Mengden R1 avtar raskt når alanin blir utsatt for temperaturer over 463 K [19]. Mengden R2 forsvinner saktere enn R1, og R3 forsvinner saktest av disse tre radikalene. Derfor forskyves den nye likevekten for R1, R2 og R3 seg til 6 %, 51 % og 43 % etter oppvarming til 480 K [9]. At andelene av radikalene forandrer seg med økende temperatur fører til at EPR-spektrene ser forskjellig ut avhengig av hvilken temperatur prøvene utsettes for. Intensiteten til EPR-spektret er lavere ved høyere temperaturer, og andelen av radikalene avtar med forskjellig rate. Figur 2-5 viser forskjellen mellom EPR-spekteret til en prøve målt ved romtemperatur og en prøve oppvarmet til 486 K i 50 minutter. Spekteret til den oppvarmede prøven er skaler med 10 for å gjøre sammenligningen lettere.



Figur 2-5, sammenligning mellom EPR-spekteret til en prøve målt ved romtemperatur (blå) og en prøve utsatt for 486 K i 50 minutter (oransje) (skalert med 10).

2.5 Dosimeter for uforutsette hendelser

Det finnes metoder for å estimere doser ved planlagte eksponeringer [5], men når en uforutsett hendelse skjer, finnes det ikke en standard metode for å estimere doser. Kravet for et material skal egen seg til EPR-dosimetri, er at materialet ikke inneholder vann, er i fast fase og har krystallinsk struktur [5]. For å gjøre analysene enklere ved en uforutsett eksponering av større folkemengder, er det en fordel at materialet som brukes som dosimeter er likt for alle personene og i umiddelbar nærhet til personen. Materialer som kan egne seg er, for eksempel: tenner [25], bein [25], klokker med mineralglass [26], mobil skjermer [10] og briller. For å analysere materialene i EPR-spektrometer må prøvene kunne settes inn i måleapparatet. Dette eliminerer muligheten for tenner og bein som en optimal løsning.

Siden Apple lanserte sin første smarttelefon i 2007 [27], har det blitt vanlig å eie en smarttelefon. En vanlig skjerm på smarttelefoner er laget av Gorilla® Glass. Gorilla® Glass finnes på totalt 4,5 milliarder håndholdte enheter (pr. april 2016) [28]. Gorilla® Glass er en bestemt type glass som brukes primært på berøringsskjermer. Gorilla® Glass blir produsert av Corning Inc., USA [28]. Gorilla® Glass er et tynt glass av alkalisk aluminiumsilikat [29].

Gorilla® Glass var gjenstand for en større EPR-dosimetri undersøkelse ledet av Fattibene et al. [10] som besto av elleve laboratorier. De kom fram til det var mulig å estimere dosen som Gorilla® Glass hadde mottatt med 20 % nøyaktighet i området 0-1,5 Gy og med 5 % nøyaktighet for doser >1,5 Gy. Gorilla® Glass datasettene i denne oppgaven er en del av denne undersøkelsen. Formålet med analysene av Gorilla® Glass i denne oppgaven er å se om det er mulig å lage modeller som er bedre til å estimere doser enn det Fattibene et al. [10] kom fram til.

3 Materialer og metoder

3.1 Alanin datasettet

Datasettet består av 18 EPR-spektre av alanin piller for bruk i dosimetre [30], bestrålt med 5,26 ± 0,09 kGy, av en røntgengenerator med doserate 404 ± 7 Gy/min [30]. Alle prøvene ble preparert og EPR-spektrene tatt opp ved Fysisk Institutt, Universitetet i Oslo. EPR-spektret til to prøver ble målt ved romtemperatur, disse er kontrollene i dette forsøket. De resterende prøvene ble satt i henholdsvis 197 ^{o}C, 205 ^{o}C og 213 ^{o}C, i 1 til maksimalt 150 minutter, etter prøverekkefølgen vist i tabell 3-1. De varmebehandlede prøvene ble avkjølt til romtemperatur før EPR-spektrene ble tatt opp. Det ble benyttet ulike temperaturer og forskjellige oppvarmingstider for å undersøke hvordan spektrene til alanin blir påvirket av oppvarming til rundt 205 ^{o}C og hvor fort en eventuell endring skjer, jamfør Heydari et al. [9] og Malinen et al. [11]. Mikrobølgeeffekten under opptakene var fra 0,20 mW til 2,0 mW. I denne oppgaven er det primært fokusert på 2 mW spektrene. Det eksterne magnetfeltet ble variert fra 336,400 mT til 354,400 mT med et intervall på 0,018 mT, totalt 1024 målepunkter for hvert spekter.

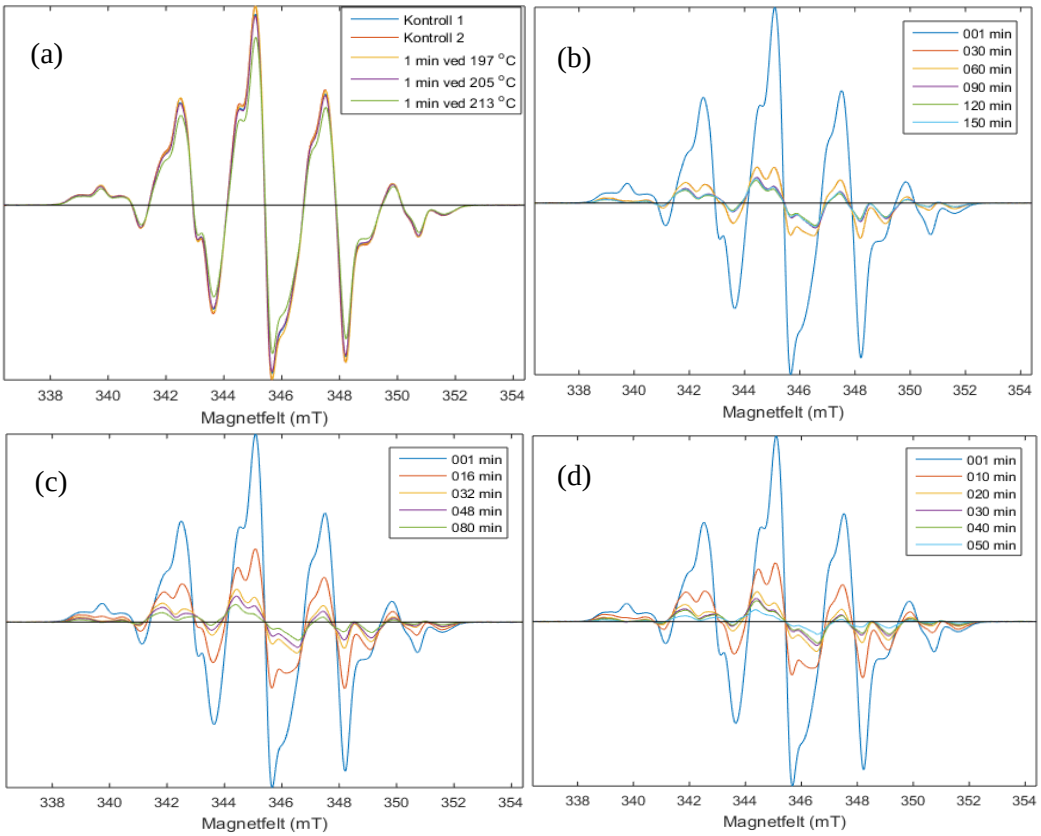
Alle spektrene er blitt baselinjekorrigert (korreksjon for drift i nullnivået) og spektrene er blitt g-sentrert for å sikre at alle målepunktene faller på felles *x*-akse [16]. Spektrene er også skalert i henhold til GAIN-opptaksparameterne [16], før spektrene ble brukt i dette arbeidet.

Datasettet er utformet som en datamatrise **X** hvor hver rad inneholder et EPR-spekteret til en alanin prøve og hver kolonne er et bestemt magnetfelt (variablene). Verdiene i datasettet er relativ intensiteter. Disse har ingen fysisk betydning, og brukes kun for sammenligning av spektrene [16]. Derfor er det ikke i denne oppgaven visst *y*-akse for rå spektrene.

Figur 3-1 viser spektrene for alanin prøvene satt i henholdsvis romtemperatur (kontroller), 197 ^{o}C, 205 ^{o}C og 213 ^{o}C. Figurene viser at intensiteten blir lavere når tiden prøvene holdes ved høyere temperatur øker. Oppvarming i et minutt har liten påvirker på spektret (se figur 3-1a), derfor ble de tre prøvene oppvarmet i et minutt brukt som kontroller [16]. Datasettet har dermed totalt fem kontroller.

Tabell 3-1, rekkefølgen på prøvene i alanin analysene, sorteringen er etter oppvarmingstid.

Indeks	Temperatur [^{o}C]	Tid [min]
1	Kontroll (romtemperatur)	0
2	Kontroll (romtemperatur)	0
3	197	1
4	205	1
5	213	1
6	213	10
7	205	16
8	213	20
9	197	30
10	213	30
11	205	32
12	213	40
13	205	48
14	213	50
15	197	60
16	205	80
17	197	90
18	197	120
19	197	150



Figur 3-1, EPR-spektrene til: (a) kontrollene og til de tre prøvene som kun har blitt oppvarmet i 1. minutt, (b) prøver oppvarmet til 197 ^{o}C, (c) prøver oppvarmet til 205 ^{o}C og (d) prøver oppvarmet til 213 ^{o}C.

3.2 Gorilla® Glass datasettet

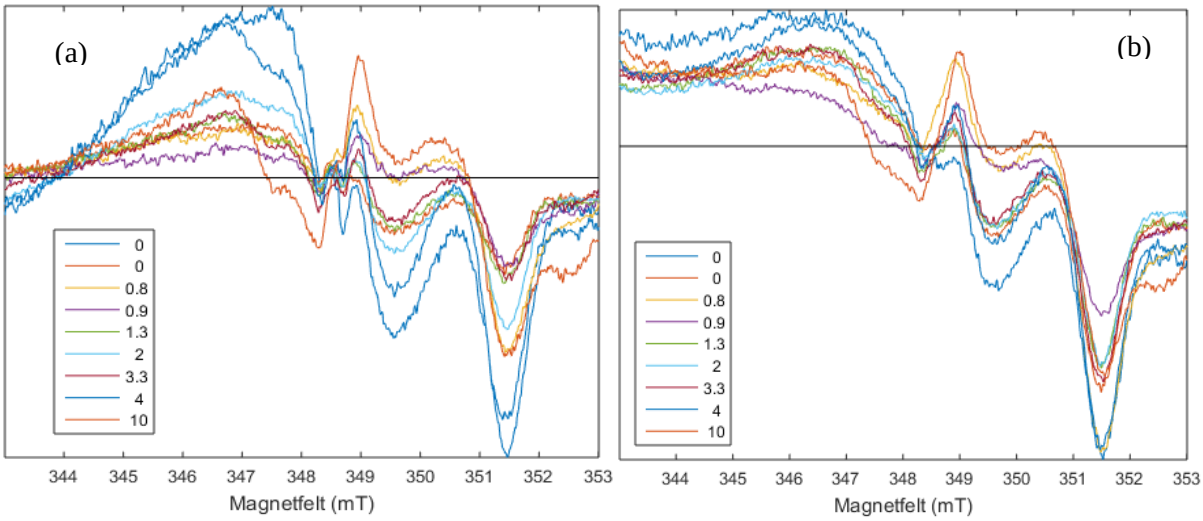
Ni prøver av Gorilla® Glass, ble bestrålt med en ^{60}Co γ-kilde med en doserate på 2 Gy/min [10], ved IRSN [31] i Frankrike. EPR-spektrene av prøvene ble tatt opp i et BRUKER Elexsys560 spektrometer [32] ved Fysisk institutt, Universitetet i Oslo, med enten en SuperX BRUKER kavitet (figur 3-2a) eller en standard rektangulært ST TE_{102} BRUKER kavitet (figur 3-2b) [32]. Opptaksinnstillingene for hver kavitet var lik for alle prøvene, men forskjellige mikrobølgeeffekter ble brukt for hver kavitet [33]. Dette datasettet er en del av en større undersøkelse utført av Fattibene et al. [10].

Datasettet er utformet som en datamatrise **X** som er organisert ved at hver rad inneholder et EPR-spekteret av Gorilla® Glass prøvene. Verdiene i datasettet er relativ intensiteter. Tilhørende hvert spekter finnes det en respons **y** som i dette tilfellet er absorbert dose målt i Gray. Responsen og klassetilhørigheten for de ulike prøvene er vist i tabell 3-2. I denne oppgaven er klassifiseringsgrensene for klasse lavdose <2 Gy og for høydose >2 Gy, dette følger linjene til Fattibene et al. [10] og MULTIBIODOSE 2013 [34]. Prøver som havner i klassen lavdose blir antatt å trenge begrenset lindrende behandling og prøver i klassen stor dose vil trenge lindrende behandling. Disse grensene blir brukt siden behovet for lindrende behandling mot stråleskader vil oppstå før LD50 dosen er mottatt.

Alle spektrene er blitt massekorrigert med masse til 10 Gy prøven som referanse [33]. Spektrene er korrigert for forskjeller i mikrobølgefrekvens.

Tabell 3-2, klassetilhørighet for de ulike prøvene av Gorilla® Glass datasettene.

Respons (y)	Klasse
0	Lav
0	Lav
0,8	Lav
0,9	Lav
1,3	Lav
2,0	Høy
3,3	Høy
4,0	Høy
10	Høy



Figur 3-2, EPR-spekteret av bestrålt Gorilla® Glass med (a) SuperX kavitet, (b) rektangulær kavitet. Etiketten angir absorbert dose i Gy, y-aksen er en intensitetsakse som er lik for alle prøvene.

3.3 Notasjon

I denne oppgaven blir vektorer skrevet som små bokstaver i fet, f.eks. **x**. Matriser blir skrevet som store bokstaver i fet, f.eks. **X**, mens skalare verdier blir skrevet som små bokstaver i kursiv, f.eks. *a*. Dette følger notasjonen i Lay [35]. SI-enheter benyttes [36].

3.4 Preprosesseringstekniker

Forskjellige metoder for å preprosessere EPR-spektrene som har blitt brukt i denne oppgaven er vist under. I vedlegg 8.1 vises figurer av hvordan preprosessering påvirker spektrene. Målet med preprosessering er at den ønskede informasjonen skal komme klarere fram. En god preprosessering tilfører ikke datasettet mere støy enn det den klarer å fjerne [37]. Det er også en fordel at spektrene ikke forandrer form ved preprosessering [37].

Normalisering

Det finnes to former for normalisering som er relevante i denne oppgaven. Den ene er 1-norm hvor hvert spekter blir normalisert med henblikk på areal [38], det vil si at arealet under kurven skal være lik 1. Den andre formen for normalisering er 2-norm som er at lengden på spekteret skal være lik 1 [38]. Begge disse metodene kan brukes for å standardisere lengden på spektrene for å sammenligne profilene, når intensiteten til spektrene i seg selv ikke er interessant.

Sentrering

I sentrering blir gjennomsnittet av hver kolonne (variablene) i datasettet trukket fra den gitte kolonnen [37]. Sentrering blir gjort for at dataene skal kunne tolkes som variasjon rundt gjennomsnittet [38]. For å kjøre en prinsipalkomponent analyse (PCA) eller Partial Least Square (PLS) analyse er det per definisjon nødvendig å sentrere dataene først.

Multiplikative signal korreksjon – MSC

Multiplikative signal korreksjon (MSC) er en metode for å korrigere for additive og multiplikative effekter i et datasett [38]. En additiv effekt er at hver måling har ulikt null-nivå for et gitt målepunkt, mens en multiplikativ effekt er at den målte *y*-verdien for en gitt *x*-verdi øker proporsjonalt med den sanne *y*-verdien. En MSC korreksjon skjer ved å sammenligne hvert spektrum mot et referansespektrum, som ofte er et gjennomsnittspekter [37]. Alle spektrene blir rotert og skalert på en sånn måte at de passer best mulig med referansespekteret. Dette skjer med en minste kvadraters tilpassing av måledataene [39].

Utvidet multiplikative signal korreksjon (EMSC) gjør det samme som MSC, bare at EMSC også korrigerer for avhengigheter i magnetfeltet for EPR-spektrene [38] [40].

Glatting

Det finnes flere forskjellige glattefiltre som kan brukes for å fjerne støy fra spekter. Mest brukt i denne oppgaven er gjennomsnittfilteret, som er et filter som bytter ut et målepunkt med gjennomsnitte av nabomålepunktene, og Savitzky-Golay, som er en metode der forskjellige polynomer blir tilpasset den delen av spekteret som skal glattes [38]. Glatting kan brukes sammen med en kombinasjon av andre preprosesseringsteknikker for å eliminere støy som kan oppstå, spesielt i områder hvor *y*-verdien er lav. Størrelsen på filteret angir grad av glatting, hvor et lite filter glatter lite og et stort glatter mye. I denne oppgaven brukes det stort sett et 15 punkters filter, der glatting er benyttet.

Gjennomsnittsfilteret brukt i kapittel 4.1.8 er gitt som:

$$f(k_j) = \frac{\sum_{i=-\Delta x}^{\Delta x} \mathbf{x}(k_j+i)}{2\Delta x+1} \quad (1)$$

Hvor **x** er en vektor som skal glattes, Δx er bredden på glattefilteret, k_j er indeksen til punktet som skal glattes og f er den nye glattede vektoren.

3.5 Prinsipalkomponent analyse (PCA)

I et datasett hvor det finnes flere variabler (målepunkter) enn det finnes målinger, vil en del av målepunktene være en lineærkombinasjon eller nesten lineærkombinasjoner av de andre variablene. Derfor er det mulig å redusere antall variabler uten å miste mye av informasjonen som er lagret i datasettet. En måte å gjøre denne reduksjonen på er ved hjelp av en Prinsipalkomponent analyse (PCA). PCA går ut på å finne de retningene i et flervariabelt datasett, som maksimerer variansen som blir forklart med noen få prinsipalkomponenter (PC) [41].

PCA går ut på at et sentrert datasett (**X**) deles opp i en matrise av skårene (**T**) og av ladninger (**P**) [37]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} \qquad (2)$$

For å beskrive fullt ut et datasett av størrelse *m* x *n* må rangen til skårene og ladningene være lik den minste av *m* og *n* [37]. For eksempel, for et datasett med 19 målinger av 1024 målepunkter vil rangen til skårene og ladningene være 19, for at datasettet fullt ut skal la seg reprodusere av skårene og ladningene ved ligning (2). I matrisene **T** og **P** er kolonnene sortert etter hvor mye av variansen de forklarer. Første kolonne forklarer mest av den totale variansen i datasettet etterfulgt av kolonne to, osv. De siste kolonnene forklarer svært lite av den observerte variansen. Det er derfor mulig å redusere rangen til **T** og **P** uten at det går utover hvor mye av variansen **T** og **P** forklarer. Ved å redusere rangen i ligning (2) innføres en residualmatrise **E** for at modellen skal fullt ut beskriver datasettet **X**:

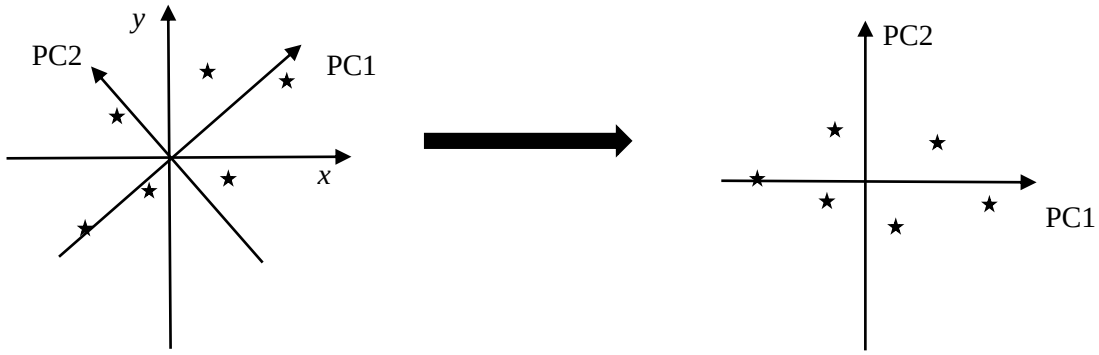
$$\mathbf{X} = \mathbf{T}\mathbf{P}^{T} + \mathbf{E} \qquad (3)$$

Skårene **T**, er relatert til hver enkelt måling. Skårene blir oppgitt som den lineære kombinasjonen av ladningene som skal til for å finne det originale datapunktet i det originale koordinatsystemet. Alle skårene er lineært uavhengige av hverandre [42], mens ladningene er ortonormale lineære kombinasjoner av de originale variablene. Ladningen $\mathbf{p}_1$ er direkte knyttet til skåren $\mathbf{t}_1$. Ladninger og skårer med samme kolonnenummer knyttets sammen når det gjøres en reduksjon av antallet ladninger og skårer.

Det finnes flere måter å bestemme hvor mange prinsipalkomponenter som må være med for at **E** i ligning (3) skal bli så liten som mulig, samtidig som det er ønskelig å ha så få prinsipalkomponenter som mulig med i modellen. En måte vil være å plotte egenverdiene til prinsipalkomponentene og se etter hvor egenverdien gjør et hopp fra et høyt til et lavt tall.

Egenverdien til en prinsipalkomponent er det samme som den absolutte variansen som den samme prinsipalkomponenten forklarer [41]. Hvis den kumulative variansen til prinsipalkomponentene blir plottet vil det være mulig å finne ut hvor mange prinsipalkomponenter som må være en del av modellen, for at modellen for eksempel skal forklare 95 % av den observerte variansen. Et annet mål for å finne riktig antall prinsipalkomponenter kan være å se på Root-Mean-Square Error of Cross Validation (RMSECV).

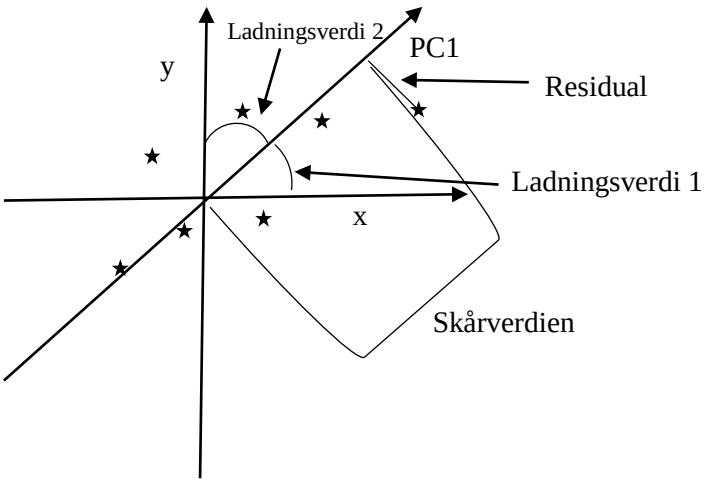
Algoritmen for å finne transformasjonen i et datasett med mange variabler er: Først bestemme den retningen som maksimerer variansen på tvers av alle målingene, denne retningen blir kalt PC1. Deretter blir alle målingene projisert ned på PC1 og residualen mellom det projiserte punktet og originalpunktet blir med videre i analysen. I residualdatasettet som nå ikke inneholder noen informasjon om PC1, blir retningen som har størst varians funnet, denne retningen blir PC2. Informasjonen om PC2 blir så projisert bort fra datasettet på samme måte som for PC1. Denne prosedyren blir gjentatt til alt av varians er blitt forklart. Transformasjonen mellom *x-y* planet og PC1-PC2 planet er skjematisk fremstilt i figur 3-3.



Figur 3-3, transformasjon for datasett med 6 målepunkter fra x-y planet og over til PC1-PC2 planet.

Ladningsverdiene er vinklene mellom originalsakene og prinsipalkomponentaksene, mens skårene er avstanden fra origo og opp til det punktet som står normalt på prinsipalkomponentaksen [42], se figur 3-4.

For alanin datasettet blir PCA brukt for å finne klynger imellom prøver med like egenskaper og finne ut hvordan de underliggende spektrene ser ut. For Gorilla® Glass datasettet blir PCA brukt til å finne klynger av prøver med like egenskaper og for å benytte prinsipal komponent regresjon (PCR).



Figur 3-4, sammenhengen mellom målinger, ladninger, skårer og residualer for en PCA analyse. Residualene blir med videre i analysen og blir brukt til å beregne PC2.

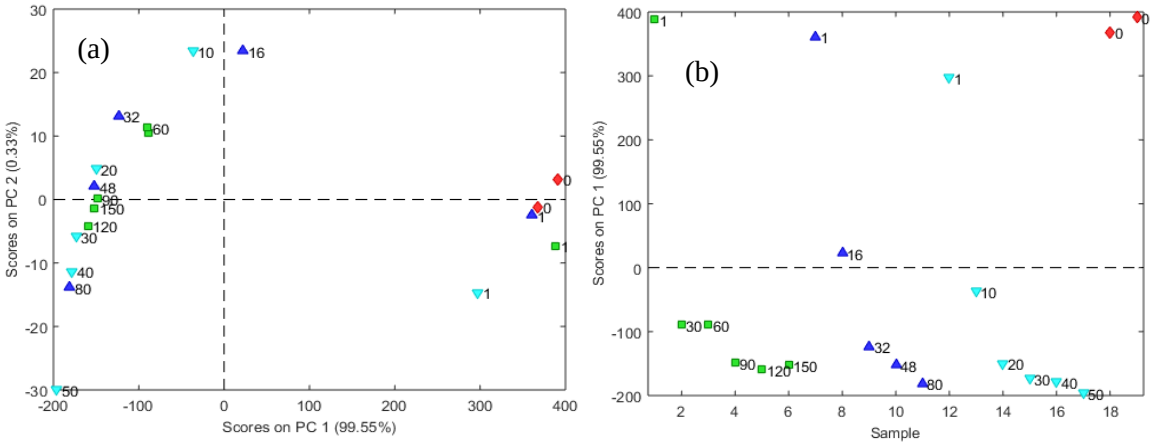
3.5.1 Ulike typer figurer

Skårplott

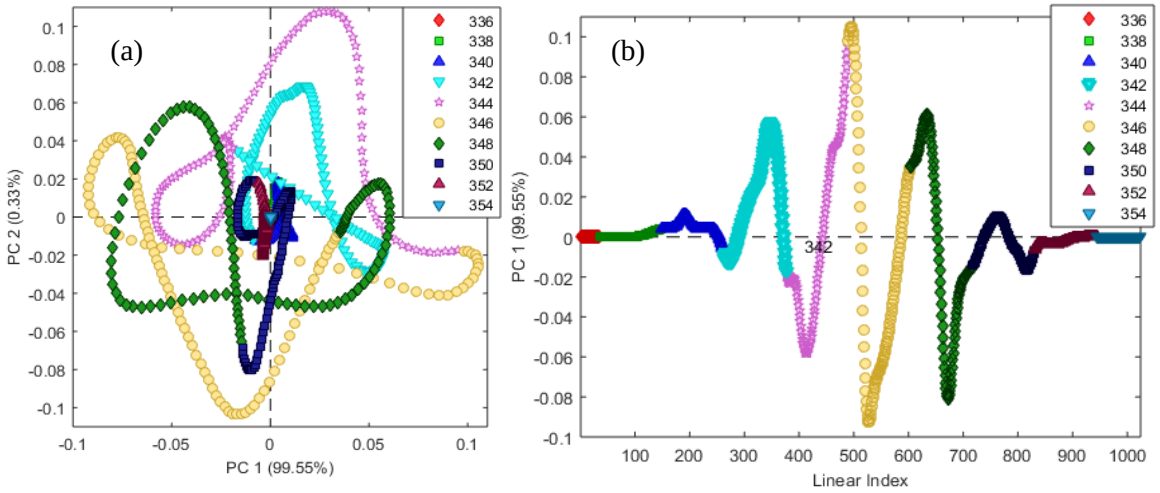
Skårplott viser sammenhengen mellom ulike prøver. Prøver med like egenskaper vil klynge seg sammen [43]. I et skårplott er PC_x plottet mot PC_y, hvor *x* og *y* er et tall imellom 1 og så mange PCer det finnes. Figur 3-5a viser et eksempel på et skårplott. Her er PC1 plottet mot PC2, det kommer fram at prøvene 0 og 1 har mange av de samme egenskapene, siden disse klynger seg sammen. Et annet relevant plot er å plotte skårene for en gitt PC mot prøve nummeret (se figur 3-5b). Dette viser størrelsen de ulike prøvene har i PC-verdi. I eksempelet vist i figur 3-5b har de fleste prøvene svak negativ PC1 verdi.

Ladningsplott

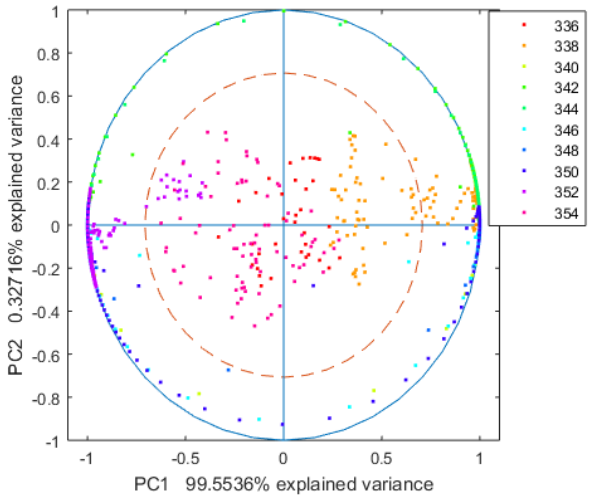
Ladningsplot blir en PC_x plottet mot en annen PC_y, og viser sammenhenger mellom de ulike variablene [43]. Ladningsplottet er viktig for å tolke skårplottet [43] siden ladningene er blitt skalerte for å passe med verdiene til skårplottet. Et eksempel på et ladningsplott er vist i figur 3-6a, hvor det vises at variablene kalt 342-350 har størst verdier for PC1/PC2. Denne type plott forteller ikke så mye om spektroskopiske data, siden spektroskopiske data har høy grad av korrelasjon mellom variablene. En annen form for ladningsplott er å plotte ladningene mot variabelnummeret, se figur 3-6b. Dette kan være nyttig for å lettere se hvilke variabler som drar PC verdien opp.



Figur 3-5, (a) viser et skårplott, her vil prøver med like egenskaper klynge seg sammen og (b) PC1 mot prøvenummer, her er det lett å se hvilke prøver som har negative verdier og hvilke som har positive verdier for PC1.



Figur 3-6, (a) viser et ladningsplott og (b) variabelnummer mot PC1, hvor det er lett å se hvilke variabelnummer som er viktigst for PC1 verdien.



Figur 3-7, korrelasjonsplott over variablene i et datasett. Den stiplede sirkelen står for 50 % korrelasjon, mens den heltrukne står for 100 % korrelasjon mellom variablene og skårene. De viktigste variablene ligger langs den heltrukne linjen.

Korrelasjonladningsplott

Korrelasjonsplott er et plott der korrelasjonen mellom skårene og variablene blir plottet [44] normalt PC1 mot PC2, som i figur 3-7. Den stiplede sirkelen står for 50 % forklart varians, mens den heltrukne linjen står for 100% forklart varians. Punktene innenfor den stiplede sirkelen er variabler som ikke er egnet til å finne sammenhenger mellom skårene [44]. Punktene som ligger på den heltrukne linjen er de målepunktene som er de viktigste variablene for å estimere skårene.

3.5.2 Fordeler med prinsipalkomponent analyse

Ved å plotte skårene i et koordinatsystem med prinsipalkomponentene som akser vil egenskaper ved datasettet kunne oppdages [42]. Målinger med lignede egenskaper vil på et skårplott kunne oppdages som små klynger av punkter, mens målinger som er mer forskjellige vil havne langt fra hverandre. På samme måte som skårplottet viser sammenhenger mellom målinger vil et ladningsplott kunne vise sammenhenger mellom variablene [37]. Hvis variablene i ladningsplottet er langt fra hverandre betyr det at variablene tilfører modellen unik informasjon. Hvis variablene i ladningsplottet er nærme hverandre betyr dette at variablene er høyt korrelerte, noe som betyr at det hadde vært mulig å få de samme resultatene uten å ha så mange variabler. Siden de første prinsipalkomponentene forklarer mesteparten av den observerte variansen i datasettet, vil det som regel holde å plotte de første prinsipalkomponentene mot hverandre for å finne egenskapene i datasettet. Jo høyere prinsipalkomponent nummer som blir studert, jo mer støy vil prinsipalkomponenten vise og følgelig ikke vil være i stand til å forklare egenskapene som målingene har tilfelles.

Siden PCA vektorene er lineært uavhengig vil ikke disse nødvendigvis beskrive sanne sammenhenger i datasettet [45]. Dette gjør at å forklare hvilke egenskaper de ulike PCene står for ikke er så lett siden de ikke forklarer noe fysisk, men bare forklarer de retningene som har mest varians.

3.5.3 Prinsipalkomponent regresjon (PCR)

Ordinær regresjon fungerer best når det finnes flere prøver enn det finnes variabler [37]. Når det skal gjøres regresjonsanalyser på EPR-spektre er det som regel mange flere variabler enn det er prøver. For å unngå dette problemet er det mulig å bruke regresjon på noen eller alle prinsipalkomponentene, dette blir kalt prinsipalkomponent regresjon (PCR).

I PCR bestemmes prinsipalkomponentene ved ordinær PCA og blir deretter brukt videre i en regresjonsanalyse. Først blir skårene og ladningene funnet ved ligning (2), deretter går disse inn i regresjonsanalysen. Regresjonskoeffisientene $\hat{\mathbf{b}}$ for PCR blir funnet ved [37]:

$$\hat{\mathbf{b}} = \mathbf{P}(\mathbf{T}^t\mathbf{T})^{-1}\mathbf{T}^t\mathbf{y} \quad (4)$$

Hvor **P** er ladningene, **T** er skårene og **y** er responsen. Regresjonskoeffisienten kan bli brukt til å estimere responsen $\hat{\mathbf{y}}$ til nye prøver **x**, ved:

$$\hat{\mathbf{y}} = \mathbf{x}\hat{\mathbf{b}} \quad (5)$$

Hvor **x** her er skårverdien til den nye prøven.

3.6 Dekomponering av spektre

Dekomponering av spektre blir i denne oppgaven benyttet for å finne karakteristikker til de underliggende radikalspektrene i alanin datasettet.

3.6.1 Multivariat kurveoppløsning (MCR)

Multivariat kurveoppløsning, (engelsk: Multivariate Curve Resolution, MCR) er samlebetegnelse på en gruppe tekniker som blir brukt til å finne konsentrasjoner og responser i et datasett uten at det trengs å gjøres mange antagelser om hvordan konsentrasjonene og responsene ser ut [38]. I MCR nøstes de underliggende kildene til variasjon i datasettet opp [38], uten nødvendigvis å finne den eller de retningene som maksimerer variansen. MCR består av å dele opp datasettet **X** i en matrise **C** som er rene konsentrasjonsprofiler og en matrise **S** som er de underliggende spektrene [38]:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \quad (6)$$

Det er mulig på samme måte som for PCA å redusere antallet vektorer i **C** og **S**. Modellen utvides da med residualmatrisen **E**:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \quad (7)$$

Ved en MCR analyse er det viktig å preprosessere dataene for å minimalisere støyen og få **X** til å inneholde mest mulig av de ønskede variasjonene i datasettet. Det er viktig at preprosesseringen ikke forandrer på formen til spektrene, dvs. et toppunkt kan ikke bli omgjort

til et bunnpunkt. Da gir ikke MCR analysen informasjon om det originale datasettet. Det betyr at sentrering ikke er en god preprosesseringsteknikk for MCR.

I analysen av alanin EPR-spektret ble MCR brukt til å estimere antallet rene komponenter som datasettet består av, identifisere komponentene og spektrene, samt å kvalitetssikre komponentene og spektrene ved å undersøke residualmatrisen **E** i ligning (7) [38].

Forskjeller mellom PCA og MCR

MCR modellen i ligning (6) ligner på PCA modellen i ligning (2). Forskjellene er at i PCA består **T** av ortogonale vektorer og **P** av ortonormale vektorer, med rekkefølge etter hvor mye av den totale variansen hver enkelt vektor beskriver [38]. I MCR er det ikke noen absolutte betingelser for **C** og **S** [38]. For PCA vil **T** og **P** være unike løsninger for matrisen **X**, men **T** og **P** vil ikke ha noen fysisk betydning. **T** og **P** gir retninger med mest varians, men forklarer ikke nødvendigvis de sanne underliggende faktorene i **X** [38]. **C** og **S** derimot har ikke unike løsninger, men har en fysisk betydning ved at **C** og **S** beskriver faktiske egenskaper ved **X**. Siden spektre ikke står normalt på hverandre er det en fordel at MCR ikke gir ortogonale løsninger [46]. I motsetning til i PCA, sorterer ikke MCR rekkefølgen til vektorene ut fra viktighet [47]. Dette betyr at dersom en av vektorene i **C** og **S** skal fjernes, må det testes om den har stor betydning for modellen eller ikke. Dette gjøres ved å lage en ny modell med færre vektorer og sammenligne den med den gamle modellen. Siden komponentene i MCR ikke er lineært uavhengige vil det måtte regnes ut et nytt sett med **C** og **S** dersom antallet komponenter endres. Dette er imidlertid ikke nødvendig i PCA, siden PCA komponentene er lineært uavhengige og det vil derfor være mulig å finne faktorene med lavest varians uten at modellen må regnes på nytt. Både PCA og MCR gir rom for tolkning av resultatene og kan bli brukt for å estimere fremtidige målinger.

Svakheter med MCR

Siden MCR ikke gir en entydig løsning av ligning (6) vil det kunne finnes mange løsninger for **C** og **S** som passer dataene like bra. Dette vil være av betydning dersom en MCR analyse skal reproduseres. Dersom en av målingene tilføres støy, vil det kunne forandre matrisene **C** og **S** som kommer ut en analysen [38].

Det finnes mange rotasjonsmatriser **T** som oppfyller:

(8) $\mathbf{X} = \mathbf{C}\mathbf{S}^{T} = \mathbf{C}(\mathbf{T}\mathbf{T}^{-1})\mathbf{S}^{T} = (\mathbf{C}\mathbf{T})(\mathbf{T}^{-1}\mathbf{S}^{T}) = \mathbf{C}'\mathbf{S}'^{T}$

Hvor **S**' og **C**' kan være like gode til å beskrive **X** som **C** og **S**. **C** og **S** kan også bli skalert med en skalar k_i:

(9) $\mathbf{X} = \mathbf{C}\mathbf{S}^{T} = \sum_{i=1}^{n} \mathbf{c}_i \mathbf{s}_i^{T} = \sum_{i=1}^{n} \left(\frac{1}{k_i}\mathbf{c}_i\right)\left(k_i \mathbf{s}_i^{t}\right) = \mathbf{C}'\mathbf{S}'^{T}$

Hvor *n* er antallet vektorer i **C** og **S**. På grunn av egenskapene vist i ligning (8) og (9) vil det alltid være mulig å finne andre og/eller bedre matriser for **C** og **S** [46]. Det er derfor mulig å spesifisere betingelser som **C** og **S** skal oppfylle, slik at matrisene **C** og **S** blir unike.

Betingelser

Betingelser er egenskaper ved **C** og/eller **S** som er antatt kjent før en MCR analyse blir gjort [47]. Betingelsene kan være kjemiske eller matematiske egenskaper som det antas at de underliggende faktorer i datasettet skal oppfylle [38]. Ved riktig valg av betingelser kan en MCR analyse gi nøyaktige **C** og **S** som er lette å tolke. Det finnes to hovedtyper av betingelser. Den ene er likhetsbetingelser som går ut på å sette alle elementer i en profil til samme verdi. Den andre hovedtypen er ikke-likhetsbetingelser som tvinger profilen til å være høyere eller lavere enn en bestemt verdi [38]. Det finnes også noen rent matematiske betingelser [47], som ikke nødvendigvis har en forankring i naturlige fenomener. Disse blir brukt for å optimalisere resultatet av MCR analysen.

Eksempler på ikke-likhetsbetingelser:

- Ikke negativitet, brukes dersom verdiene i profilen alltid er positive [38]. Denne betingelsen kan brukes hvis det blir antatt at det bare kan være positive konsentrasjoner eller spektre, sånn som når et spekter er bygget opp av noen ukjente spektre i en vis prosentandel og det er denne prosentandelen som det er ønskelig å finne.
- Unimodality, spesifiserer at det er et maksimum i konsentrasjonsprofilen **C** [47].

Eksempler på likhetsbetingelser:

- Konsentrasjoner **C** i systemet skal summeres opp til en bestemt verdi [38], for eksempel, 100 %, eller at det skal være like mye av alle komponentene i absoluttverdi.

- Noen av spektrene eller konsentrasjonene er kjent [47]. Denne betingelsen kan bruktes når noen/alle underliggende spektrene er kjente og konsentrasjonene skal bestemmes eller at prøvene inneholder en kjent mengde av komponentene og spektrene skal bestemmes fra MCR analysen.

Eksempler på matematiske betingelser:

- Selektivitet, angir at bare noen av variablene brukes i MCR analysen [47] og kan, for eksempel, brukes for å selektere bort variabler med mye støy.
- Sortering av prøvene i stigende eller synkende konsentrasjonsrekkefølge [47].
- Det er mulig å ønske at spektrene eller konsentrasjonene skal ha maksimal varians imellom seg. Denne betingelsen fører til at spektrene eller konsentrasjonenes varians vil bli maksimert innad. Den andre responsen vil da kunne bli liten og være vanskelig å tolke.

Betingelsene velges kun dersom spesifikasjonen oppfylles. Enkelte betingelser vil kunne påvirke spektrene og komponentene ulikt [47], slik at spektrene blir bedre, men komponentene blir vanskeligere å tolke. Betingelsene skal ikke innføre mere støy til modellen enn det betingelsene klarer å fjerne fra modellen [47].

Valg av startpunkt

Dersom de underliggende spektrene eller de sanne konsentrasjonene er kjent, kan disse bruktes som startpunkter for MCR itereringen. Hvis brukeren legger inn sin egen gjetning på konsentrasjonsprofilene **C** og/eller spektrene **S**, er det en fordel at gjettingen er så nærme det virkelige svaret som mulig, for å unngå lokale minimumspunkter [38]. Dersom ikke alle/noen av de underliggende spektrene eller konsentrasjonene er kjent benyttes en algoritme for å estimer de underliggende spektrene, før en MCR analyse kjøres. Den vanligste algoritmen for å finne startpunkter, er utviklende faktoranalyse (Evolving Factor Analysis, EFA) [46] [48] [49]. EFA er en metode basert singulær verdi dekomponering (SVD) transformasjon av dataene [46].

Algoritmer for å regne ut MCR

Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) er en algoritme som bruker en interaktiv metode for å finne **C** og **S** fra ligning (7) og samtidig sørge for at **E** blir så liten som mulig. MCR-ALS tar utgangspunkt i et startpunkt bestemt enten av brukeren eller EFA. Deretter iterere algoritmen seg til et globalt minimumspunkt for **E**. Dette gir [46]:

$$\mathbf{S} = \mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} \quad (10)$$

$$\mathbf{C} = \mathbf{X}\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1} \quad (11)$$

Disse to stegene gjentas inntil **S** og **C** konvergerer. Det kan være nødvendig å kjøre algoritmen med forskjellige startbetingelser for å unngå lokale maksimumpunkter/ minimumspunkter.

PLS_Toolbox

For beregninger med MCR i denne oppgaven er PLS_Toolbox [50] benyttet (se kapittel 3.10). De viktigste betingelsene som er standard i PLS_Toolbox er:

- Konfidensnivå 0,95.
- Ingen antatt kjente underliggende spektre eller konsentrasjonsprofiler.
- Ikke negativitet til konsentrasjonene og de underliggende spektrene.
- Det er ikke maksimal kontrast verken for de underliggende spektrene eller konsentrasjonene.
- Det er ikke bestemt at konsentrasjonene skal summeres til enhetsverdi.
- Spektrene blir ikke normalisert.
- Toleransen til de underliggende spektrene og til konsentrasjonen er satt til $1 * 10^{-5}$.
- Initialiseringsmetode er å velge en måling utenfor datarommet etter normalisering av dataene.

I denne oppgaven (kapittel 4.1.4 og vedlegg 8.3) blir betingelsene over testet, samt at det blir lett etter to-fem komponenter.

Bruk av MCR

I MCR analysene antas R1, R2, både R1 og R2 eller ingen komponenter kjent. De teoretiske spektrene gitt i figur 2-4 ble brukt og normalisert til enhetsareal. Videre ble det antatt at ladningene kunne være negative og at disse var normalisert til enhetsareal. EPR-spektrene ble preprosessert med MSC, EMSC, normalisering eller ingen preprosessering.

3.6.2 Selvmodellerings miksturanalyse (SMA)

Selvmodellerings miksturanalyse (Self-modelling Mixture Analysis, SMA) er en teknikk som minner om MCR. Forskjellen mellom MCR og SMA er at MCR bygger på kontinuitet i dataene, mens SMA bygger ikke på kontinuitet [50]. SMA har som mål å finne de underliggende spektrene og konsentrasjoner (eller noe som er proporsjonalt med konsentrasjonene) til målespektre uten at det må gjøres noen antagelser [37]. Modellen for SMA er lik modellen for MCR vist i ligning (7) [37]. For SMA er det viktig å finne variablene som gir informasjon om de unike underliggende spektrene. Disse variablene kalles rene variabler og finner variabler hvor *y*-verdier til de underliggende spektrene, er ulik null for kun et av spektrene. Disse variablene er rene ved at de kun er koblet til et unikt spekter [37]. For å kunne bruke SMA må det antas at det finnes minst en ren variabel for hver av de underliggende spektrene. Dersom det ikke finnes tilstrekkelige rene variabler, vil de underligendespektrene bli lineær kombinasjoner av de rene variablene [37].

Algoritme for å regne ut SMA

Den generelle formelen for SMA er lik ligning (7), hvor **X** er datamatrisen som er målt, **C** er de ukjente konsentrasjonene, **S** de ukjente spektrene og **E** er en residualmatrisen.

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad (12)$$

SMA tar utgangspunkt i en minste kvadraters løsning av problemet i ligning (12) [51] som minimerer residualen **E**. De reneste variablene i et datasett, kan bestemmes ved beregningene en renhetsvektor **p** definert ved [52]:

$$p_{i,1} = \sigma_i/\mu_i \qquad (13)$$

Hvor i er et tall mellom 1 og antallet variabler i **X**, σ_i er standardavviket for variablen i og μ_i er tilhørende gjennomsnitt for variabelen. Siden μ_i kan bli veldig liten, kan en liten konstant α

bli innført i ligning (13) for å unngå å dele på null. α må være liten sammenlignet med den største verdien i $\boldsymbol{\mu}$ for å forhinder at α dominere modellen:

$$(14) \qquad p_{i,1} = \frac{\sigma_i}{\mu_i + \alpha}$$

Den første rene variabelen er den variabelen som har høyest p-verdi [52]. De neste rene variablene blir funnet ved skalere datapunktene med lengden [52]:

$$(15) \qquad d_{i,j} = \frac{x_{i,j}}{\sqrt{\mu_i^2 + (\sigma_i + \alpha)^2}}$$

Hvor $x_{i,j}$ er datapunktene i $\mathbf{X}$ og $d_{i,j}$ er korrigerte datapunkter til matrisen $\mathbf{D}$. Fra ligning (15) kan en kovariansmatrise bli funnet:

$$(16) \qquad \mathbf{Q} = \frac{1}{n}\,\mathbf{D}^T\mathbf{D}$$

Hvor n er antall prøver. Deretter kan en skaleringsvariabel $\omega_{i,2}$ bli funnet ved determinanten:

$$(17) \qquad \omega_{i,2} = \begin{vmatrix} Q_{i,i} & Q_{i,p_1} \\ Q_{p_1,i} & Q_{p_1,p_1} \end{vmatrix}$$

Hvor p_1 er indeksen til den første rene variabelen og Q_{i,p_1} er korrelasjonen som står på rad i og kolonne p_1 i kovariansmatrisen $\mathbf{Q}$. $\omega_{i,2}$ brukes til å bestemme den andre renhetsvektoren:

$$(18) \qquad p_{i,2} = \left(\frac{\sigma_i}{\mu_i + \alpha}\right) * \omega_{i,2}$$

Den andre rene variablene blir funnet ved å finne indeksen til det høyeste tallet til $p_{i,2}$. Flere rene variabler blir funnet ved å utvide ligning (17) til:

$$(19) \qquad \omega_{i,j} = \begin{vmatrix} Q_{i,i} & Q_{i,p_1} & \cdots & Q_{i,p_{j-1}} \\ Q_{p_1,i} & Q_{p_1,p_1} & \cdots & Q_{p_1,p_{j-1}} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{p_{j-1},i} & Q_{p_{j-1},p_1} & \cdots & Q_{p_{j-1},p_{j-1}} \end{vmatrix}$$

Hvor j er den ønskede rene variablene. Deretter blir ligning (18) gjentatt med $\omega_{i,j}$. Når alle j rene variablene er funnet kan andelen av konsentrasjonene $\mathbf{C}$ regnes ut ved å finne intensitetene i hvert av målespektrene ved de rene variablene.

Når de rene variablene er identifisert og $\mathbf{C}$ er approksimert fortsetter algoritmen ved å kjøre minste kvadraters metode inntil resultatet konvergerer:

(20) $$\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}$$

(21) $$\mathbf{C} = \mathbf{X}\mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}$$

Bestemme antall rene variabler

Antall rene variabler i et datasett kan bli bestemt fra normen til determinanten i ligning (19) [52]. Hvis normen er nær null, er uavhengig informasjon i datasettet hentet ut og den gjenværende informasjonen er støy. Antallet rene variabler blir bestemt ved å sette nedre terskelverdi for størrelsen til determinanten i ligning (19). Rene variabler har $\omega_{i,j}$ større enn denne terskelverdien.

Bruk av SMA

SMA ble utført to ganger. Den ene gangen ble spektrene preprosessert ved normalisering til enhetsareal. Den andre gangen er verdien 54,66 lagt til alle spektrene i datasettet, for å forskyve disse til positive verdier. α-verdien i ligning (14), ble valgt til 0,0001. Dette er 2,5 % av den største μ-en. Variablene 1 – 100, (336,400 - 338,142 mT) og variablene 890 - 1024 (352,042-354,400 mT) ble fjernet, siden de målte spektrene har lav intensitet i disse områdene.

3.6.3 Faktoranalyse (MLCFA)

Maksimum sannsynlighets felles faktoranalyse (Maximum Likelihood Common Factor Analysis – MLCFA) tar utgangspunkt i at målevariablene er lineært avhengige av et mindre antall felles faktorer [53]. En faktor som er felles er lineært avhengig med minst to andre variabler. Korrelasjonen mellom variablene må kunne forklares ved hjelp av disse felles faktorene [53]. Modellen som MLCFA bygger på er at variabel x_i, forholder seg til de andre faktorene (f_r) og ladningene λ_{if} ved:

(22) $$x_i = \sum_{r=1}^{k} \lambda_{ir} f_r + e_i, \qquad i = 1, 2, \dots, p$$

Hvor k er antallet faktorer, p er antallet variabler som måles og e_i er residualen til x_i. Ligning (22) kan skrives på matrise form:

(23) $$\mathbf{X} = \mathbf{\Lambda F} + \mathbf{E}$$

Hvor $\mathbf{\Lambda}$ er ladningsmatrisen, $\mathbf{F}$ er en matrise bestående av faktorene, $\mathbf{X}$ er de datamatrisen og $\mathbf{E}$ er en residualmatrise.

Verdien av de forskjellige parameterne blir estimert ved å finne den største verdien til sannsynlighetsfunksjonen $L(\mathbf{\Lambda}, \mathbf{\Psi})$, gitt ved [53]:

(24) $$\ln L(\mathbf{\Lambda}, \mathbf{\Psi}) = \frac{1}{2}[\ln|\mathbf{\Sigma}| + tr(\mathbf{S}, \mathbf{\Sigma}^{-1})]$$

Hvor $\mathbf{S}$ er kovariansmatrisen til datasettet, gitt ved:

(25) $$\mathbf{S} = \frac{1}{N-1}\mathbf{X_s}\mathbf{X_s^T}$$

Hvor N er antallet variabler og $\mathbf{X_S}$ er den sentrerte data matrisen. $\mathbf{\Sigma}$ er gitt ved:

(26) $$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^t + \mathbf{\Psi}$$

Hvor $\mathbf{\Psi}$ er residualenes kovariansmatrise.

Antallet felles faktorer blir bestemt, ved å gradvis innføre flere felles faktorer inntil modellen ikke blir bedre [19]. Når antallet felles faktorer er bestemt, transformeres disse til komponentspektre, ved av å finne lineær kombinasjoner av felles faktorene og spektrene [1].

Bruk av MLCFA

I denne oppgaven blir MLCFA brukt på alanin datasettet med ingen preprosessering, glatting og normalisering, normalisering og EMSC som preprosessering. MLCFA blir også utført med et redusert datasett som besto av langtidsoppvarmede prøver. Disse er 32, 48 og 80 minutter oppvarmet ved 205 ^{o}C og prøvene varmet i 20, 30, 40 og 50 minutter oppvarmet ved 213 ^{o}C.

3.6.4 Uavhengig komponent analyse (ICA)

Uavhengig komponent analyse (Independent Component Analysis, ICA) er en teknikk for å løse interferensproblemer når det finnes flere kilder som påvirker et målt spekter [54]. Modellen til ICA er [54]:

(27) $$\mathbf{X} = \mathbf{AS}$$

Hvor $\mathbf{X}$ er det målte spektermatrisen av størrelse n x t, hvor n er antall målinger og t er antall variabler, $\mathbf{S}_{n \text{ x } t}$ er matrisen bestående av de underliggende spektrene og $\mathbf{A}_{n \text{ x } n}$ er matrisen som

angir bidragene av hvert underliggende spekter [55], **S** og **A** er skjulte i den forstand at de ikke kan måles, de må estimeres. Målet med ICA er å finne de originale signalene **S** og blandingsmatrisen **A** som inneholder informasjon om konsentrasjonene av **S** i datamatrisen **X** [56].

De viktigste antagelsene som gjøres i ICA er at de underliggende spektrene **S** er uavhengige, **S** kan ikke være normalfordelte [54] og blandingsmatrisen **A** skal være kvadratisk og invertibel [55]. I ICA er to vektorer uavhengig når, $\mathbf{x}_1$ og $\mathbf{x}_2$ ikke er proporsjonale eller parallelle.

Algoritme for ICA

ICA algoritmen har mange av de samme stegene som for PCA. Med utgangspunkt i ligning (27), brukes SVD for å få den sentrerte matrisen **A** på formen**:**

(28) $$\mathbf{A} = \mathbf{U\Sigma V}^{\mathrm{T}}$$

Hvor **U** og **V** er ortonormale og henholdsvis en basis for radrommet og kolonnerommet til matrisen **A**, mens **Σ** er en diagonalmatrise med singulærverdiene til **A** langs diagonalen [35].

Ved å sette ligning (28) inn i ligning (27) og løse for alle yterprodukter for den sentrerte varianten av **X** får vi [56]:

(29) $$\langle \mathbf{XX}^{\mathrm{T}} \rangle = \langle (\mathbf{AS})(\mathbf{AS})^{\mathbf{T}} \rangle = \langle (\mathbf{U\Sigma V}^{\mathrm{T}}\mathbf{S})(\mathbf{U\Sigma V}^{\mathrm{T}}\mathbf{S})^{\mathbf{T}} \rangle = \mathbf{U\Sigma V}^{\mathrm{T}} \langle \mathbf{SS}^{\mathbf{T}} \rangle \mathbf{V\Sigma U}^{\mathrm{T}}$$

Ved å anta at **S** er bygget opp av uavhengige spektre ($\langle \mathbf{SS}^{\mathbf{T}} \rangle = \mathbf{I}$), hvor **I** er identitetsmatrisen, forenkles ligning (29) til:

(30) $$\langle \mathbf{XX}^{\mathrm{T}} \rangle = \mathbf{U\Sigma}^2\mathbf{U}^{\mathrm{T}}$$

Ligning (30) viser at det er mulig å finne **U** og **Σ** ved å løse et standard egenvektor/egenverdi problem for ytreproduktet $\langle \mathbf{XX}^{\mathrm{T}} \rangle$. Deretter beregnes **V**, ved å «hvite» (engelsk: whitening) **X**, hvor avhengigheten i datasettet **X** fjernes, og variansen i alle retninger normaliseres [56]. Dette kan gjøres ved PCA. Hvittingen gir $\langle \mathbf{X}_{\mathrm{w}}\mathbf{X}_{w}^{\mathrm{T}} \rangle = \mathbf{I}$, og følgende ligninger (31) - (34):

(31) $$\boldsymbol{X}_{\mathbf{w}} = \mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{X}$$

Ved å sette ligning (28) og (31) inn i ligning (27):

(32) $$\mathbf{X} = \mathbf{AS}$$

(33) $$\mathbf{U\Sigma X}_{\mathrm{w}} = \mathbf{X} = \mathbf{AS} = \mathbf{U\Sigma V}^{\mathrm{T}}\mathbf{S}$$

(34) $$\mathbf{X}_w = \mathbf{V}^T\mathbf{S}$$

Som igjen lar seg løse for **S** som et egenvektor/egenverdi problem.

Svakheter ved ICA

ICA har mange av de samme svakhetene som MCR. Siden **A** og **S** er ukjente, er det ikke mulig å si at de beregnede **A** og **S** matrisene virkelig gir de sanne underliggende faktoren. Dersom **A** multipliseres med en skalar, vil **S** kunne deles på den samme skalaren og modellen vil kunne være like god [54]. Derfor er det kun mulig å finne størrelsesforholdet mellom **A** og **S**. Det vil heller ikke være mulig å bestemme rekkefølgen på komponentene i **A** og **S**, siden størrelsen på dem er usikker og de kan roteres på samme måte som rotasjonsmatrisen **T** i ligning (8) [54].

De fleste ICA algoritmer bruker et tilfeldig startpunkt, noe som betyr at ICA kan finne lokale maksima og minima som ikke er globale [55]. For å unngå dette, kjøres analysen med flere forskjellige startpunkter for å lette etter et globalt maksimum eller minimum.

Bruk av ICA

I denne oppgaven blir ICA utført med sikte på å finne to og tre komponenter på et ikke preprosessert datasett.

3.7 Regresjon

Regresjons metodene blir i denne oppgaven benyttet både til å estimere de gitte stråledosene i Gorilla® Glass datasettet og beregne andeler av de ulike radikalkomponentene i alanin datasettet.

3.7.1 Minste kvadraters metode

Regresjonsproblemer går ut på å finne beste løsning for **x** i ligningen:

(35) $$\mathbf{Ax} = \mathbf{y}$$

Hvor **y** er en responsvektor. Hvis matrisen **A** inneholder flere variabler enn målinger, vil det ikke være mulig å finne en eksakt løsning på ligning (35). Beste tilpassing for **x** i ligning

(35), kan bestemmes med minste kvadraters metode, med mål å få residualvektoren **e** så liten som mulig [35], gitt ved:

(36) $$\|\mathbf{y} - \mathbf{A}\mathbf{x}\| = \|\mathbf{e}\|$$

I denne oppgaven er **y** den målte dosen, radene i **A** består av målte EPR-spektre, **x** er minste kvadraters løsning av problemet, som er den lineær kombinasjon mellom variablene som gir den beste tilpasning til responsen **y**. Ligning (36) kan løses ved:

(37) $$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{y}$$

Da vil den estimerte $\hat{\mathbf{x}}$ vektoren bli regnet ut ved:

(38) $$\hat{\mathbf{x}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}$$

3.7.2 Lasso

Least Absolute Shrinkage and Selection Operator (lasso) er en metode for å redusere antallet variabler i et datasett [57]. Dette gjøres ved å minimerer summen av kvadratet til residualen, ved å sette en del av regresjonskoeffisientene, som er mindre i absoluttverdi enn en bestemt verdi, til null [58]. Det betyr at en lasso modell er lettere å tolke siden bare de viktigste variablene har en koeffisient forskjellig fra null.

Lasso algoritmen tar utgangspunkt ligning (39):

(39) $$\mathbf{A}\mathbf{x} = \mathbf{y}$$

Hvor i denne oppgaven er **A** en matrise med de målte EPR-spektrene og **y** er dosene. For lasso blir det antatt at vektorene i **A** er standardiserte og oppfyller kravene $\Sigma_{\mathrm{i}=1}^{n} A_{ij}/n = 0$ og $\Sigma_{\mathrm{i}=1}^{n} x^2/n = 1$ [58], hvor $i = 1, 2, \dots, n$ og n er antallet prøver i datasettet og $j = 1, 2, \dots, p$ og p er antallet variabler som måles.

Lasso går ut på å finne de beste estimatene av **x** og konstanten α ved å løse [58]:

(40) $$(\hat{\alpha}, \hat{\mathbf{x}}) = \arg\min\left\{\Sigma_{i=1}^{n}\left(y_i - \alpha - \Sigma_j^p x_j A_{ij}\right)^2\right\}, \text{ sånn at } \Sigma_j^p |x_j| \leq \lambda$$

Hvor $\lambda \geq 0$ er en reguleringsparameter. Normalt er $\hat{\alpha} = \bar{\mathbf{y}}$ [58], og for sentrerte datasett vil denne parameteren forsvinne.

Størrelsen til reguleringsparameteren λ bestemmer hvor stor andel av variablene som skal bidra til modellen. Jo lavere λ, jo flere av variabler i **x** vil få verdien 0 og følgelig ikke bidra i

modellen. For $\lambda = \Sigma\hat{\mathbf{x}}$, hvor $\hat{\mathbf{x}}$ gis i ligning (38), vil lasso regresjonsmodellen være lik minste kvadraters regresjonsmodell. Derfor er minste kvadraters metode, vist i kapittel 3.7.1, et spesialtilfelle av lasso.

Bruk av lasso

I denne oppgaven blir lasso brukt på begge Gorilla® Glass datasettene, med leting etter ni variabler og de 2-3 viktigste variablene.

3.7.3 Delvis minste kvadraters metode (PLS)

Delvis minste kvadraters metode (Partial Least Squares, PLS) går ut på å finne de lineære kombinasjonene av variablene, som maksimere kovariansen mellom variablene og responsen [45]. PLS er en måte å gjøre minste kvadraters metode på, hvor det tas høyde for at variablene kan være høyt korrelerte eller at det finnes flere variabler enn det finnes prøver [45].

Algoritme

Det finnes mange ulike algoritmer for å regne ut PLS [46]. En av de vanligste algoritmene er Nonlinear Iterative Partial Least Squares (NIPLS) [45]. NIPLS algoritmen leter etter underliggende sammenhenger mellom høyt korrelerte variabler og responsen på en iterativ metode [45], samt regner ut skårene, ladningene og koeffisientene [37]. Dataene blir preprosessert med sentrering og skalert [45], men også andre preprosesseringer kan være aktuelle.

Fordeler med PLS

PLS trenger normalt en til to færre komponenter i regresjonsmodellen enn det PCR trenger, for å oppnå samme kryssvalideringsfeil [46]. Dette er en fordel siden modellen blir enklere uten at det går utover nøyaktigheten til modellen. PLS benytter seg av mellomgruppe variasjoner og innad i gruppe variasjoner, noe PCA ikke gjør [59]. Dette gjør at PLS er å foretrekke framfor PCA for å klassifisere gruppe med forskjellige varians mellom og innad i gruppene, siden PCA utelukkende vil finne retninger som maksimerer variansen uavhengig om det er mellom gruppene, innad i gruppene eller en retning midt i mellom.

Intervall PLS

Intervall PLS (IPLS) er en metode for å redusere antallet variabler i et datasett, på en sånn måte at de gjenværende variablene er bedre til å estimere prøver enn det hele datasettet er [37]. IPLS tar utgangspunkt i et intervall med variabler. Dette intervallet kan enten være en samling av frittstående variabler eller et vindu av variabler [37]. IPLS finner den optimale samlingen av variabler for å optimaliser RMSECV verdien for datasettet [37]. Variabelseleksjon av typen IPLS gir ikke nødvendigvis de viktigste variablene for egenskapene til det sanne systemet, men bare de viktigste egenskapene til datasettet som foreligger. Færre variabler betyr mindre informasjon om systemet. Selv om noe av denne informasjonen er støy kan også noe være av viktighet for det sanne systemet. Med færre variabler vil det også bli vanskeligere å oppdage uteliggere i kalibreringsdatasettet og bli lettere å finne falske uteliggere i nye målinger. IPLS kan gjøre en modell bedre, men det må utvises forsiktighet ved bruk.

Bruk av PLS

PLS blir benyttet på SuperX kavitet og rektangulær kavitet datasettene med sentrering og sentrering og MSC som preprosessering. Antallet komponenter som blir benyttet blir bestemt ved å finne det antallet som minimerer RMSECV.

IPLS ble utført på SuperX kavitet og rektangulær kavitet datasettene med sentrering som preprosessering, intervallbredden ble satt til 1, 10 og 100. Algoritmen bestemte hvor mange intervaller som var optimalt. Disse variablene ble brukt til å finne minste kvadraters løsninger med konstantledd.

IPLS ble også utført på Gorilla® Glass datasettene delt opp i et kalibreringsdatasett bestående av prøvene som har blitt bestrålt med dosene 0 Gy, 0,8 Gy, 2 Gy, 4 Gy og 10 Gy og et valideringsdatasett som besto av prøvene 0 Gy, 0,9 Gy, 1,3 Gy og 3,3 Gy.

3.8 Klassifisering

Det finnes to hovedtyper av klassifiseringsmodeller overvåka (supervised) og ikke overvåka (unsupervised). For overvåka klassifisering har treningsdatasettet en kjent respons [57] og det er størrelsen på responsen til nye prøver som det er ønskelig å finne. Eksempler på overvåka metoder er PLS og regresjonsmodeller. Ikke overvåka klassifisering krever ikke kjent respons

og det er mønsteret i dataene som er interessante [57]. Eksempler på ikke overvåk metoder er PCA og K-gjennomsnitt (K-means clustering).

Klassifisering blir i denne oppgaven benyttet på Gorilla® Glass datasettet for å dele datasettet opp i to klasser, definert ut fra høy og lavdose.

3.8.1 Avstandsmål

I en klassifisering må et avstandsmål defineres for avstand mellom to punkter i et multivariat koordinatsystem. I dette arbeidet bruktes euklidsk avstand og Mahalanobis avstand som avstandsmål.

Euklidsk avstand er definert som radius mellom to punkter [41], gitt ved:

$$d(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})} \quad (41)$$

Hvor **x** er det ene datapunktet og **μ** det andre datapunktet. I de fleste klassifiseringsmodeller vil **μ** være gjennomsnittet for alle punktene i en klasse.

Mahalanobis avstand er et avstandsmål der alle punkter (**x**) med samme avstand til et sentrum (**μ**) ligger langs en ellipse bestemt av kovariansmatrisen (**Σ**) til datasettet som bli undersøkt. Ligningen for Mahalanobis avstand er [46]:

$$d(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \quad (42)$$

For enkelte typer målinger og modeller vil Mahalanobis avstand gi en bedre klassifisering enn euklidsk avstand.

3.8.2 Lineær diskriminant analyse (LDA)

Lineær diskriminant analyse (LDA) er en overvåka klassifiseringsteknikk. LDA gjør antagelsen om at alle klassene har samme kovariansmatrise [57]. LDA tar utgangspunktet i den gruppen som en prøve har størst sannsynlighet til å befinne seg i, gitt ved Bayes teorem [60]. Det vil si å finne den klassen *k* som maksimerer sannsynligheten for å observere *k* gitt dataene **x**, $\Pr(k|\mathbf{x})$ gitt ved [57]:

$$\Pr(k|\mathbf{x}) = \frac{f_k(\boldsymbol{x})\,\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\,\pi_l} \quad (43)$$

Hvor $k \in [1, K]$ er klassen som blir undersøkt, π den på forhånd kjente sannsynligheten for å finne $\mathbf{x}$ i klasse k, hvor π er definert som $\sum_{l=1}^{K} \pi_l = 1$. π blir enten satt til å være lik for alle gruppene eller bestemt ut fra hvor mange prøver det er i hver klasse i kalibreringsdatasettet. $f(\mathbf{x})$ er sannsynlighetsfunksjonen for å finne $\mathbf{x}$ definert av:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}-\boldsymbol{\mu}_k)} \quad (44)$$

Hvor $\boldsymbol{\Sigma}$ er kovariansmatrise, p er dimensjonen til hyperplanet som skiller klassene og $\boldsymbol{\mu}$ er klassesentrum.

Ligninger (43) og (44) er en generell ligning som fungerer for mange klassifiseringsproblemer. For LDA er $\Sigma_k = \Sigma$ og lik for alle klassene. Det som skiller LDA fra andre klassifiseringsmetoder er at i LDA er klassegrensene rette hyperplan [57], dvs. lineære klassegrenser. Et nytt punkt blir klassifisert med LDA ved å finne det nærmeste klassesentret til punktet. Klassesentret blir bestemt ved ta gjennomsnittet av alle målingene som er i en bestemt klasse i kalibreringsdatasettet.

3.8.3 Delvis minste kvadraters metode for diskriminant analyse (PLSDA)

Delvis minste kvadraters metode for diskriminant analyse (Partial Least Squares Discriminant Analysis, PLSDA) bygger på PLS for regresjon (se kapittel 3.7.3). I PLS for regresjon er det eksakte verdier for responsen, mens det i PLSDA er klassetilhørighet som er responsen (klasse 0 / klasse 1) [45]. PLSDA prøver å finne den optimale gruppeseparasjonen ved hjelp av mellom gruppekovariansmatrisen [45]. Dette gjøres ved å reduserer antallet variabler til noen få ved hjelp av PLS algoritmen.

3.8.4 K-gjennomsnitt (K-means)

K-gjennomsnitt (K-means clustering) er en ikke overvåka klassifiseringsmetode [41], som går ut på å klassifisere målinger som ligger nærme hverandre, som samme klasse. Algoritmen startes enten ved at alle prøvene blir tilfeldig inndelt i de gitte antallet klasser, eller ved at noen punkter blir tilfeldig valgt som gjennomsnitter for klassene [41]. Prøver som ligger nærmest gjennomsnittet settes inn i den aktuelle klassen, så regnes det ut et nytt gjennomsnitt. Dette gjentas til det ikke oppstår noen forandringer i klassetilhørigheten [41]. I denne oppgaven ble algoritmen kjørt tusen ganger med initialisering: 50% av gangene tilfeldig inndeling i klasse og

50 % av gangen tilfeldig startpunkt, for å være sikker på at beste gruppetilpasning ble funnet. K-gjennomsnittalgoritme er effektiv, som regel gir bare de første itereringene endringer i gruppetilhørighet [41].

Svakheter

Siden K-gjennomsnitt er en metode som består av å prøve og feile, er det en sannsynlighet for at resultatet er et lokalt optimum og ikke et globale optimum. Derfor kjøres K-gjennomsnittalgoritmen mange ganger med ulike startpunkter, for at det skal være sannsynlig at det globale optimum er funnet. Hvilket avstandsmål som brukes i algoritmen påvirker resultatet, og derfor kan flere ulike avstandsmål undersøkes for å finne den beste klassifiseringen.

Uteliggere i datasettet kan enten bli satt i en egen klasse eller kan påvirke klassegjennomsnittet slik at klassegjennomsnittet ikke gjenspeiler det sanne klassegjennomsnittet. Av den grunn kan K-gjennomsnitt være en dårlig metode for å identifisere klasser for datasett med uteliggere.

3.9 Metodevalidering

For å vurdere kvaliteten til analysene trengs det noen statistiske mål, som kan brukes til å sammenligne ulike modeller. Metodene ble validert og evaluert, med teknikkene beskrevet nedenfor. Disse gir et mål på hvor godt en modell kan forutsi utfallet til nye prøver, som ikke er inkludert i kalibreringsdatasettet.

Kryssvalidere

Å kryssvalidere (Cross-Validation, CV) vil si at en modell laget av datasett $\mathbf{X}_{kal}$, som blir kalt kalibreringsdatasett, blir brukt på et tilsvarende datasett $\mathbf{X}_{test}$, som er et uavhengig testsett, for å teste om modellen er riktig [38]. For noen systemer tar prøvetakningen lang tid eller er kostbar. Det kan derfor bli for få prøver til å dele opp i et $\mathbf{X}_{kal}$ og $\mathbf{X}_{test}$. En mulighet for å validere modellen blir da å gjøre en K-fold kryssvalidering [57], hvor prøvene i $\mathbf{X}_{kal}$ deles i K like store deler. Den ene av delene blir lagt til siden, mens K-1 delene blir brukt til å lage en modell, som deretter testes på den delen som er utelatt i modellbyggingen. Deretter byttes det ut hvilken del som blir lagt til siden og en ny modell lages og testes igjen. Dette gjøres til alle K delene har blitt brukt til testing. Kryssvalidering blir gjort for å finne ut om modellen laget på $\mathbf{X}_{kal}$

overtilpasser datasettet. Modellen beskriver egenskaper ved datasettet $\mathbf{X}_{kal}$ bra, men ikke egenskaper til andre prøver som ikke er inkludert i modellbyggingen. Hvis $\mathbf{X}_{kal}$ består av svært få prøver er det vanlig å utelate en måling av gangen [57] (Leave-One-Out Cross-Validation, LOOCV). Dette gjøres ved at en modell blir laget av $\mathbf{X}_{kal}$, men uten rad $\mathbf{x}_i$, deretter blir modellen brukt til å estimere $\mathbf{x}_i$. Dette gjøres for alle målingene i datasettet $\mathbf{X}$. LOOCV residualene blir brukt for å validere modellen.

Estimerte målinger og residualer

En estimert måling er et målepunkt som beregnes ut fra en modell, for eksempel fra en regresjonsmodell, PCA eller MCR modell. En estimert verdi vil aldri være helt lik de målte punktene. En residual, e, er en vektor eller en skalar og er avviket mellom den målte verdien y og den estimerte verdien $\hat{y}$, gitt ved:

$$e = y - \hat{y} \quad (45)$$

For å validere en modell er det mulig å se på hvor stor summen av alle residualene er og finne den modellen som minimerer summen av alle residualene. Residualene viser hvor treffsikkert en modell treffer målepunktene. En tommelfingerregel er at det ikke skal være mer enn 5 % av alle residualene som skal ha en høyere verdi enn to standardavvik [61].

Uteligger

En uteligger er et datapunkt som har en residual som er ekstremt stort [61] i forhold til de andre residualene. Ofte defineres ekstremt stort som tre standardavvik for alle residualene, vek fra gjennomsnittet [61]. Uteliggere er målepunkter som kan ha stor betydning for modellen, siden de drar gjennomsnittet til alle punktene kraftig opp eller ned. Uteligger kan skyldes en målefeil eller at modellen ikke klarer å fange oppførselen til systemet i dette målepunktet, med andre ord kan modellen være for dårlig. En uteligger kan enten fjernes hvis det er flere faktorer som indikerer at uteliggeren skyldes en målefeil, eller hvis ikke, tas med i modellen, siden punktet kan tilføre modellen ny kunnskap som de andre punktene ikke fanger opp.

Q residual og Hotelling T^2

En Q residual er kvadratet av residualen for hver prøve [37], gitt ved:

(46) $$\mathbf{Q} = \mathbf{e}_i \mathbf{e}_i^{\mathbf{t}}$$

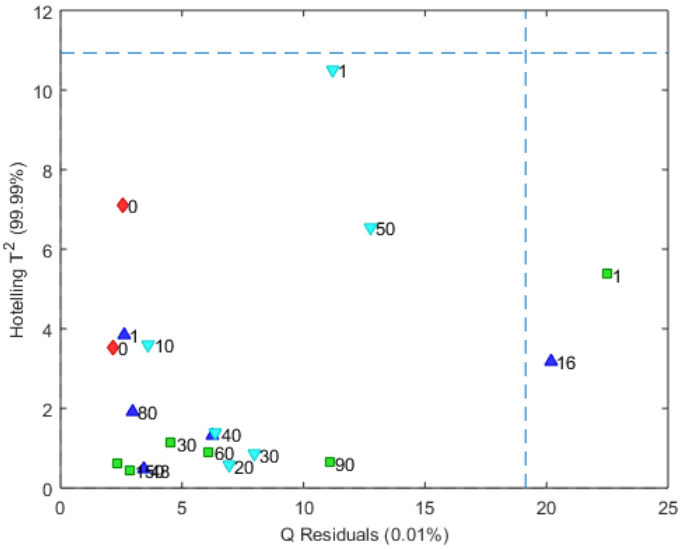
Hvor **Q** er matrisen med Q residualen til prøve i og $\mathbf{e}_i$ er den tilhørende residualenvektoren. **Q** er et mål for hvor godt modellen passer med de faktisk målte verdiene.

Hotelling T^2 er normaliserte og kvadrerte skårverdier, og er et mål for variansen innad i modellen [37]. Hotelling T^2 blir regnet ut ved:

(47) $$T_i^2 = \mathbf{t}_i \boldsymbol{\lambda}^{-1} \mathbf{t}_i^{\boldsymbol{t}}$$

Hvor $\mathbf{t}_i$ er den i skårvektoren til modellen og $\boldsymbol{\lambda}$ er en diagonalmatrise med egenverdiene på diagonalen. T_i^2 beskriver hvordan hver enkelt variabel bidrar til hver prøve [37].

Mens Q er et mål for variasjonen som ikke er forklart av modellen, er T^2 et mål for hvor langt unna gjennomsnittet en prøve er [37]. Disse to målene kan hjelpe oss til å finne uteliggere i datasettet, for eksempel, i et Q mot T^2 plott som er vist i figur 3-8. De stripede linjene representerer signifikansnivået, normalt 95 %. Målingene som ligger innenfor signifikansnivået, det vil si i det nederste, venstre kvadratet, blir beskrevet godt av modellen. Målingene som ligger utenfor de stripede linjene er målinger som passer dårligere inn i modellen, og kan være et tegn på at de er uteliggere.



Figur 3-8, eksempel på et Q residual mot Hotelling T^2 plott, med signifikansgrenser grenser (stiplet linjer). Alle prøver bortsett fra grønn 1 og blå 16, ligger innenfor signifikansgrensene for modellen.

Determinanskoeffisienten R^2

Determinanskoeffisienten R^2 (Coefficient of determination) er et mål på hvor mye av informasjonen i et datasett som blir forklart av modellen [45], og gis ved [61]:

(48) $$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Hvor y_i er målepunktene, $\bar{y}$ er gjennomsnittet av alle målepunktene og $\hat{y}$ er de estimerte målepunktene. R^2 er et tall mellom 0 og 1, hvor 0 betyr at modellen ikke passer til dataene, mens 1 betyr perfekt tilpasning mellom modellen og datasettet.

RMSE verdier

Root Mean Square Error (RMSE) er et mål som brukes i modellvalideringen og beregnes ved at residualen mellom det estimerte punktet $\hat{x}_i$ og det målte punktet x_i blir kvadrert og summert opp [45]. RMSE gis ved:

(49) $$RMSE = \sqrt{\sum_{i=1}^{m}(\hat{x}_i - x_i)^2/m}$$

Hvor *m* er antallet prøver som er med i datasettet.

Det finnes forskjellige typer RMSE, med lik matematiske grunnlag. De vanligste typene er Root Mean Square Error of Prediction (RMSEP), Root Mean Square Error of Calibration (RMSEC) og Root Mean Square Error of Cross Validation (RMSECV). RMSEC er et mål for gjennomsnittsdifferansen mellom de estimerte og de målte verdiene i kalibreringsdatasettet [38], mens RMSEP er et mål for gjennomsnittsdifferansen mellom de estimerte og de målte verdiene for fremtidige målinger [38], regnet ut ved hjelp av et treningsdatasett. RMSECV er et mål mellom de estimerte og de målte verdiene på de utelatte prøvene i en kryssvalidering.

Korrelasjon

Det finnes flere former for korrelasjon. En av de vanligste er Pearson korrelasjon [62]. Pearsons korrelasjon blir regnet ut ved:

(50) $$Cov_{xy} = \sum_{i=1}^{N} \frac{(\mathbf{x}_i - \bar{\mathbf{x}}) * (\mathbf{y}_i - \bar{y})}{N-1}$$

(51) $$Corr_{xy} = \frac{Cov_{xy}}{var(\mathbf{x})\, var(\mathbf{y})}$$

Hvor Cov er kovariansen og $Corr$ er korrelasjonen mellom vektor **x** og vektor **y**, bestående av *N* tall, $\bar{x}$ er gjennomsnittet for **x**-vektoren og $var(\mathbf{x})$ er variansen til vektor **x**. Korrelasjonen vil alltid bli et tall mellom -1 og 1, hvor -1 betyr at **x** og **y** er identiske, men med motsatt fortegn, 0 betyr at **x** og **y** er helt ulike og 1 betyr at **x** og **y** er helt like [62].

Manglende tilpasning

Manglende tilpasning (engelsk: Lack Of Fit, LOF) er et mål for hvor mye av inputmatrisen **X** som finnes igjen i modellen [63]. LOF er definert ved:

$$LOF = \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} x_{i,j}^2}} \quad (52)$$

Hvor $e_{i,j}$ er en residual mellom målepunktet $x_{i,j}$ og det predikerte punktet $\hat{x}_{i,j}$. Siden LOF er et mål for hvor dårlig modellen passer med måledataene er det en fordel at LOF verdien er så lav som mulig. I denne oppgaven blir LOF regnet ut med residualene og målingene basert på en ferdig preprosessert **X**, for å unngå skaleringsfeil dersom **X** skaleres som en del av preprosessering.

Nøyaktighet

Resultatene fra klassifisering kan settes inn i en klassifiseringstabell. Tabell 3-3 vise et eksempel for en klassifiseringstabell med to klasser. Antall riktigklassifiserte er (SP+SN) og antall feilklassifiserte er (FN+FP). Nøyaktigheten til en klassifiseringsmodell kan regnes ut ved [64]:

$$Nøyaktighet = \frac{SP+SN}{SP+SN+FP+FN} \quad (53)$$

Tabell 3-3, eksempel på en klassifiseringstabell med to klasser.

		Gitt verdi	
		Positiv	**Negativ**
Estimert verdi	**Positiv**	Sann Positiv (SP)	Falsk Positiv (FP)
	Negativ	Falsk Negativ (FN)	Sann Negativ (SN)

3.10 Programvare

I denne oppgaven har Matlab® (versjon R2015a, The MathWorks®, Inc, Natick, Massachusetts, USA [65]) blitt brukt, sammen med tilleggspakkene:

- Statistics and Machine Learning Toolbox™ (versjon 10.0, The MathWorks®, Inc, Natick, Massachusetts, USA [66]), brukt til lasso og MLCFA.
- PLS_Toolbox (versjon 8.1, Eigenvector Research, Wenatchee, Washington, USA [50]) brukt til å preprosessere dataene, IPLS, MCR, PCA, PCR, PLS og PLSDA.
- FastICA for Matlab® 7.x and 6.x (versjon 2.5, Hugo Gävert, Jarmo Hurri, Jaakko Särelä, and Aapo Hyvärinen, Department of Computer Science, Aalto University, Finland [67]) brukt til å finne ICA.

I tillegg er det brukt en del egenutviklede skript basert på teorien som er beskrevet over.

4 Resultater

4.1 Analyser av alanin datasettet

4.1.1 Minste kvadraters tilpasning av alanin spektrene

Tilpasning basert på to radikalkomponenter

Alle spektrene ble tilpasset de teoretiske R1- og R2-spektrene (figur 2-4) ved bruk av minste kvadraters metode for å estimere andelen av R1 og R2. Resultatene er vist i tabell 4-1, hvor kolonnene merket R1 og R2 viser beste andel av komponent R1 og R2 i målespektrene. For kontrollprøvene gir fordelingen 60 % R1 og 40 % R2 beste tilpasning, som er i tråd med funnene til Heydari et al. [9]. LOF verdiene er relativt høye for alle prøvene, noe som kan indikere at det finnes flere komponenter i spektrene. Korrelasjonen og R^2 er høy for de fleste spektrene og følger utviklingen til LOF. LOF verdiene er høyest for målingene utsatt for høy temperatur over lang tid, og lavest for kontrollene. Dette tyder på at det er lettest å tilpasse R1 og R2 til kontrollene, og at tilpassing av spektrene er dårligere for prøvene som har blitt oppvarmet ved høy temperatur lenge. For disse prøvene er bruk av teoretiske basisspektere for R1 og R2 ikke tilstrekkelig for å tilpasse de målte spektrene.

Tabell 4-1, tilpassing av målespektrene til teoretiske basisspektere for R1 og R2 ved bruk av minste kvadraters metode. LOF, korrelasjon og R^2, mellom estimert spekter og måltspekter er vist. Pilene viser at LOF øker for prøver oppvarmet over lengre tid.

Temperatur [°C]	Tid [min]	R1	R2	LOF	Korrelasjon	R^2
197	1	0,618	0,382	0,147	0,989	0,978
197	30	0,240	0,760	0,264	0,965	0,930
197	60	0,238	0,762	0,259	0,966	0,933
197	90	0,124	0,876	0,365	0,931	0,866
197	120	0,106	0,894	0,385	0,923	0,852
197	150	0,121	0,879	0,366	0,931	0,866
205	1	0,605	0,395	0,137	0,991	0,981
205	16	0,360	0,640	0,200	0,980	0,960
205	32	0,156	0,844	0,329	0,944	0,892
205	48	0,109	0,891	0,380	0,925	0,856
205	80	0,046	0,954	0,488	0,873	0,762
213	1	0,613	0,387	0,152	0,988	0,977
213	10	0,286	0,714	0,235	0,972	0,945
213	20	0,108	0,892	0,383	0,924	0,853
213	30	0,056	0,944	0,466	0,885	0,783
213	40	0,048	0,952	0,478	0,879	0,772
213	50	0,017	0,983	0,581	0,814	0,662
Kontroll 1	0	0,601	0,399	0,137	0,991	0,981
Kontroll 2	0	0,598	0,402	0,135	0,991	0,982

Tilpasning basert på tre radikal komponenter

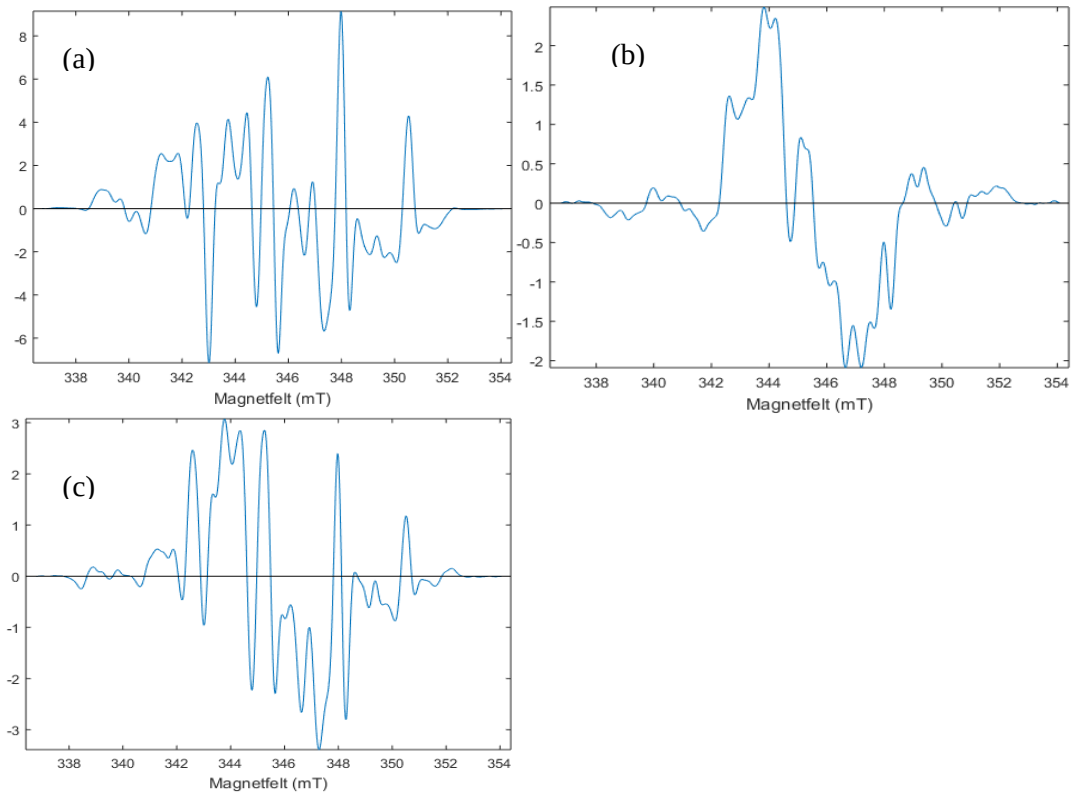
Residualene mellom målespektrene og de estimerte spektrene fra tabell 4-1 ble brukt som estimater for radikalet R3, en tredje mulige alanin radikal. Tre forskjellige estimater ble vurdert.

Gjennomsnittet av residualen til de fem kontrollprøvene ble brukt som et estimat av R3*-spekteret (figur 4-1a). Deretter ble minste kvadraters metode gjentatt på alle målingene ved bruk av R1-, R2- og R3*-spektrene. Tabell 8-1 vedlegg 8.2 viser at LOF verdien er lav og korrelasjonen og R^2 mellom målespektrene og de estimerte spektrene er høy for kontrollene, mens LOF for de andre spektrene er høy. Dette betyr at kontrollspektrene blir overtilpasset, og at residualspekteret er en støykomponent av kontrollene. Derfor er ikke denne R3* et godt estimat på den sanne R3*.

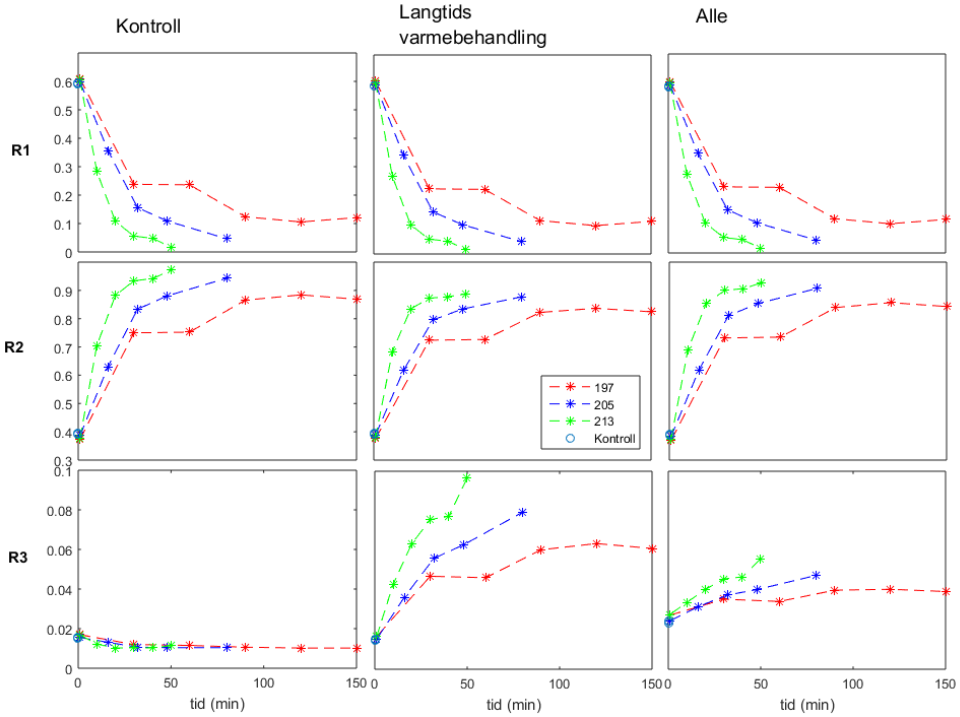
Siden LOF verdien blir ganske høy for målingene som har blitt utsatt for høy temperatur over lengre tid (tabell 4-1), ble gjennomsnittsresidualen for de fem prøvene som er antatt å inneholde mest R3 funnet. Disse prøvene er 197 ^{o}C 150 minutter, 205 ^{o}C 80 minutter og 213 ^{o}C 30, 40 og 50 minutter. Gjennomsnittsresidualen vises i figur 4-1b. Tabell 8-2 vedlegg 8.2, viser at det er høy korrelasjonen for alle målingene. Siden korrelasjonen er høyest for de fem prøvene som er brukt til å lage residualen er det grunn til å tro disse prøvene er overtilpasset når dette estimatet av R3*-spekteret blir brukt. Siden korrelasjonen blir høyere når denne residualkomponenten inkluderes i tillegg til R1 og R2 i minste kvadraters metode, kan dette tyde på at det er behov for en tredje komponent for å fullt ut beskrive formen til EPR-spektrene.

Minste kvadrater metode ble brukt på gjennomsnittet av alle residualene som et tredje estimat for R3, denne residualen er vist i figur 4-1c. Resultatene fra minste kvadraters analysen er vist i tabell 8-3 vedlegg 8.2, her korrelasjonen ganske høy for alle prøvene.

Figur 4-2 viser estimert andel av R1-, R2- og R3*-radikalene basert på de tre forskjellige R3* estimater. Andelen R3* er størst dersom R3 beregnes fra langtids varmebehandlede prøvene. Andelen av R1 og R2 er omtrent lik uavhengig av hvilken metode som er benyttet til å estimere R3. Mengden R1 avtar med økende tid. Raskest går det for prøvene behandlet ved 213 ^{o}C, R2 øker med tiden, mens R3* holder seg relativt konstant når R3* estimeres fra kontrollene og alle prøvene.



Figur 4-1, gjennomsnittsresidualene mellom de målte og estimerte spektrene fra minste kvadraters metode: (a) for de fem kontrollene, (b) for de fem målingene som ble antatt å inneholde mest R3. Dette er prøvene 197 °C 150 minutter, 205 °C 80 minutter og 213 °C 30, 40 og 50 minutter og (c) for alle spektrene. Spektrene (a) og (c) ser ut til å inneholde mye støy.



Figur 4-2, andeler av de ulike radikalkomponentene, funnet ved bruk av minste kvadraters metode. R3 ble funnet ved å beregne residualen til kontrollspektrene (venstre kolonne), ved residualen til prøvene med langtids varmebehandling (midterste kolonne) og ved alle residualene (høyre kolonne).*

4.1.2 Prinsipalkomponent analyse (PCA)

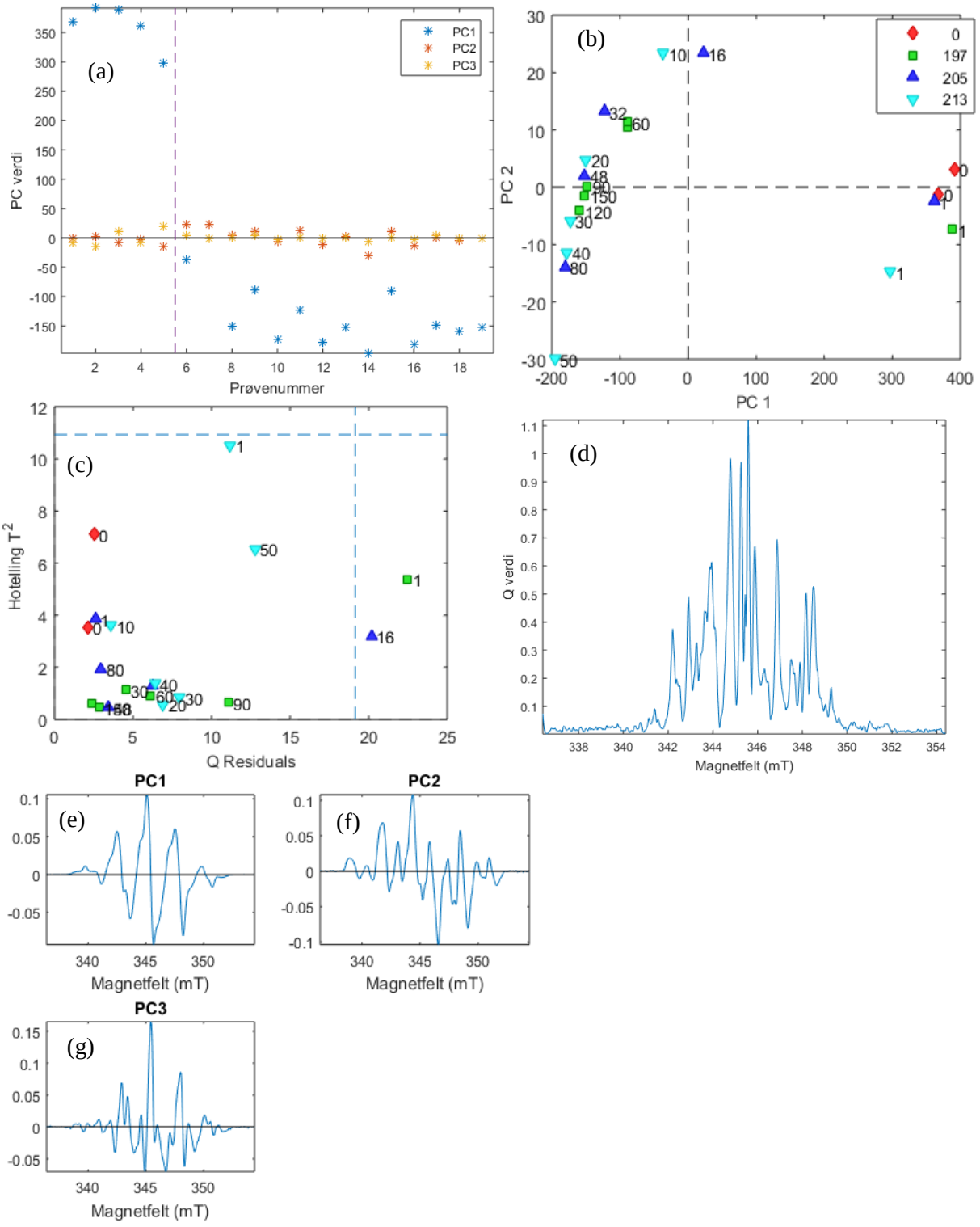
PCA med sentrering som preprosessering

Prinsipalkomponent analyse ble utført på et sentrert alanin datasettet. Figur 4-3a viser skårene til de tre første prinsipalkomponentene mot prøvenummer. Nesten all variansen i datasettet blir forklart av PC1 som forklarer 99,55 % av variansen, PC2 forklarer 0,33 % av variansen og PC3 forklarer kun 0,10 % av variansen. Dette viser at det kun er nødvendig med en komponent (PC1) for å reprodusere datasettet. Siden et mål i dette arbeidet er å finne tre radikal komponenter vises de tre første PCene. Skårplottet i figur 4-3b viser at kontrollene og prøver som har blitt oppvarmet i 1 minutt klynger seg sammen ved positive PC1 verdier, mens prøver varmebehandlet i lengre tid klynger ved negative PC1. PC2 blir lavere/mer negativ for prøver som har høyere LOF verdier i tabell 4-1. Q residualene mot Hotelling T^2 i figur 4-3c viser at residualen til prøve 197 ^oC 1 min og 205 ^oC 16 min er større enn 95 % konfidensintervallet til residualene, men ingen prøver har store Hotelling T^2 verdier. Figur 4-3d viser Q residualene til ladningene. Det er mange spisse topper, noe som tyder på at det kun er enkelte av magnetfeltmålingene som har stor betydning for PCA modellen.

Ladningene til PCA analysen er vist i figur 4-3e-g. Korrelasjonsmatrisen mellom PC1, PC2 og PC3 mot de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4), samt et av kontrollspektrene er vist i tabell 4-2. PC1 har høy korrelasjon med det teoretiske R1-spekteret. PC2 korrelerer til en viss grad med R2 og PC3 er ikke godt korrelert med noen av de tre spektrene. Siden PC1 har en 99,5 % korrelasjon med kontrollspekteret, kan det virke som at PC1 kun gjengir kontrollspektrene.

Tabell 4-2, korrelasjonsmatrisen mellom PC1, PC2 og PC3 mot de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4), samt til en av kontrollene, ved PCA av sentrert alanin datasett.

	PC1	PC2	PC3
R1	0,987	-0,059	-0,033
R2	0,536	0,746	0,070
R3	0,111	0,495	0,096
Kontroll	0,995	0,094	0,003



Figur 4-3, resultatene fra PCA utført på sentrert alanin datasett: (a) Skårene til PC1, PC2 og PC3 mot prøvenummer (tabell 3-1). Skårverdiene til PC1 er positiv og høye for de fem kontrollene og negativ for resten, mens PC2 og PC3 verdiene svinger rundt 0 for alle prøvene. (b) Skårplott, PC1 mot PC2, etiketten angir hvilken temperatur prøvene er oppvarmet til, 0 betyr kontroll/ingen oppvarming. Her kommer det tydelig fram at kontrollene klynger seg sammen ved positive PC1 verdier. (c) Q residualer mot Hotelling T^2 til skårene. (d) Q residualene til ladningene og ladningene til PC1 (e), PC2 (f) og PC3 (g).

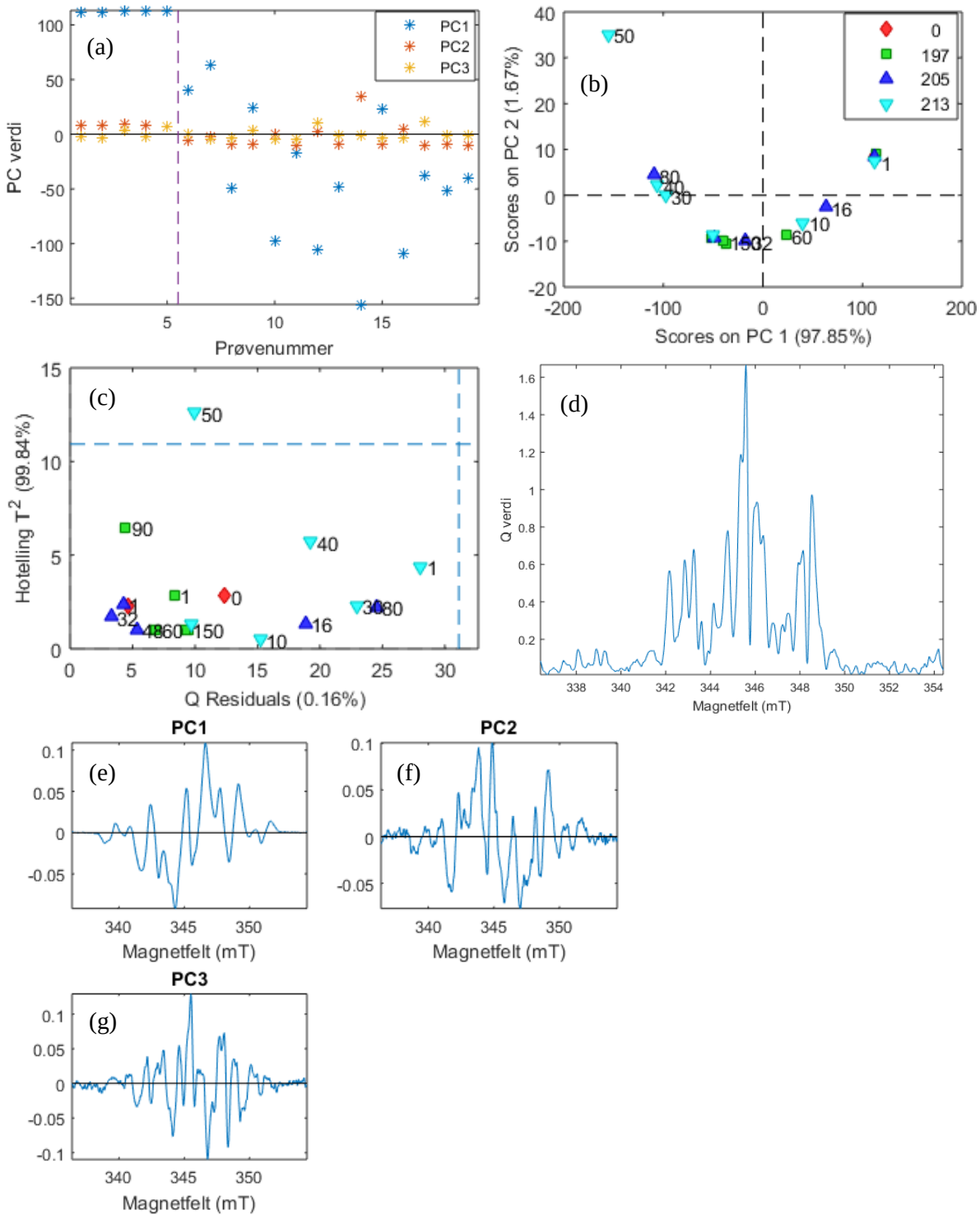
PCA med MSC og sentrering som preprosesseringsteknikk

En tilsvarende prinsipalkomponent analyse ble gjort med preprosesseringsteknikken (i rekkefølge) MSC, glatting med 15 målepunkters glattefilter og sentrering. I dette tilfellet forklarer PC1, PC2 og PC3 henholdsvis 97,85 %, 1,87 % og 0,31 % av den forklarte variansen. Skårene mot prøvenummer i figur 4-4a, viser at størrelsen til PC2 og PC3 varierer rundt null. Skårplottet vist i figur 4-4b viser at kontrollene og 1 minuttsoppvarmede prøvene klynger ved positive PC1 verdier. De varmebehandlede prøvene følger en parabel fra positive PC1 til negative PC1 verdier, i rekke følge etter antatt økende mengde R3. Q residualene mot Hotelling T^2 (figur 4-4c) viser at det kun er prøve 213 ^oC 50 minutter som har unormalt høy T^2 verdi. Q residualen til variablene gitt i figur 4-4d, viser at residualen er størst midt i det målte intervallet.

Ladningene til PC1, PC2 og PC3 er vist i figur 4-4e-g. Korrelasjonsmatrisen for PC1, PC2 og PC3 mot teoretiske spekteret til R1, R2, R3 og en kontroll er vist i tabell 4-3. Det er lav korrelasjon mellom PCene og de teoretiske basisspektrene til R1 og R2 (figur 2-4). Dette betyr at det ikke er noen sammenheng mellom PCene og de antatte profilene til R1 og R2.

Tabell 4-3, Korrelasjonsmatrisen mellom PC1, PC2 og PC3 mot de teoretiske basisspektrene til R1, R2, R3 og en kontroll som vist i figur 2-4, samt til en av kontrollene. MSC og sentrering ble brukt som preprosessering for alanin datasettet.

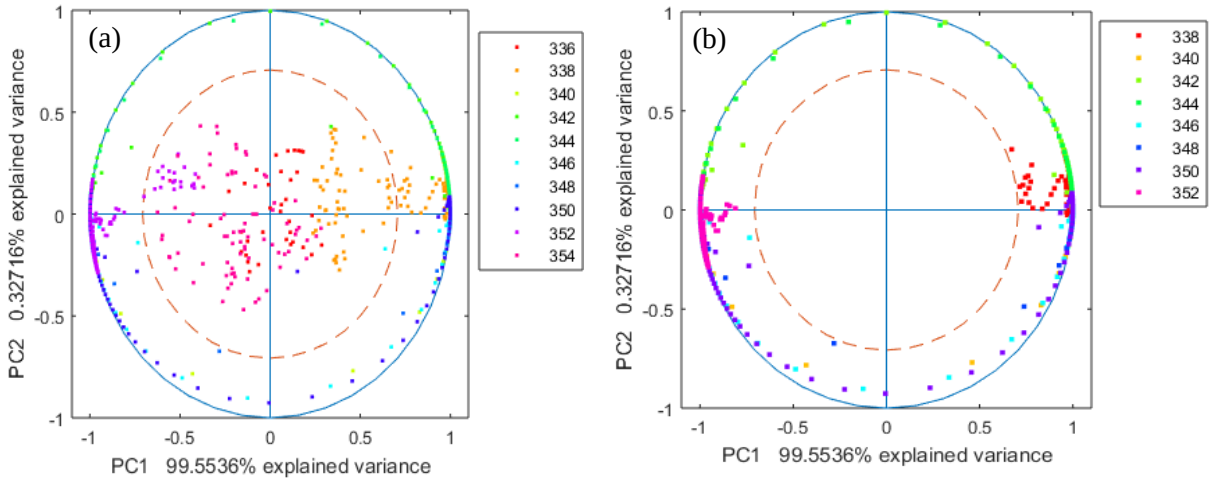
	PC1	PC2	PC3
R1	0,301	0,097	-0,001
R2	-0,507	-0,380	-0,016
R3	-0,716	0,574	0,047
Kontroll	0,163	0,035	-0,004



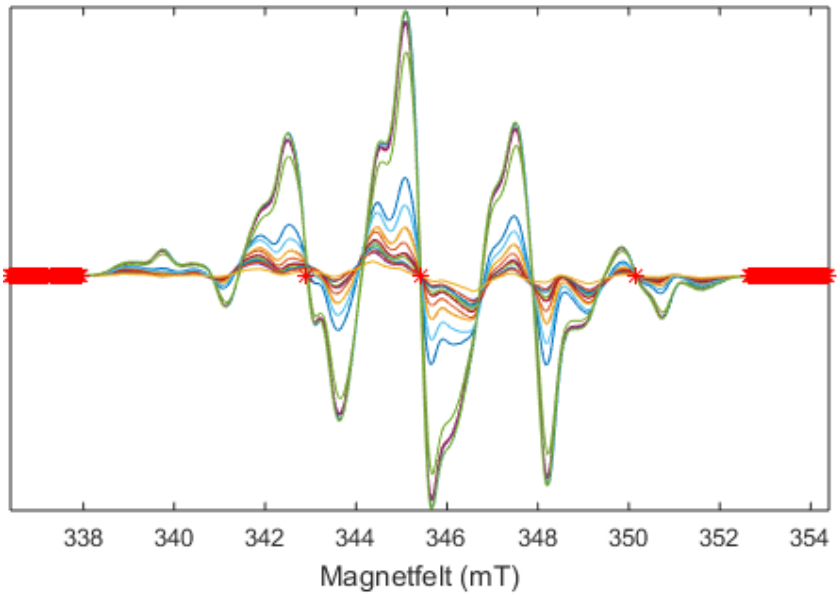
Figur 4-4, resultatene fra PCA med MSC og sentrering som preprosessering av alanin datasettet: (a) Skårene til PC1, PC2 og PC3 mot prøvenummer (tabell 3-1). (b) Skårplott med PC1 mot PC2. (c) Q residualer mot Hotelling T^2 til målingene. (d) Q residualene til ladningene og ladningene til PC1 (e), PC2 (f) og PC3 (g).

Korrelasjon mellom variablene

Korrelasjonladningsplottet (figur 4-5a) over det sentrerte alanin datasettet viser at variablene innenfor den stiplete linjen gir mindre enn 50 % av den forklarte variansen og er derfor mindre viktige for modellen. Disse variablene er de laveste og de høyest magnetfeltene (336-338 mT og 352-354 mT). I figur 4-5b vises kun de viktigste variablene. Figur 4-6 viser hvilke magnetfelt som har blitt fjernet fra figur 4-5b markert med stjerne. Som det kommer fram er det de minste og de største magnetfeltene samt tre plasser hvor spektrene krysser null aksen.

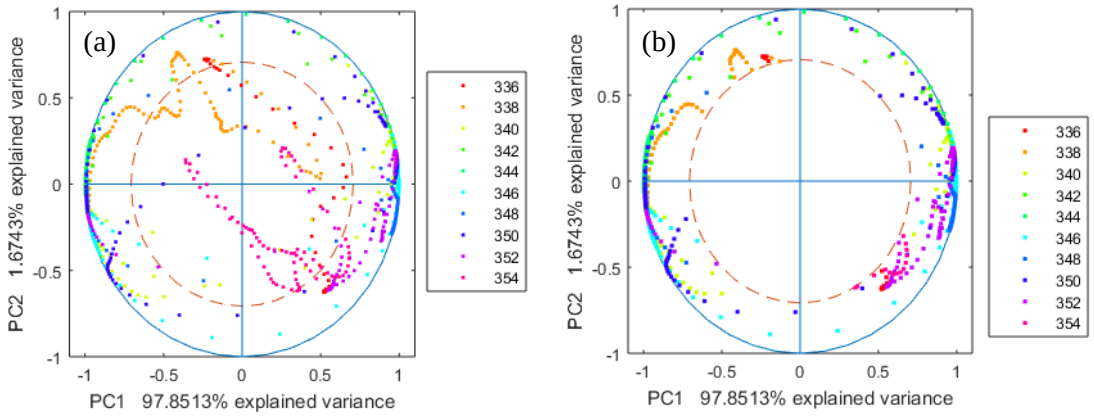


Figur 4-5, korrelasjonladningsplott, med alle variablene (a) og med kun de variablene som er egnet til å finne sammenhenger mellom PC1 og PC2 (b), for det sentrerte alanin datasettet.

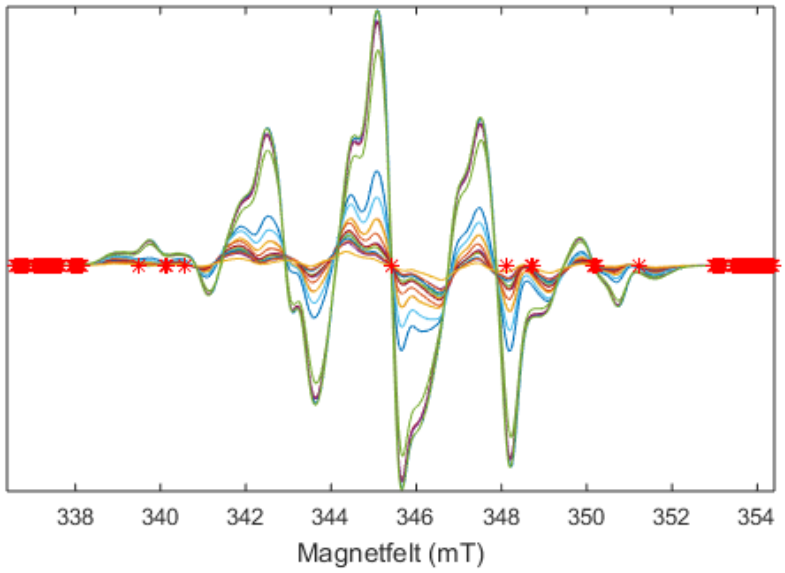


Figur 4-6, variablene som har lav korrelasjon er markert med stjerner, for det sentrerte alanin datasettet.

Korrelasjonladningsplottet figur 4-7a for spektrene preprosessert med MSC, glatting og sentrering, viser ikke en like klar gruppe av variabler rundt origo, sånn som i figur 4-5a. Dette skyldes preprosesseringen MSC, som gjør at spektrene får samme form. Igjen er det de laveste og høyeste magnetfeltene som er minst viktige for modellen (figur 4-8).



Figur 4-7, korrelasjonladningsplott, med alle variablene (a) og med kun de variablene som er egnet til å finne sammenhenger mellom PC1 og PC2 (b), for de MSC og sentrerte alanin datasettet.

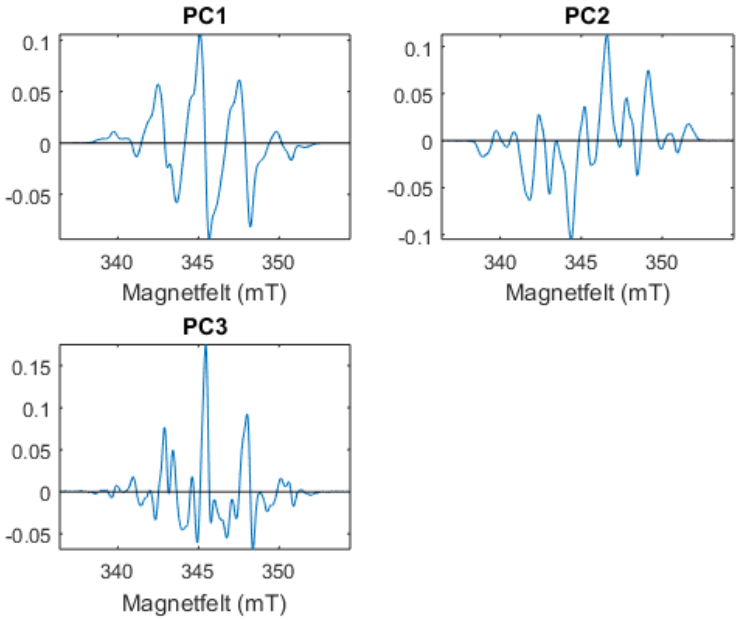


Figur 4-8, variablene som har lav korrelasjon er markert med stjerner, for MSC og sentrerte alanin datasettet.

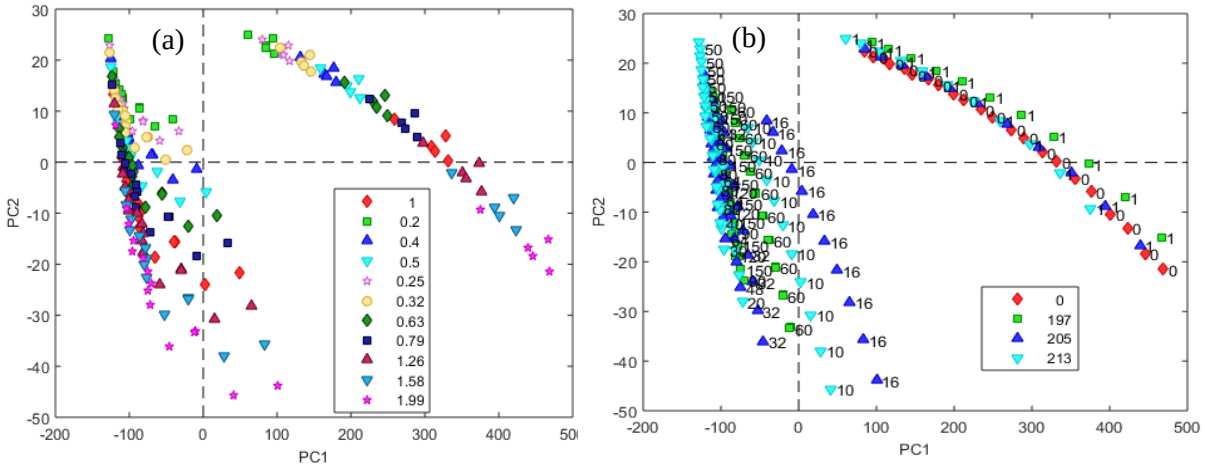
<u>Prinsipalkomponent analyse av alanin spekter ved forskjellige mikrobølgeeffekter</u>

PCA analyse ble gjort for et utvidet alanin datasettet som består av alanin prøvene tatt opp ved forskjellige mikrobølgeeffekter. Ladningene er vist i figur 4-9 og er relativt like ladningene funnet for standard datasettet (figur 4-4b) bortsett for PC2 som ser speilvendt ut. Også for dette utvidete datasettet står PC1 for mest forklart varians (henholdsvis 98,3 %, 0,90% og 0,10% for

PC1, PC2 og PC3). Skårplottene i figur 4-10 viser at skårene fra de samme prøvene tatt med forskjellige mikrobølgeeffekter danner bånd hvor mikrobølgeeffekten avtar med økende PC2 verdi. Igjen danner kontrollene og de varmebehandlete prøvene to adskilte klynger. Dette kan tyde på at det ikke er ny informasjon i prøvene tatt opp ved forskjellig effekt. Derfor blir ikke datasettet med flere mikrobølgeeffekter benyttet noe mer i dette arbeidet.



Figur 4-9, ladningene mot magnetfelt for PC1, PC2 og PC3, for sentrert alanin datasett tatt opp over flere mikrobølgeeffekter.



Figur 4-10, skårplott av PC1 mot PC2 for alanin datasettet tatt opp over flere mikrobølgeeffekter. Etiketten står for (a) mikrobølgeeffekt (mW) og (b) temperatur for varmebehandlingen, tallene ved hvert punkt er oppvarmingstid. PC2 verdien er mest negativ for høyest mikrobølgeeffekt, og positiv for de laveste. Prøvene følger klare bånd fra lavest mikrobølgeeffekt til høyest.

4.1.3 Faktoranalyse (MLCFA)

Resultatene fra faktoranalysen av alanin datasettet preprosessert ved forskjellige metoder, er vist i figur 4-11. Valg av preprosesseringsteknikk endrer ikke de identifiserte faktorene, mest sannsynlig fordi de underliggende faktorene som faktoranalysen finner oppfører seg likt uavhengig av preprosesseringsteknikk.

Tabell 4-4 viser at det er høy korrelasjon mellom faktor F2 og det teoretiske R1-spekteret, samt at det er høy korrelasjon mellom F1 og det teoretiske R2-spekteret.

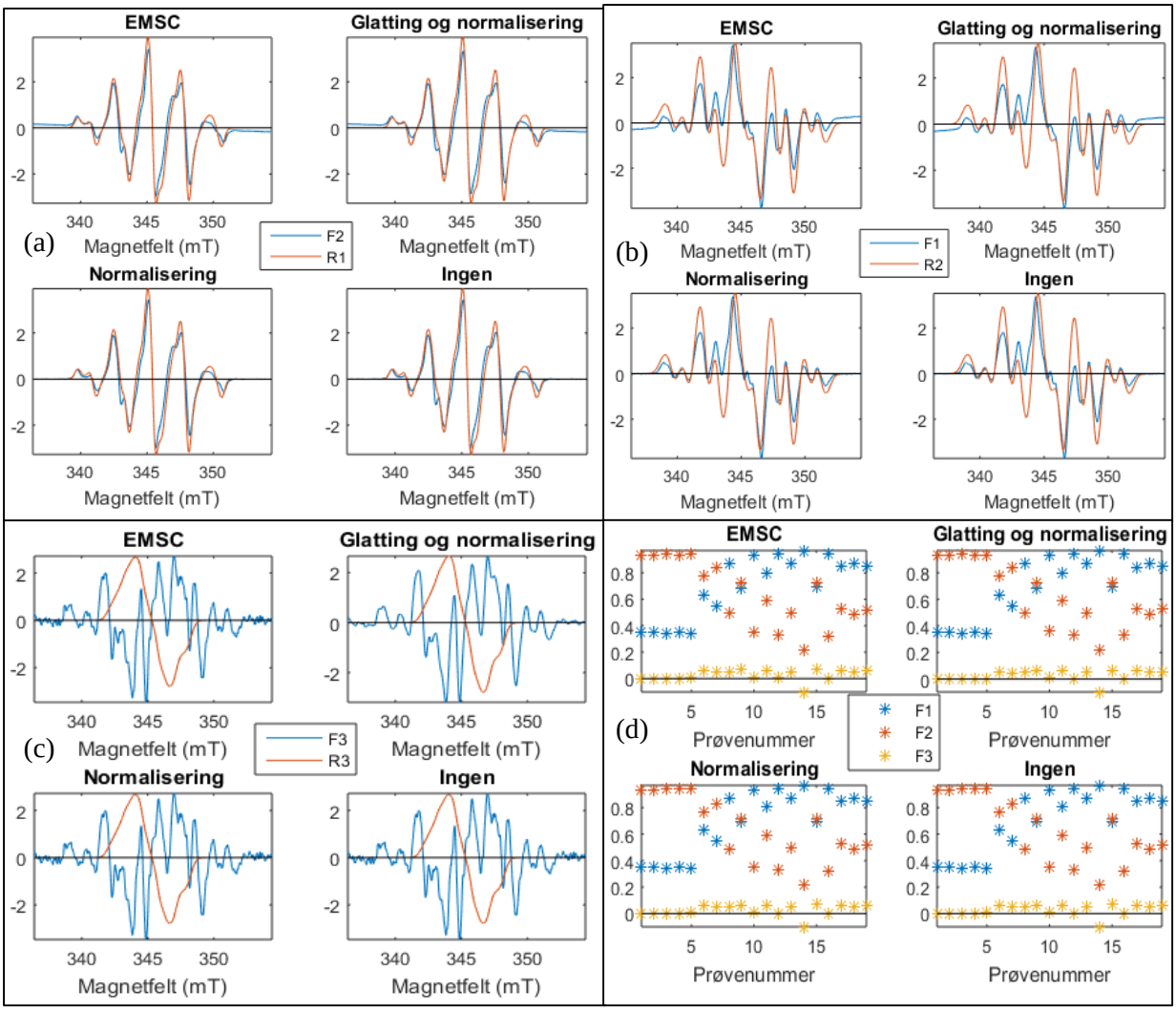
Tabell 4-4, korrelasjonsmatrise mellom faktor 1-3 og mellom det teoretiske R1-, R2- og R3-spekteret (figur 2-4), for alanin datasettet uten preprosessering.

	F1	F2	F3
R1	0,205	0,966	-0,061
R2	0,801	0,342	0,355
R3	0,742	-0,084	-0,583

Figur 4-11c viser faktor 3 som i dette tilfellet er den komponenten som er igjen og således kan være et estimat for R3*, ut fra fasongen til spekteret kan det antas at faktor 3 inneholder en del støy. Skårene til de ulike faktorene er vist i figur 4-11d. Det kommer frem at skårene er like for alle preprosesseringene og at faktor 3 er omtrent null for alle prøvene. Dette støtter opp om hypotesen at faktor 3 er en støy komponent.

Figur 4-12 viser faktor 1-3 fra en redusert faktoranalyse med prøvene som er varmet i 32, 48 og 80 minutter ved 205 ^{o}C og prøvene varmet i 20, 30, 40 og 50 minutter ved 213 ^{o}C. Disse målingene er antatt å inneholde prosentvis mer av radikal R3 enn de andre målingene og dermed kunne gi et estimat for R3. En svakhet med denne modellen er at faktoranalysen gir kun to faktorer, på grunn av at F3 er null for alle prøvene som vist figur 4-12. Derimot kan faktor 1 være en mulig kandidat for radikalet R3. Faktor 2 vil da være en blanding av R1- og R2-radikalspekteret. Korrelasjonsmatrisen mellom F1-2 og R1-3 er vist i tabell 4-5.

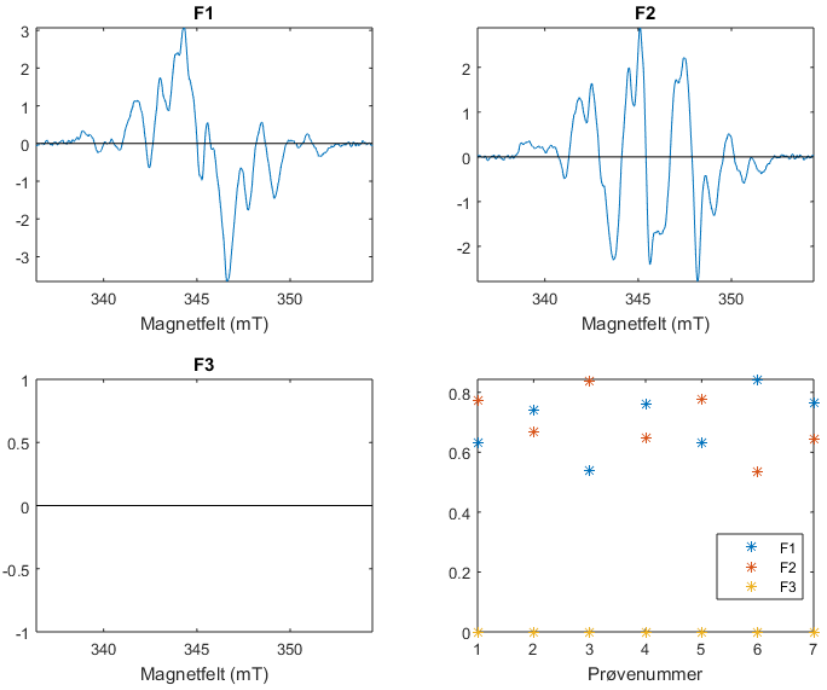
De estimerte R1*, R2* og R3* fra MLCFA analysen ble brukt til å estimere andeler av hver av komponentene i målespektrene (vedlegg 8.5, tabell 8-5). R1* og R2* er valgt som henholdsvis F2 og F1 fra analysen uten preprosessering, R3* ble satt lik F1 fra analysen med de langtids varmebehandlete prøver. R1*, R2* og R3* ble normalisert til enhetsareal og glattet med et gjennomsnittsfilter med henholdsvis størrelse 10 for R1*, 15 for R2* og 100 målepunkter for R3*. Korrelasjonen mellom R1*, R2* og R3* mot basisspektrene for R1, R2 og R3 (figur 2-4) er henholdsvis 97 %, 77% og 89 %. De glattete spektrene er vist i figur 4-13.



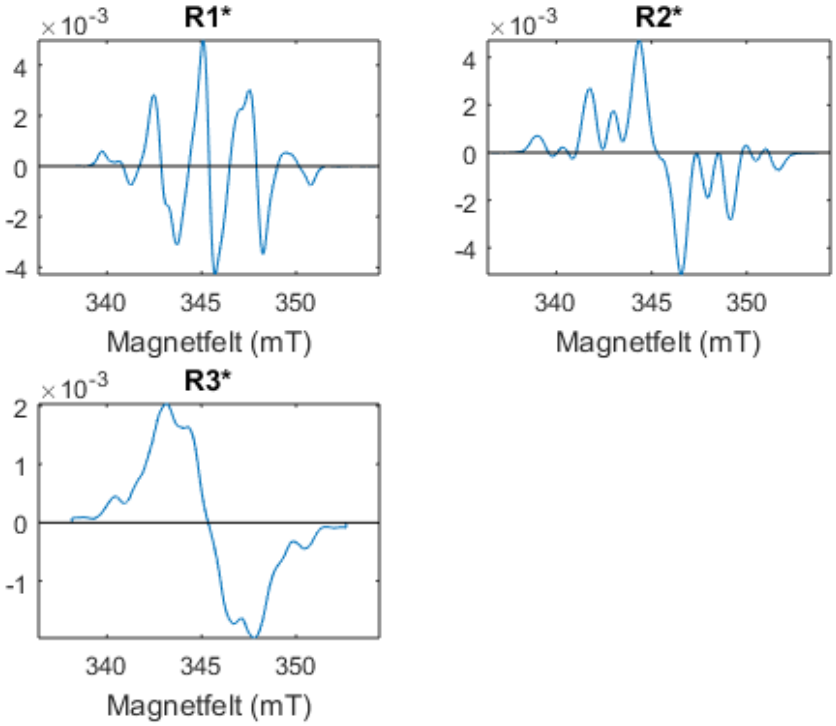
Figur 4-11, resultatene fra faktoranalysen av alanin datasettet med preprosessering: EMSC, glatting og normalisering, normalisering og ingen preprosessering, som angitt. Preprosessering (vist i tittelen) påvirker resultatet i liten grad. (a) Faktor 2 (F2) er relativt lik det teoretiske spekteret til R1 (figur 2-4). Spekteret til R1 er skalert med 3 for å ha samme størrelse som F2. (b) Faktor 1 (F1) og det teoretiske spekteret til R2 (figur 2-4). Spekteret til R2 er skalert med 10 for å ha samme størrelse som F1. (c) Faktor 3 (F3) og (d) skårene som funksjon av prøvenummer (se tabell 3-1).

Tabell 4-5, korrelasjonsmatrise mellom faktor 1 og 2, og mellom det teoretiske R1-, R2- og R3-spekteret (figur 2-4), fra faktoranalyse av prøvene som er varmet i 32, 48 og 80 minutter ved 205 °C og varmet i 20, 30, 40 og 50 minutter ved 213 °C.

	F1	F2
R1	-0,105	0,917
R2	0,526	0,702
R3	0,856	0,035



Figur 4-12, F1-3 fra faktoranalyse av prøvene som er varmet i 32, 48 og 80 minutter ved 205 °C og varmet i 20, 30, 40 og 50 minutter ved 213 °C. Skårene er vist nederst i høyre hjørne og er null for faktor F3. Prøvenummeret er fra kortest tid til lengst oppvarmingstid.



Figur 4-13, approksimerte basisspektere for R1, R2 og R3, glattet med henholdsvis 10, 15 og 100 målepunkters glattefilter, funnet ved faktoranalyse (MLFCA).

4.1.4 Flervariabel kurveoppløsning (MCR)

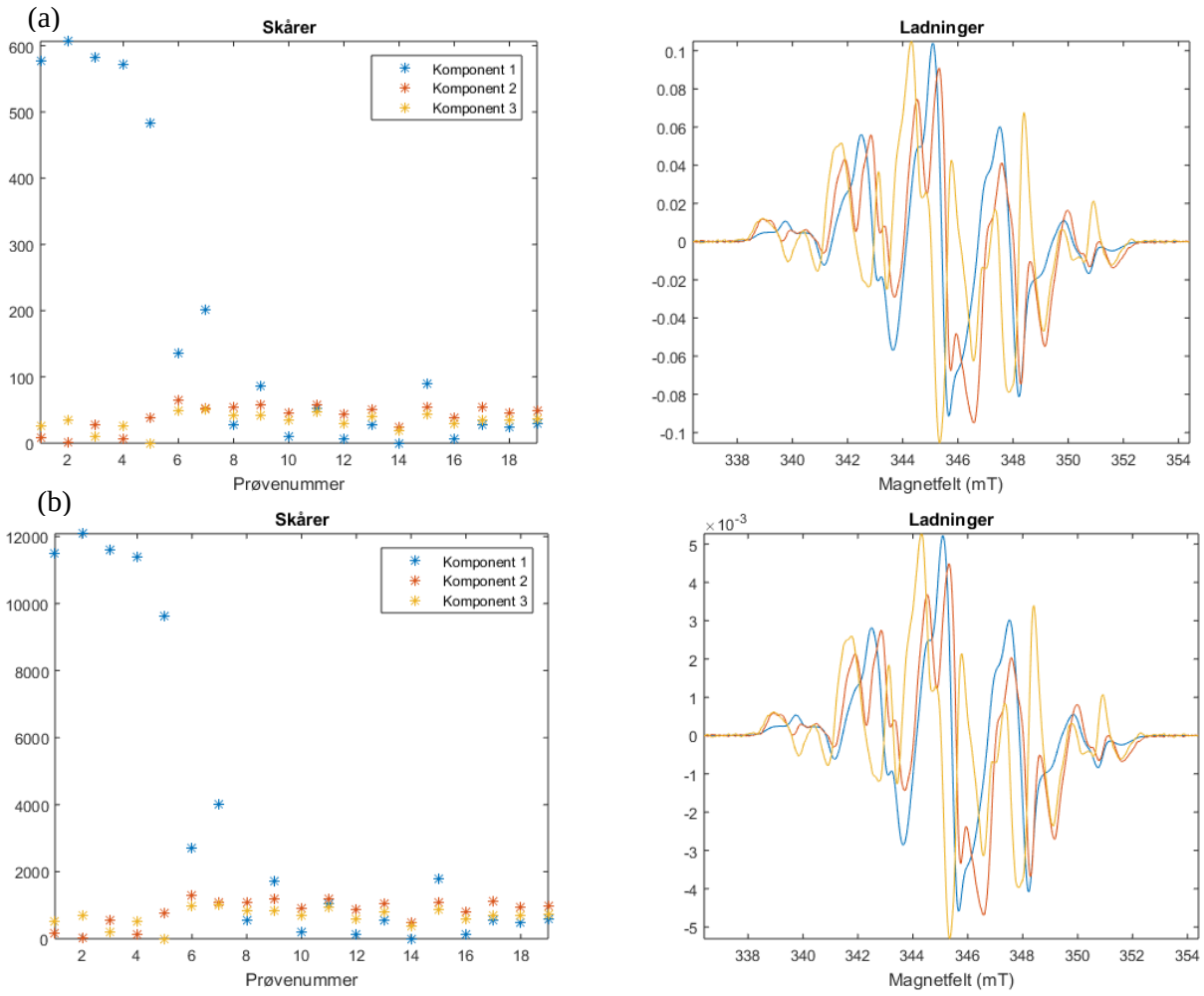
Testing av ulike betingelser

MCR med forskjellige betingelser ble brukt til å finne to til fem komponenter fra alanin datasettet uten preprosessering. Det er lett etter tre spektre bortsett fra der det kommer helt tydelig fram at det er lett etter et annet antall spektre. Standardbetingelsene, samt betingelsene om negative skårer, maksimal kontrast mellom ladninger og skårer, skårer summert til 1 og andre startpunkter gav ikke resultater som gjenspeilet EPR-spektrene eller komponent andelene. Disse resultatene vises i vedlegg 8.3 og brukes ikke videre i oppgaven

Enhetsareal og negative ladninger

Siden EPR-spektrene er første deriverte av absorpsjonsspektrene er det rimelig å anta at ladningene kan være negative, og betingelsen at spektrene kan være negative ble testet. Denne MCR betingelsen gir en liten forandring på skårene (figur 4-14a), spesielt for komponent 1 ved lave *y*-verdier, i forhold til standard betingelsene (figur 8-2a, vedlegg 8.3), mens ladningene har negative verdier og et utseende som ligner mere på EPR-spektre (figur 4-14a) enn ladningene beregnet fra standard betingelsene (figur 8-2a, vedlegg 8.3).

I figur 4-14b vises resultatene av MCR analyse med negative ladninger som er normalisert til enhetsareal. Ladningene (figur 4-14b) er tilnærmet identiske med ladningene funnet med betingelsen negative ladninger (figur 4-14a). Normalisering av ladningene påvirker skaleringen til ladningene, men ikke formen. Dette gjør det enklere å sammenligninger spektrene. Derfor blir normalisering til enhetsareal brukt videre i oppgaven.



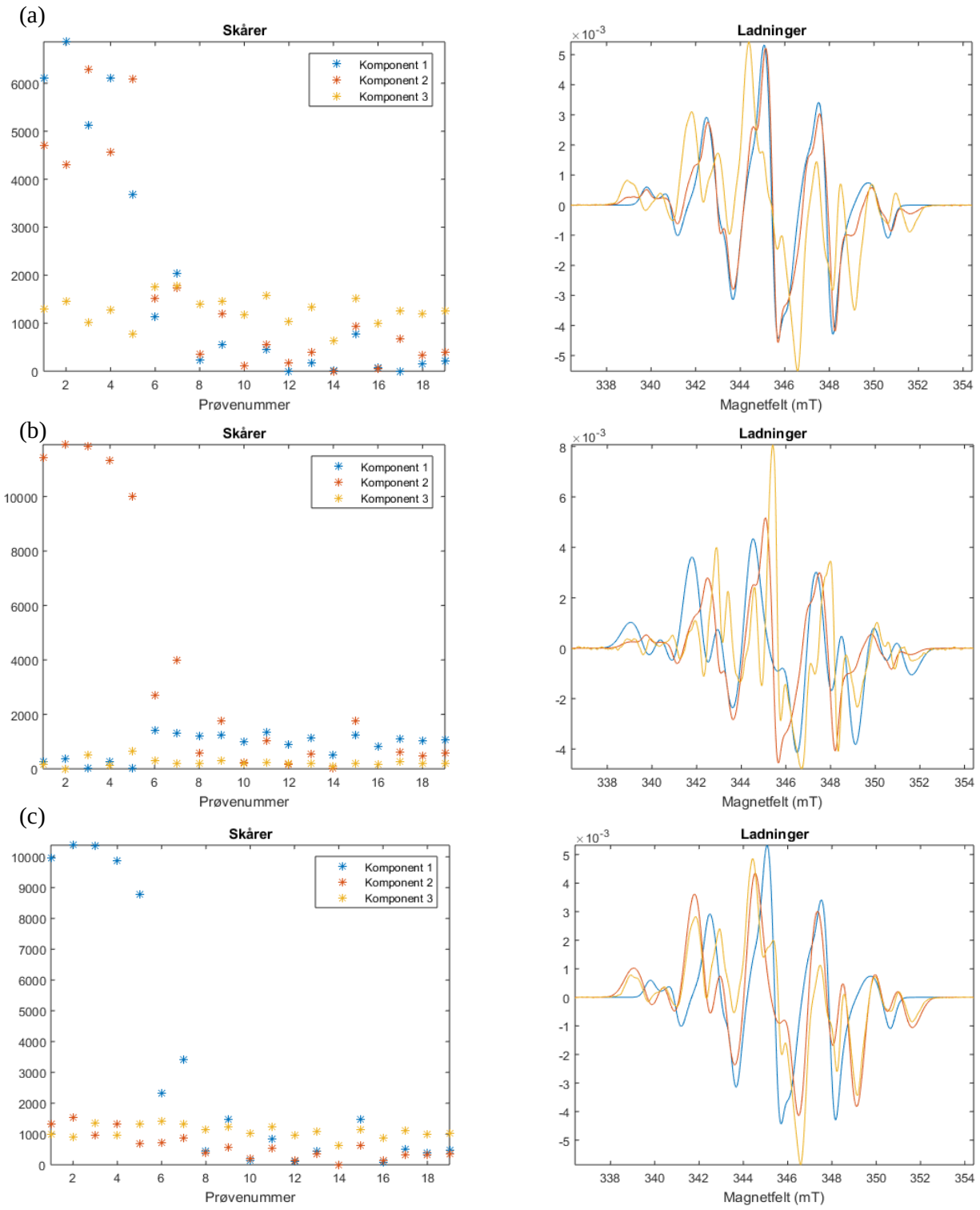
Figur 4-14, resultater fra MCR analysen med (a) negative ladninger og (b) negative ladninger som er normalisert til enhetsareal.

Kjente R1- og R2-spektre

MCR analyse med det teoretiske spekteret til R1 (figur 2-4a) antatt kjent er vist i figur 4-15a. R1 ble normalisert til enhetsareal før den ble brukt som betingelse. Ladningene til MCR komponent 1 og 2 ble likere enn det de var når R1 ikke var antatt kjent (figur 4-14b).

I figur 4-15b er det teoretiske basisspektere til R2 (figur 2-4b) normalisert til enhetsareal og antatt kjent. Komponent 3 får en betydelig høyere sentral topp enn det komponent 3 har ved standard betingelser (figur 8-2a, vedlegg 8.3). For skårene har komponent 1 og komponent 2 (figur 4-15b) byttet plass som den viktigste komponenten i forhold til standard betingelsen.

Figur 4-15c viser resultatene av MCR analysen der de teoretiske basisspektrene til R1 og R2 er normalisert til enhetsareal og antatt kjente. Videre antas det at ladningene kan være negative og at de skal være normalisert til enhetsareal.

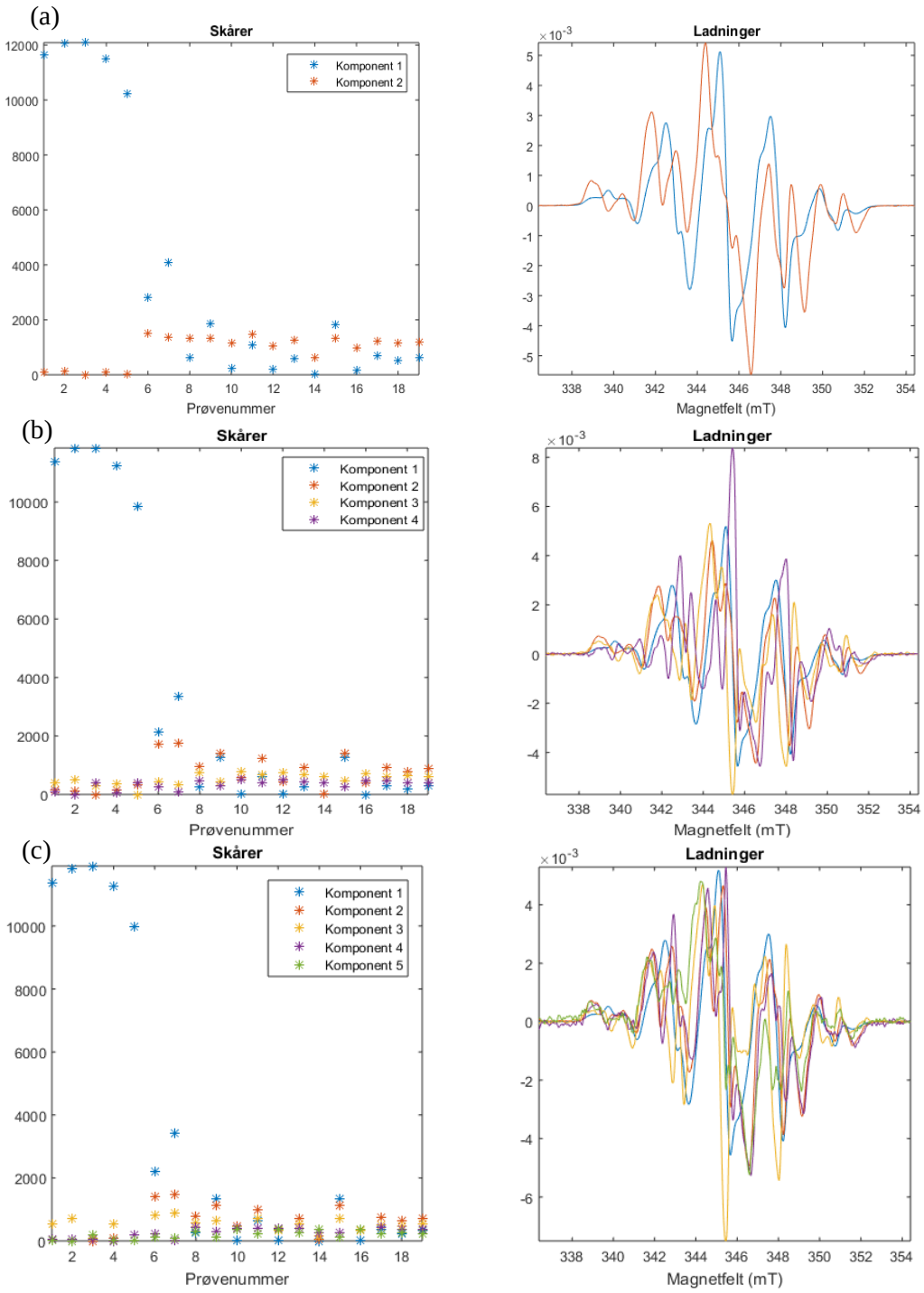


Figur 4-15, resultat av MCR analyse med betingelsen negative ladninger som er normalisert til enhetsareal og det teoretiske basisspektere til (a) R1, (b) R2 og (c) R1 og R2 antatt kjent (figur 2-4).

To, fire eller fem komponenter

MCR analyser hvor det er antatt to, fire eller fem komponenter er vist i figur 4-16. Komponenter fire og fem har skårverdier nesten lik null og kan anses å være støy og brukes dermed ikke

videre. I denne oppgaven er et mål å identifisere tre underliggende komponenter, derfor brukes ikke MCR med kun to komponenter videre.



Figur 4-16, resultater av MCR med (a) to, (b) fire og (c) fem komponenter med betingelsene negative ladninger som er normalisert til enhetsareal.

Oppsummering

Analysene over viser at et valg av tre komponenter er rimelig gitt at komponentene skal ha en betydning knyttet til alanin radikalene, og ikke bare forklare støy i målingene. Betingelsen ladninger normalisert til enhetsareal gir resultater som kan lettere sammenlignes med de målte EPR-spektrene. Betingelsen negative ladninger gir resultater som ligner på EPR-spektrene. Derfor blir disse betingelsen brukt videre i de følgende delkapitlene under.

R1 og R2 er antatt kjent

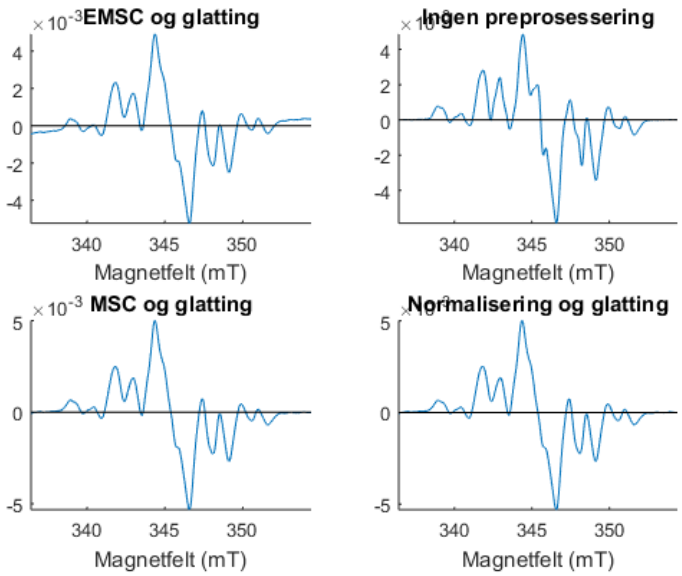
Basert på foregående analyse ble følgende betingelser valgt for videre analyse: Tre komponenter, negative ladninger som er normalisert til enhetsareal og R1 og R2 antatt kjent. Datasettet ble preprosessert på fire måter: ingen preprosessering, EMSC og glatting, MSC og glatting og normalisering og glatting. Glatting ble utført med et 15 målepunkters gjennomsnittsfilter. Tabell 4-6 viser korrelasjonen mellom komponent 1-3 og de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4). Korrelasjonen mellom komponent 1 og R1 og komponent 2 og R2 er 1, siden disse er antatt kjent. Komponent 3 antas å være et estimat på R3*-spekteret og korrelasjonen mellom komponent 3 og R3 er 67-77 % (tabell 4-6).

Preprosessering med EMSC, MSC og normalisering påvirket ikke nevneverdig MCR resultatene. LOF (tabell 4-9 side 82), residualene (figur 4-19), ladningene (figur 4-17) og skårene (figur 4-18) til prøvene samt den estimerte R3*-spekteret (figur 4-17), ved forskjellige preprosesseringer var tilnærmet like. LOF og residualene for tilfellet ingen preprosessering var noe høyere, skårene til prøvene (figur 4-18) var i dette tilfellet også ulike. Skårene (figur 4-18) viser for de preprosesserte modellene at andelen R1 minker, R2 er konstant og R3 øker med økende oppvarmingstid.

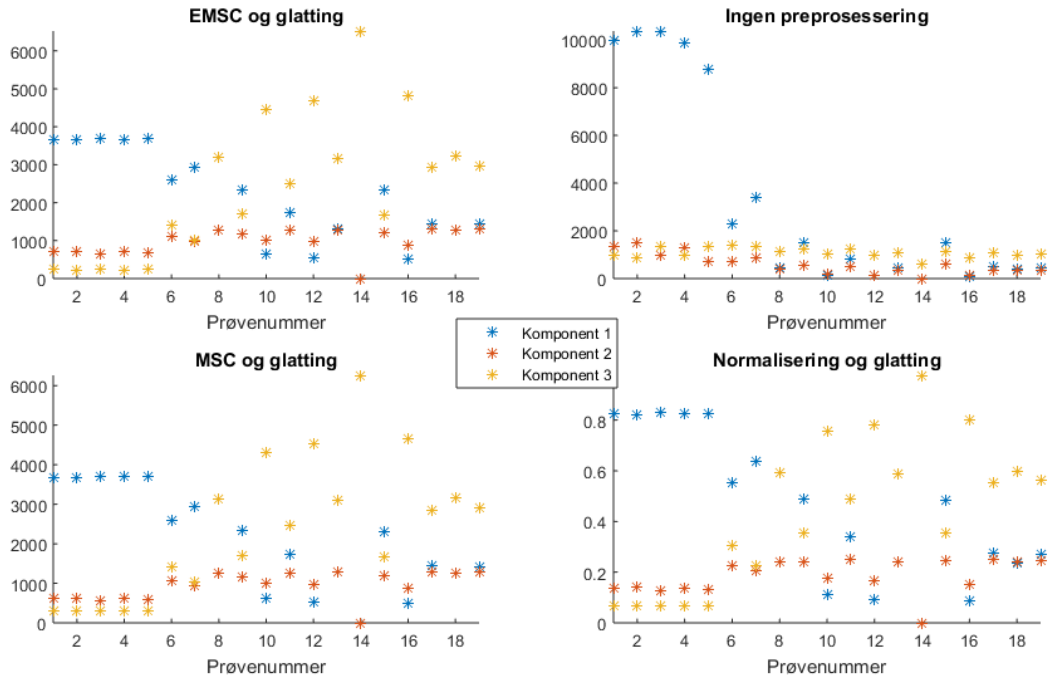
Figur 4-20 viser et kontrollspekter, residualspekteret og en linje for to standardavvik til residualene. Denne linjen viser hvilke og hvor mange residualer det er som har høyere verdi enn 2 standardavvik. Det er få residualer med en så høy verdi. Residualen er høyest der hvor kontrollspekteret har toppunkter.

Tabell 4-6, korrelasjonsmatriser mellom komponent 1-3 og de teoretisk R1, R2 og R3 basisspektere, når R1 og R2 er antatt kjent, for preprosesseringsteknikkene: ingen preprosessering, EMSC, MSC og normalisering, alle med glatting.

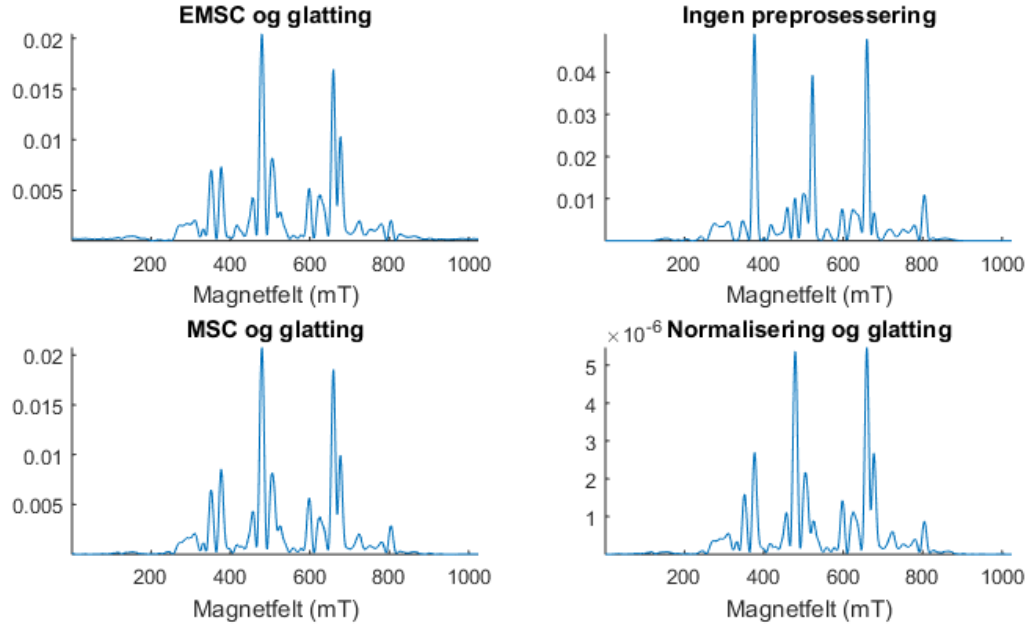
Ingen preprosessering	Komponent 1	Komponent 2	Komponent 3
R1	1	0,472	0,434
R2	0,472	1	0,843
R3	0,106	0,275	0,669
EMSC og glatting			
R1	1	0,472	0,419
R2	0,472	1	0,797
R3	0,106	0,275	0,769
MSC og glatting			
R1	1	0,472	0,418
R2	0,472	1	0,808
R3	0,106	0,275	0,774
Normalisering og glatting			
R1	1	0,472	0,418
R2	0,472	1	0,808
R3	0,106	0,275	0,774



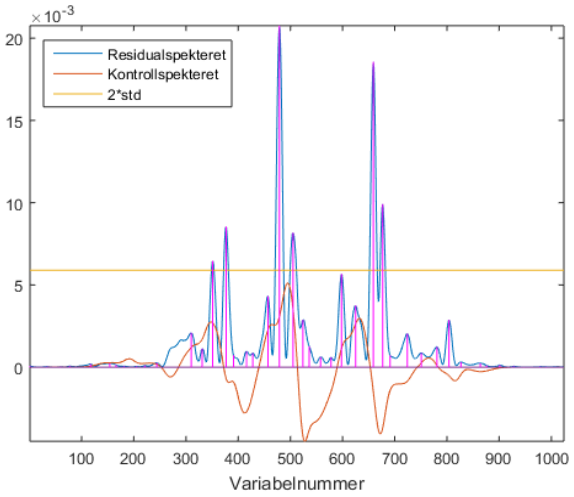
Figur 4-17, ladningene fra MCR av alanin datasettet, titlene angir preprosesseringsteknikk som ble brukt, med R1 og R2 som kjente spektre.



Figur 4-18, skårene fra MCR av alanin datasettet. Titlene angir preprosesseringsteknikk som ble brukt, R1 og R2 er antatt kjente.



Figur 4-19, Q residualen fra MCR analysen for de ulike variablene for de fire ulik preprosesseringsteknikkene. R1 og R2 er antatt kjent. y-aksen er en relativ prosent akse i forhold til størrelsen til den største skårverdien (figur 4-18), for hver preprosesseringsteknikk.

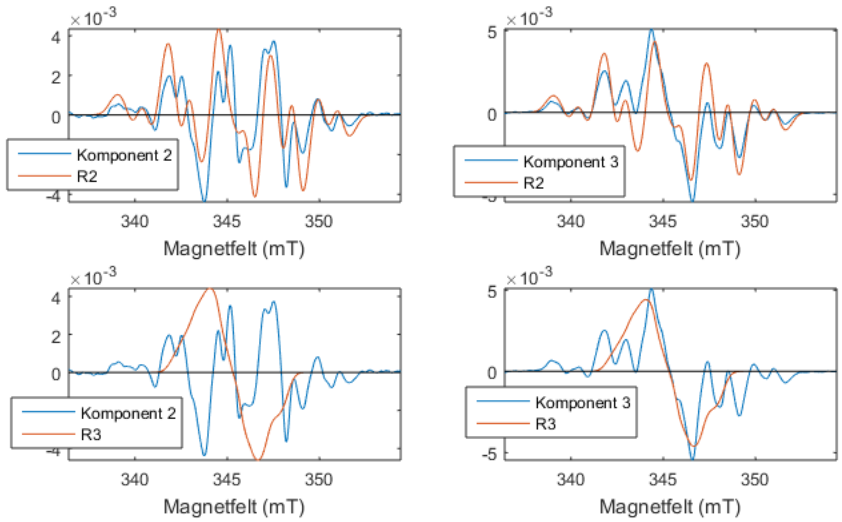


Figur 4-20, residualspekteret, kontrollspekteret og linje for to standardavvik størrelse for residualen, av MCR med R1 og R2 antatt kjent, preprosessering MSC og glatting. Kontrollspekteret er normalisert. y-aksen er en relativ prosent akse i forhold til størrelsen til den største skårverdien (figur 4-18).

<u>R1 er antatt kjent</u>

I denne MCR analysen antas kun R1 kjent, og R2 og R3 estimeres. Korrelasjonsmatrisen (tabell 4-7), LOF (tabell 4-9 side 82), ladningene (vedlegg 8.4 figur 8-5) og skårene (vedlegg 8.4 figur 8-6) viser at de tre preprosesserte modellene igjen er ganske like, og skiller seg fra den ikke preprosesserte modellen. Figur 4-21 viser sammenligning mellom komponent 2 og 3, og R2 og R3. Korrelasjonen mellom R2 og komponent 2 og 3 er henholdsvis 62 % og 80 %. Komponent 3 kan være et estimat både på R2 og R3 siden korrelasjonen er henholdsvis 80 % og 78 %.

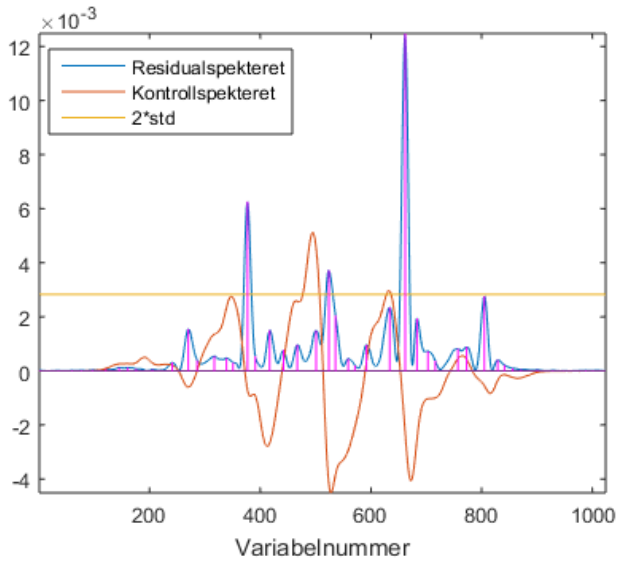
Residualspekteret til MCR modellen med MSC som preprosessering er vist i figur 4-22. De største residualene sammenfaller med nullpunktene kontrollspekteret.



Figur 4-21, sammenligning mellom det teoretiske R2- og R3-spekteret (figur 2-4) og komponent 2 (venstre) og komponent 3 (høyre) fra MCR med MSC som preprosesseringsteknikk og R1 som antatt kjent spekter.

Tabell 4-7, korrelasjonsmatriser mellom komponent 1-3 og de teoretisk R1, R2 og R3 basisspektere (figur 2-4), for preprosesseringsteknikkene: ingen preprosessering, EMSC, MSC og normalisering, alle med glatting, fra MCR av alanin datasettet med R1 antatt kjent.

Ingen preprosessering	Komponent 1	Komponent 2	Komponent 3
R1	1	0,963	0,408
R2	0,472	0,577	0,886
R3	0,106	0,174	0,648
EMSC og glatting			
R1	1	0,833	0,375
R2	0,472	0,629	0,789
R3	0,106	-0,232	0,779
MSC og glatting			
R1	1	0,837	0,373
R2	0,472	0,624	0,800
R3	0,106	-0,236	0,784
Normalisering og glatting			
R1	1	0,853	0,363
R2	0,472	0,610	0,798
R3	0,106	-0,229	0,787



Figur 4-22, residualspekteret, et kontrollspekteret og linjen for to standardavvik størrelse på residualen, med MSC som preprosesseringsteknikk, av MCR analysen med R1 antatt kjent. Kontrollspekteret har blitt normalisert. y-aksen er en relativ prosent akse i forhold til størrelsen til den største skårverdien (vedlegg 8.4 figur 8-6).

Ingen spektre antatt kjent

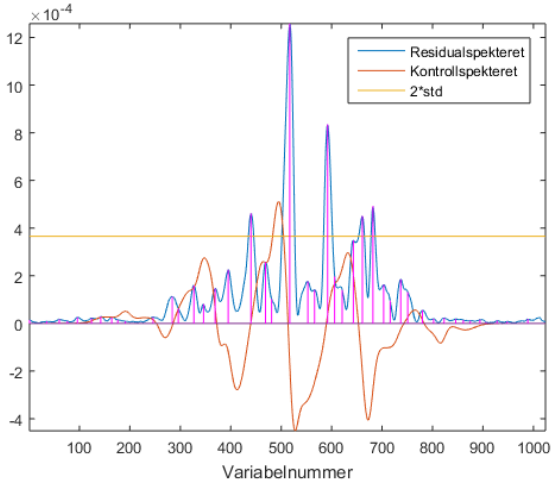
Her antas ingen spektre kjent og R1, R2 og R3 estimeres med tilsvarende MCR analyse og preprosesseringer som over. Korrelasjonsmatrisen (tabell 4-8), LOF (tabell 4-9 side 82) ladningene (vedlegg 8.4 figur 8-7) og skårene (vedlegg 8.4 figur 8-8) viser at resultatet er likt for de preprosesserte modellene, bortsett fra for normalisering. Modellen med ingen preprosessering skiller seg fra de preprosesserte ved at denne gir kun en komponent med høye skårverdier (vedlegg 8.4 figur 8-8), hvor det bare er komponent 1 som har en skårverdi som er høy. Her har komponent 1 og 2 byttet rekkefølge sammenlignet med MSC og EMSC. Residualspekteret (figur 4-23) viser at den største residualen oppstår når kontrollspekteret har et nullpunkt.

Tabell 4-8, korrelasjonsmatriser mellom komponent 1-3 og de teoretisk R1, R2 og R3 basisspektere (figur 2-4), for preprosesseringsteknikkene: ingen preprosessering, EMSC, MSC og normalisering, alle med glatting. For alanin datasettet uten noen spektre kjent.

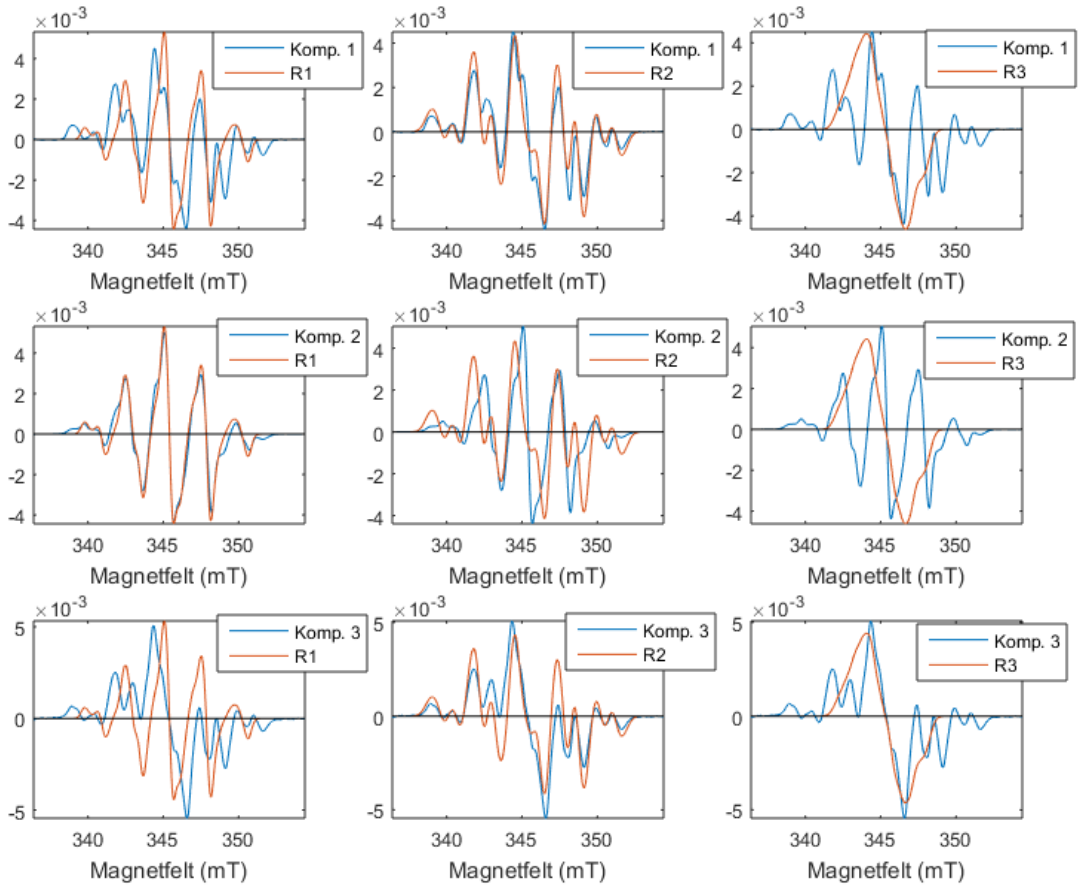
Ingen preprosessering	Komponent 1	Komponent 2	Komponent 3
R1	0,983	0,640	-0,120
R2	0,572	0,735	0,555
R3	0,149	0,489	0,467
EMSC og glatting			
R1	0,660	0,979	0,378
R2	0,905	0,599	0,786
R3	0,510	0,170	0,781
MSC og glatting			
R1	0,656	0,981	0,377
R2	0,909	0,594	0,797
R3	0,517	0,164	0,787
Normalisering og glatting			
R1	0,981	0,661	0,369
R2	0,594	0,908	0,795
R3	0,165	0,516	0,790

Komponent 1 er et estimat for R1* for ingen preprosessering og normalisering, mens komponent 2 er et estimat for R1* for EMSC og MSC. Figur 4-24 viser de estimerte

komponentene 1-3, med preprosessering MSC, vist mot basisspektrene til R1, R2 og R3. Det kommer fram at komponent 2 og R1 er like (korrelasjon 98%), komponent 1 og 3 er ganske like med R2 (korrelasjon 91 % og 80 %) og komponent 3 og R3 er ganske like (korrelasjon 79%).



Figur 4-23, residualspekteret sammen med et kontrollspekter og to standardavviksgrense, fra MCR med MSC som preprosessering og ingen antatt kjente spektre. Kontrollspekteret er normalisert og skalert med 0,1. y-aksen er en relativ prosent akse i forhold til størrelsen til den største skårverdien (vedlegg 8.4 figur 8-8).



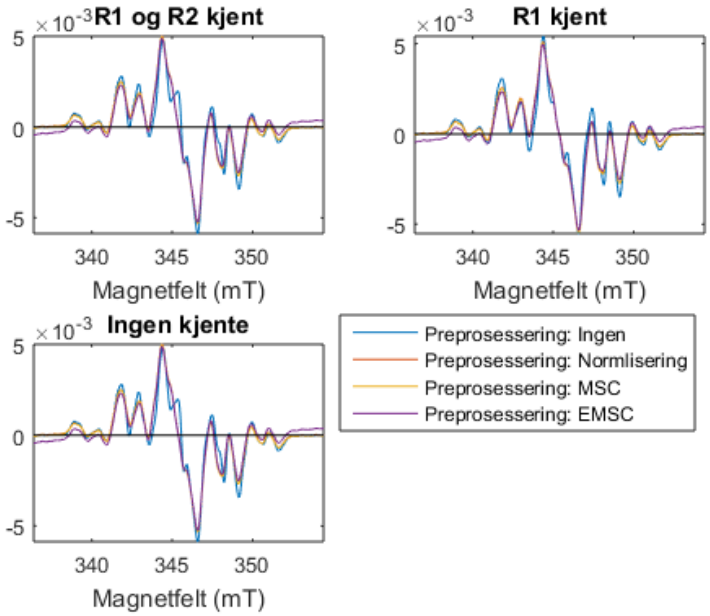
Figur 4-24, sammenligning mellom de estimerte komponentene 1-3 fra MCR analyse med MSC som preprosessering og ingen kjente spektre, mot de teoretiske basisspektere for R1, R2 og R3 (figur 2-4).

Oppsummering MCR

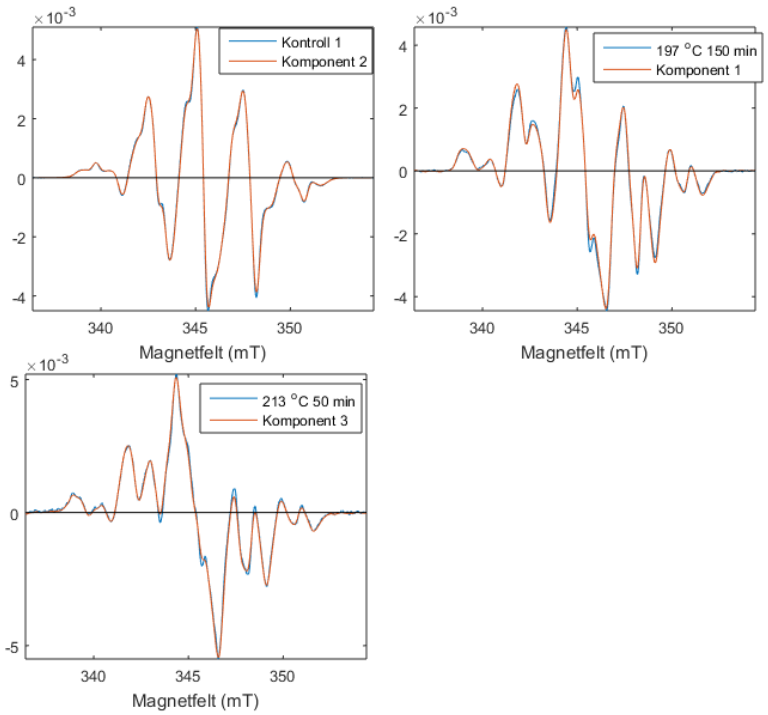
De ulike estimatene av R3-spekteret er vist i figur 4-25. Disse estimerte R3* er ganske like uavhengig av hvor mange komponenter som er antatt kjent i modellen og preprosessering som er brukt.

I figur 4-26 er spektrene fra MCR analysen med betingelsen ingen kjente spektre og MSC som preprosessering vist. EPR-spekteret til prøven kontroll 1 er ganske lik komponent 2 (korrelasjon på 0,997), spekteret til prøven 197 ^{o}C 150 minutter er lik komponent 1 (korrelasjon 0,999) og spekteret til prøven 213 ^{o}C 50 minutter er lik komponent 3 (korrelasjon 0,998). Dette viser at komponentene fra en MCR analyse er like med enkelte av spektrene som er inkludert i analysen. Dette er en svakhet med MCR analysene.

LOF avtar når færre komponenter er antatt kjent i modellen (tabell 4-9). Dette kommer av at med færre kjente spektre blir det flere estimerte spektre som kan approksimere variansen i datasettet. Med R1 og R2 antatt kjent vil det kun være komponent 3 som kan forklare variansen som ikke er beskrevet med R1 og R2. Det er ikke sikkert at den modellen som har lavest LOF har den sanne R3, da modellene med lav LOF har dårligere tilpasning av R1 og R2 enn modellene med høy LOF.



Figur 4-25, de estimerte R3 spektrene fra MCR analyse med enten R1 og R2 kjent, R1 kjent eller ingen kjente spektre, med ulike preprosesseringer.*



Figur 4-26, sammenligninger av spektrene for komponent 2 og kontroll 1, komponent 1 og 197 ºC 150 minutter og komponent 3 og 213 ºC 50 minutter fra MCR analyse med betingelse ingen kjente spekter og MSC som preprosesseringsteknikk.

Tabell 4-9, sammendrag av LOF av de ulike MCR modellene med null til to kjente spektre. LOF er beregnet fra tilpasning av ladningene og skårene til hvert enkelt målespekter.

	Ingen	Norm	MSC	EMSC
R1 og R2 kjent	0,131	0,098	0,092	0,097
R1 kjent	0,094	0,067	0,061	0,069
Ingen kjente	0,013	0,025	0,024	0,024

4.1.5 Selvmodellerings miksturanalyse (SMA)

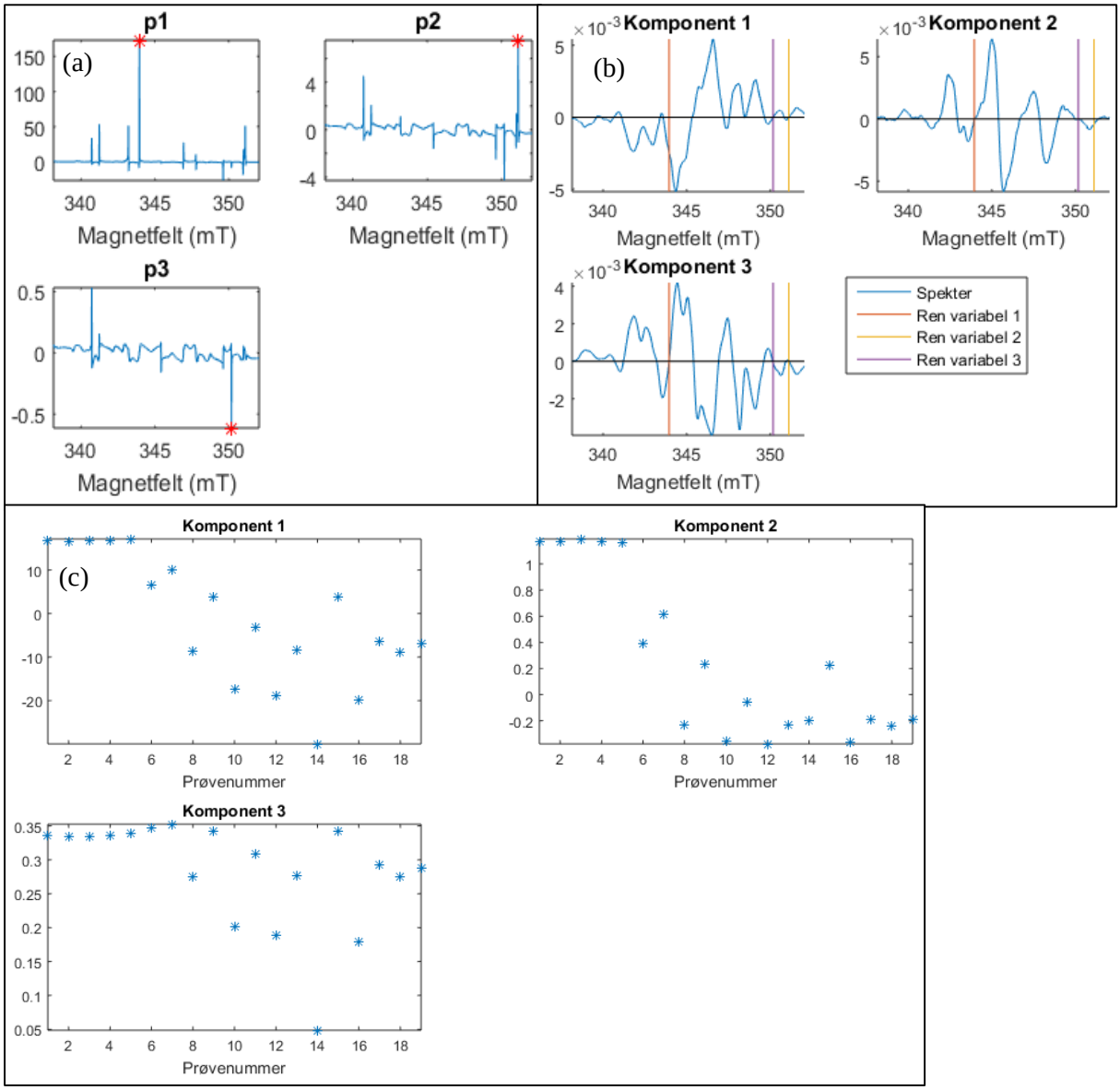
Normerte EPR-spektre

Figur 4-27a viser renhetsvektorene for SMA analysen basert på normerte spektre, de reneste variablene er vist med stjerne, dette er 342,206 mT, 348,435 mT og 349,368 mT. Komponentspektrene er vist i figur 4-27b. Kun ren variabel 1, har en verdi forskjellig fra null for komponent 1. Det samme gjelder for komponent 2 og ren variabel 2 og for komponent 3 og ren variabel 3. De tilhørende skårene er vist i figur 4-27c mot prøvenummer (dvs. oppvarmingstid), skårene minker for komponent 1 og 2, og er ganske konstante for komponent 3. Korrelasjonsmatrisen mellom komponent 1-3 og de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4) er vist i tabell 4-10. Korrelasjonen mellom komponent 2 og R1 og komponent 3

og R2 ganske bra (0,95 og 0,86). Komponent 1 vil derfor være et estimat for R3* (korrelasjon 0,78).

Tabell 4-10, korrelasjonsmatrise, mellom komponent 1-3 og de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4) for SMA analysen basert på normerte EPR-spektre.

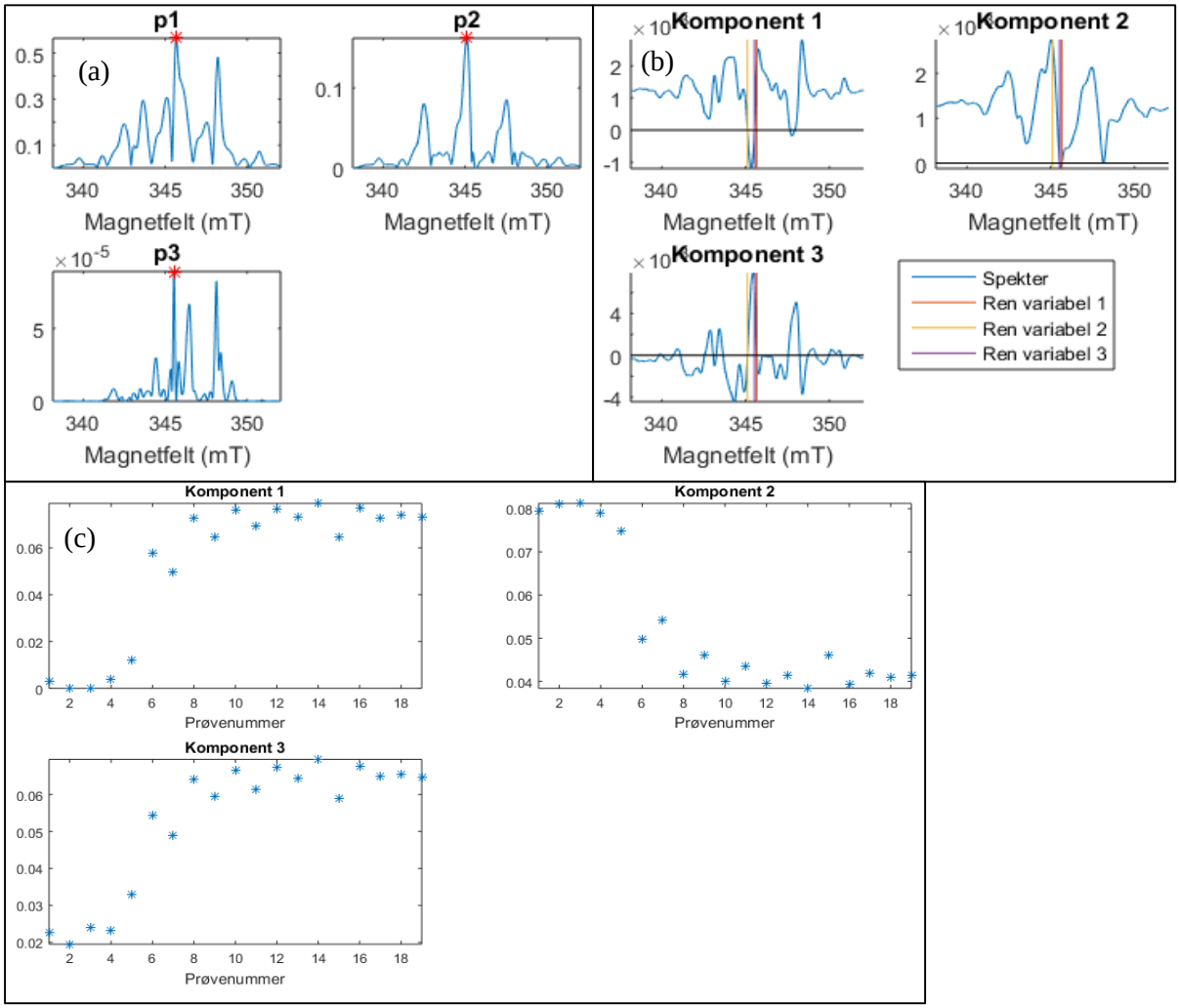
	Komponent 1	Komponent 2	Komponent 3
R1	-0,422	0,950	0,766
R2	-0,791	0,378	0,863
R3	-0,784	0,282	0,466



Figur 4-27, resultatene fra SMA basert på normaliserte EPR-spektre. (a) Renhetsvektorene, hvor de rene variablene er vist med stjerne. (b) Komponentspektrene, med vertikalstreker ved de rene variablene. Kun komponent 1 har en verdi ulik null for den ren variabelen 1 og (c) skårene som funksjon av prøvenummer (tabell 3-1).

EPR-spektre forskjøvet til positive verdier

Renhetsvektorene **p**1, **p**2 og **p**3 er vist i figur 4-28a. Verdiene til **p**3 på *y*-aksen er veldig mye mindre enn *y*-verdiene til **p**1 og **p**2. Dette betyr at den tredje komponenten inneholder mer støy enn de to andre komponentene. De rene variablene er 343,368 mT, 343,790 mT og 343,931 mT. Disse er nærme hverandre noe som betyr at komponentspektrene har ganske stor variasjon rundt disse verdiene. Komponentspektrene 1-3 er vist i figur 4-28b. Komponent 1 og 2 har få ikke negative verdier, mens komponent 3 svinger rundt null. Skårene er vist i figur 4-28c og øker med prøvenummer (dvs. oppvarmingstid) for komponentene 1 og 3 og avtar for komponent 2. Korrelasjonsmatrisen mellom komponentspektrene 1-3 og de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4) er vist i tabell 4-11. Kun komponent 2 korrelerer godt med R1. Ingen av komponentene korrelerer med R2 og R3. Derfor vil det her være vanskelig å konkludere med et estimat for R3.



Figur 4-28, resultatene fra SMA EPR spekteret forskjøvet til positive verdier. (a) Renhetsvektorene, markert med rød stjerne for de rene variablene, (b) komponentspektrene, i de rene variablene er vist med en rett streker og (c) skårene som funksjon av prøvenummeret (tabell 3-1).

Tabell 4-11, korrelasjonsmatrise, mellom komponent 1-3 og de teoretiske basisspektrene til R1, R2 og R3 (figur 2-4). For SMA analysen basert på EPR-spektre forskjøvet til positive verdier.

	Komponent 1	Komponent 2	Komponent 3
R1	-0,504	0,949	-0,136
R2	0,035	0,648	-0,376
R3	0,121	0,202	-0,167

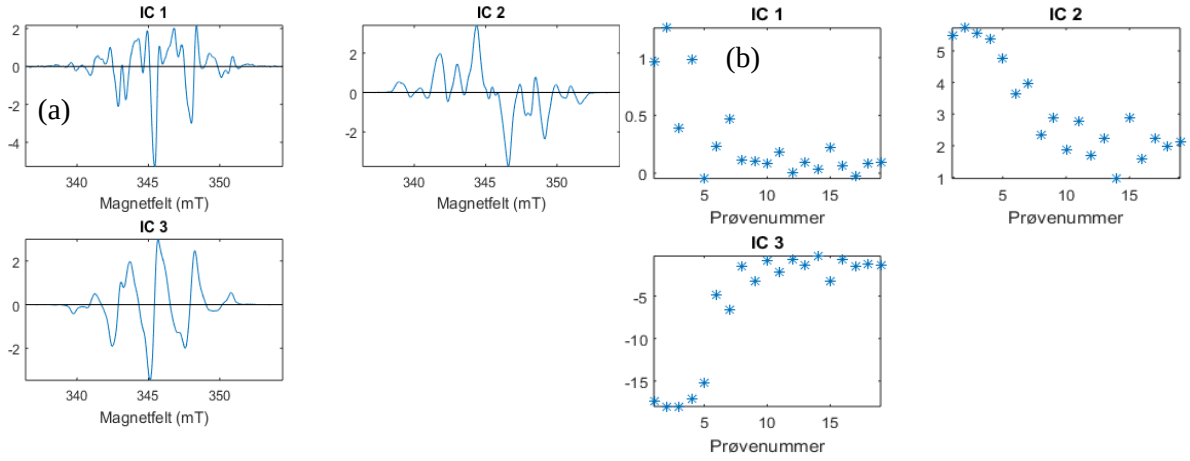
4.1.6 Uavhengig komponent analyse (ICA)

ICA ble utført på alanin datasettet uten preprosessering av EPR-spektrene, med standard betingelser gitt i FastICA algoritmen. FastICA algoritmen kjørt med tre komponenter klarte ikke å reprodusere de samme uavhengige komponentene i samme rekkefølge og skalering. Dette kan komme av at det kun er en egenverdi som har betydning for datasettet og at det har blitt funnet forskjellige rotasjonsmatriser hver gang.

Figur 4-29 viser tre uavhengige komponentene som ble funnet, disse komponentene har en korrelasjonsmatrise, som er vist i tabell 4-12. Korrelasjonsmatrisen viser at IC 3 skalert med -1, kan være estimat for R1*, IC 2 kan være et estimat for R2*, dette medfører at det er vanskelig å finne et godt estimat på R3*.

Tabell 4-12, korrelasjonsmatrise mellom IC 1-3 og teoretiske basisspektere til R1, R2 og R3 (figur 2-4).

	IC 1	IC 2	IC 3
R1	0,058	0,152	-0,975
R2	0,099	0,812	-0,371
R3	-0,066	0,684	0,028

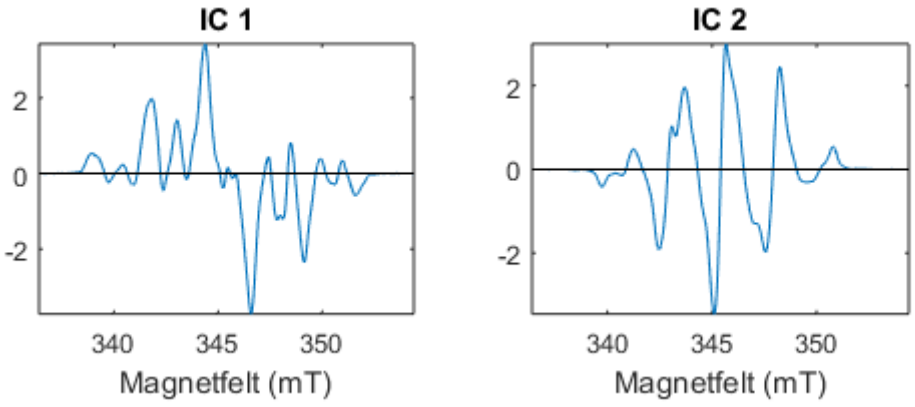


Figur 4-29, Resultatene fra FastICA med tre komponenter, uten preprosessering av alanin datasettet, (a) komponentenespektrene og (b) skårene mot prøvenummer (tabell 3-1).

ICA med to komponenter

FastICA kjørt med to komponenter er vist i figur 4-30, IC1 er lik IC2 fra figur 4-29a, og IC2 er lik IC3 fra figur 4-29a. Dette betyr at disse to komponentene er de viktigste i komponentene i de målte alanin spektrene.

For ICA med to komponenter er algoritmen stabil og samme resultat kommer ved flere kjøringer.



Figur 4-30, de to uavhengige spekter komponentene funnet ved FastICA av alanin datasettet.

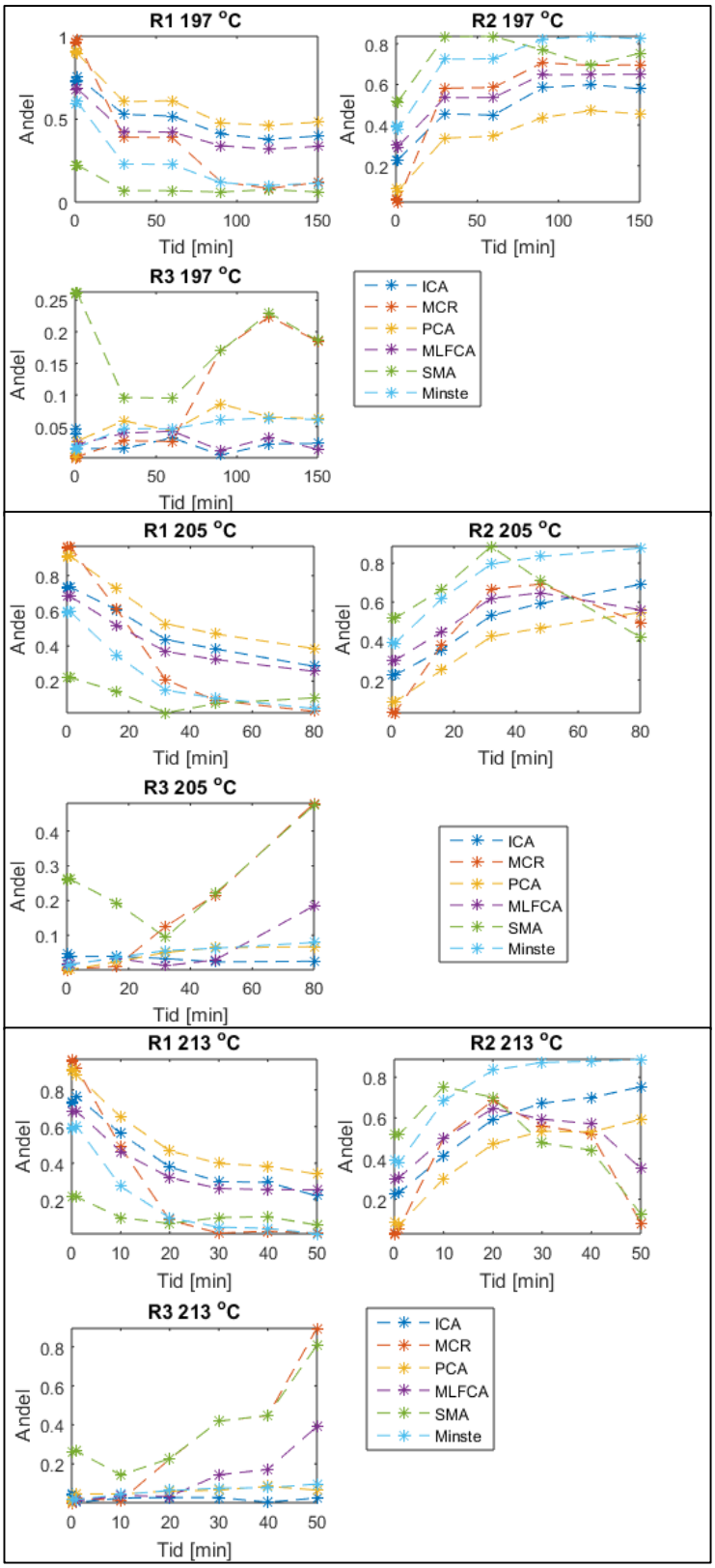
4.1.7 Andeler av de ulike komponentene

For metodene i kapittel 4.1.1-4.1.6 er det blitt funnet andeler av de estimerte R1*-, R2*- og R3*-radikalene i målespektrene (vedlegg 8.5). Figur 4-31 viser andelene av komponent R1*, R2* og R3* for henholdsvis 197 ^{o}C, 205 ^{o}C og 213 ^{o}C. Modellene som er vist er (1) MCR med MSC preprosessering og ingen kjente spektre (kapittel 4.1.4, figur 4-24), (2) ICA med tre komponenter (kapittel 4.1.6, figur 4-29a), (3) residualen til de fem EPR-spektrene antatt å inneholde mest R3 funnet ved minste kvadraters metode (kapittel 4.1.1, figur 4-1b0), (4) komponentenespektrene med høyest korrelasjon funnet ved MLCFA (kapittel 4.1.3, figur 4-13), (5) PCA med sentrerte spektre (kapittel 4.1.2, figur 4-3e-g) og (6) SMA basert på normerte spektre (kapittel 4.1.5, figur 4-27b). De forskjellige metodene estimerer forskjellige andeler av komponentene for de samme prøvene. For eksempel, vil andelen R1 for 197 ^{o}C og kontrollene variere mellom 97 % for MCR og 22 % for SMA. Utviklingen er lik for alle temperaturene og modellene, ved at andelen R1 avtar med økende oppvarmingstid, R2 øker med økende oppvarmingstid og R3 øker for metodene MCR, SMA og MLCFA, og er relativt stabil med økende oppvarmingstid for de andre metodene. Andelen R1 avtar raskere for 213 ^{o}C enn andelen for 197 ^{o}C, dette stemmer med kjent teori [9].

(a)

(b)

(c)



Figur 4-31, andelene av de estimerte R1, R2 og R3 for (a) 197 °C, (b) 205 °C og (c) 213 °C oppvarming, for metodene (1) MCR med MSC preprosessering og ingen kjente spektre, (2) ICA med tre komponenter, (3) residualen til de fem EPR-spektrene antatt å inneholde mest R3, (4) komponentenespektrene med høyest korrelasjon MLCFA, (5) PCA med sentrerte spektre og (6) SMA basert på normerte EPR-spektre.

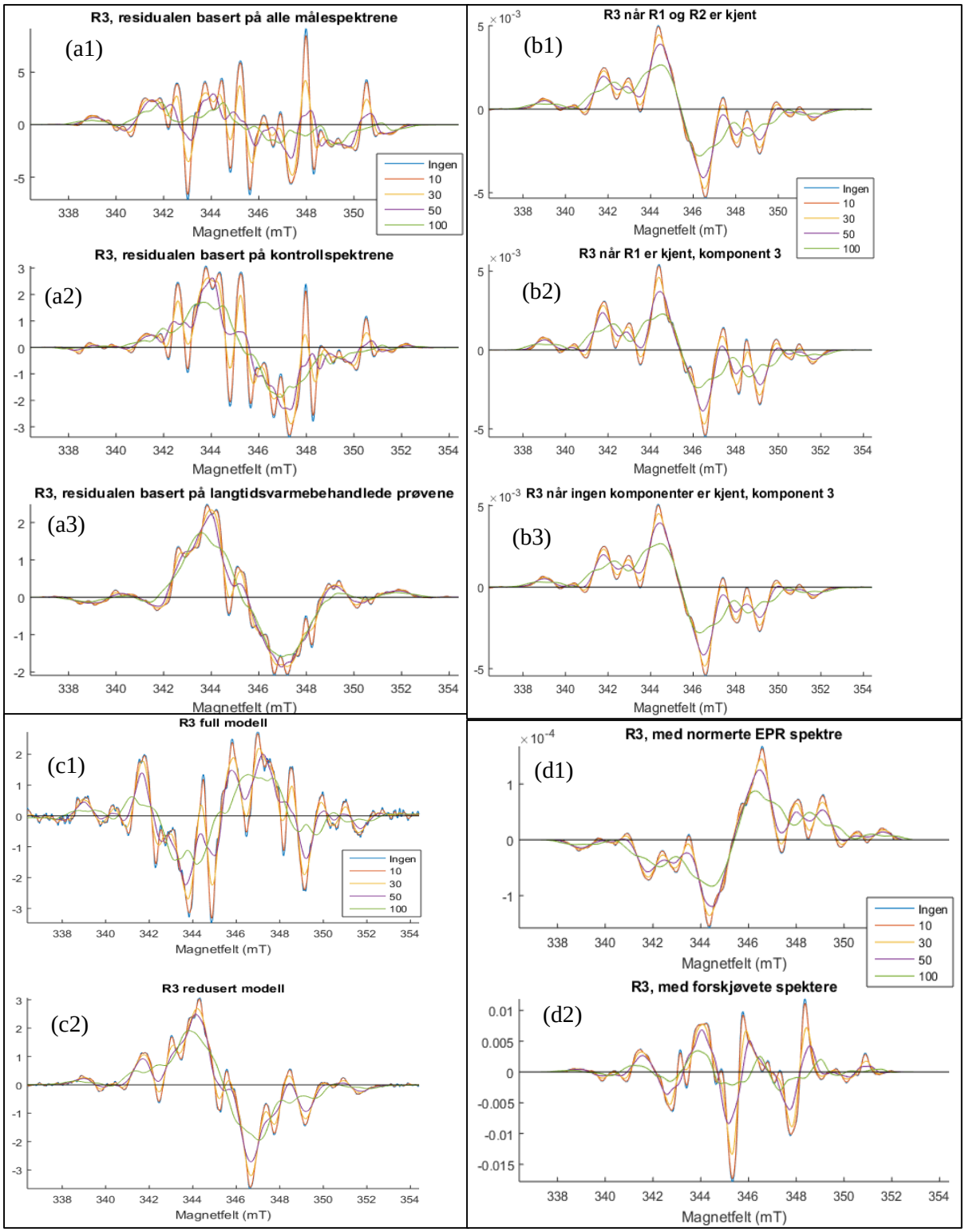
4.1.8 Estimater av R3.

I dette kapittelet blir den teoretiske basis spekteret til R3 (figur 2-4c) sammenlignet med de estimerte R3*-spektrene funnet fra de seks metodene i kapitel 4.1.1-4.1.6. Alle estimater på R3*-spekteret har blitt postprosessert med et gjennomsnittsglattefilter med bredde på 0, 10, 30, 50 og 100 målepunkter. R3* estimatet postprosessert med et 100 målepunkters glattefilter skiller seg fra de andre R3* estimatene, samtidig som de ligner mer på den teoretiske R3. Dette skyldes at den teoretiske R3-spekteret er mye glattere enn samtlige av de estimerte R3*-spektrene.

Estimater av R3 med høy korrelasjon

De statistiske metodene som har resultert i et estimat på R3*-spekteret som ligner på den teoretiske R3-spekteret (korrelasjon >70 %) er vist i figur 4-32. Tilhørende korrelasjon, LOF og R^2-tabell er vist i vedlegg 8.6, tabell 8-12.

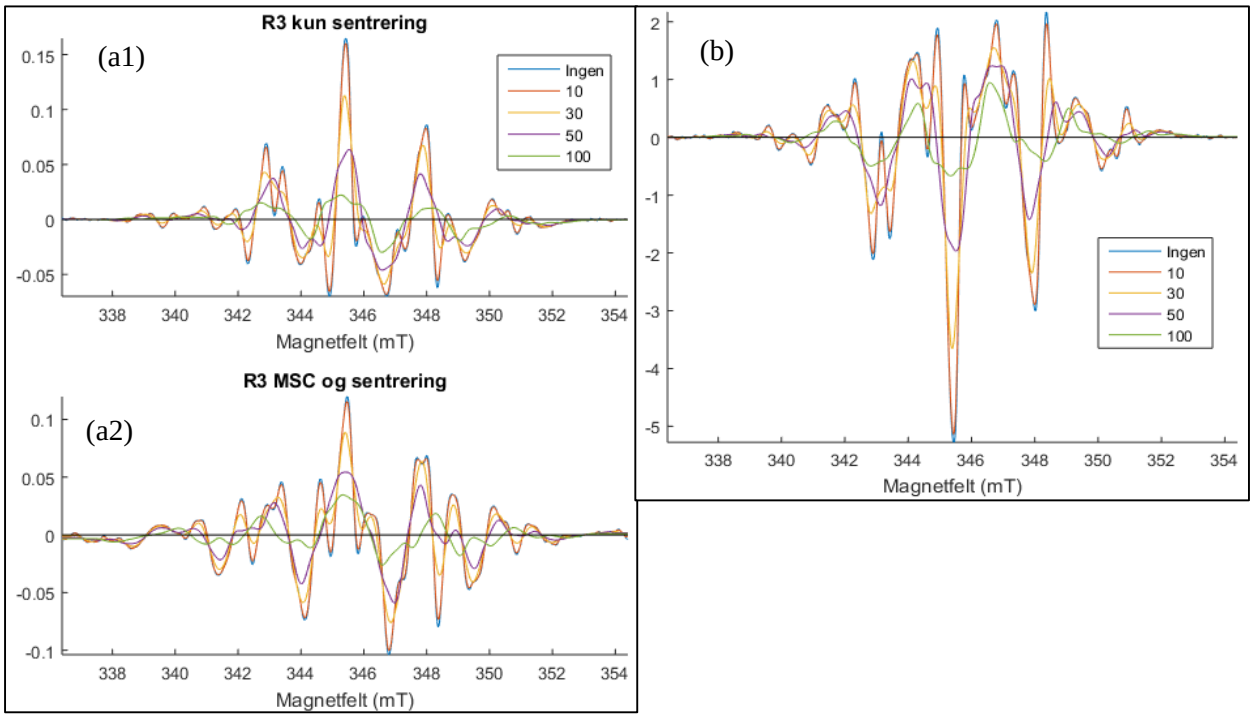
Figur 4-32a viser estimatene av R3*-spekteret fra residualanalysen i kapittel 4.1.1. Korrelasjonen mot det teoretiske R3-spekteret (figur 2-4c), er høyest dersom residualen bestemmes fra EPR-spektrene antatt å inneholde mest R3 (utvalgte). Korrelasjonen for R3*-spekteret estimert fra MCR (figur 4-32b) er høyest når ingen basisspektere er antatt kjent i modellen. R3*-spekteret fra MLCFA analysen (kapittel 4.1.3) vist i figur 4-32c, har høyest korrelasjon når MLCFA er gjort med langtids oppvarmede prøver. Korrelasjonen mellom R3 og R3* estimert fra SMA (Figur 4-32d, kapittel 4.1.5) er høyest for modellen basert på normerte spekterverider. Spekteret forskjøvet til positive verdier passer ikke med det teoretiske R3-spekteret.



Figur 4-32, glattede estimater R3-spektre, glattet med: Ingen filter og 10, 30, 50 og 100 målepunkters gjennomsnittsfilter. R3* estimert fra: (a)Minste kvadraters metodes estimering basert på residualen til alle målespektrene (a1), fra kontrollspektrene (a2) og spektre varmebehandlet lenge (a3). (b) MCR analyse med MSC preprosessering med to (b1), en (b2) og ingen (b3) kjente radikalspektre. (c) MLCFA analyse med tre komponenter (c1) og redusert modell basert på langtidsvarmede spektre (c2). (d) SMA analyse med normerte spekterverider (d1) og forskjøvet spekterverider til positive verdier (d2).*

Estimater av R3 med lav korrelasjon

Metodene PCA (kapittel 4.1.2) og ICA (kapittel 4.1.6) ga ingen komponenter som lignet på det teoretiske R3-spekteret (figur 4-33). Tilhørende korrelasjon, LOF og R^2-tabell er vist i vedlegg 8.6 tabell 8-13.



Figur 4-33, glattede R3-spektre, glattet med: Ingen filter og 10, 30, 50 og 100 målepunkters gjennomsnittsfilter. R3* estimater fra: (a) PCA med preprosesserings kun sentrering (a1) og MSC og sentrering (a2). (b) ICA med tre komponenter*

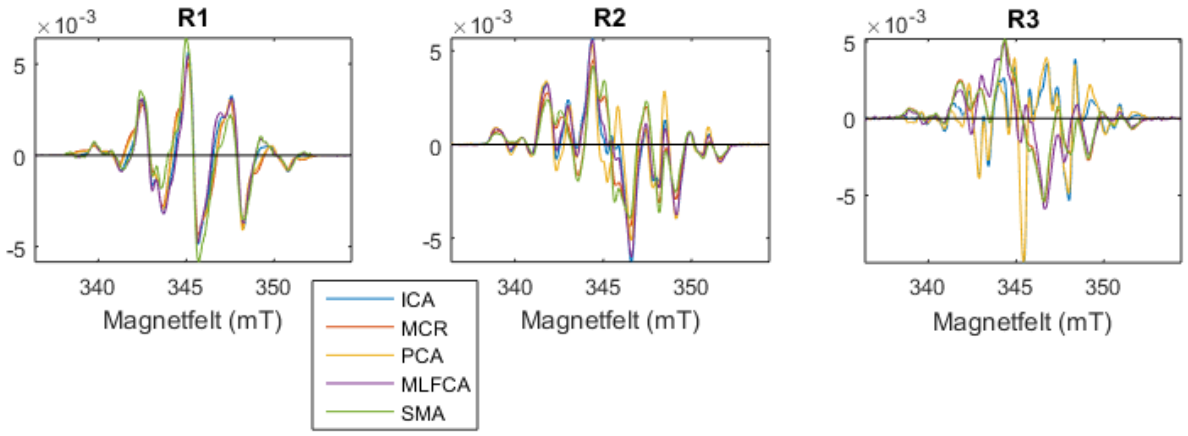
4.1.9 Korrelasjon mellom de teoretiske og estimerte spektrene

Tabell 4-13 viser korrelasjonen mellom de estimert R1*-, R2*- og R3*-spektrene mot de teoretiske R1-, R2- og R3-spektrene (figur 2-4). Korrelasjonen er høy for R1 (0,95-0,99), noe lavere for R2 (0,75-0,91) og ujevn for R3 (0,066-0,91). Korrelasjonen er høyere dersom estimerte R3* postprosesseres med et 100 målepunkters glattefilter. Grunnen til dette er at det teoretiske R3-spekteret er glatt, mens de fleste av de estimerte R3* er mer hakkede i formen.

Figur 4-34 viser hvor like de estimert R1*-, R2*- og R3*- spektrene er, uavhengig av estimerings metode. Alle de estimerte R1* og R2* ligner på hverandre. R3* estimatene kan deles inn i to grupper med lignende form. Den ene gruppen består av estimatene fra PCA og ICA, mens den andre består av estimatene fra MLCFA, MCR og SMA.

Tabell 4-13, korrelasjonene mellom de beste estimatene av R1-, R2*- og R3*- spektrene funnet fra minste kvadrater metode, PCA, MLCFA, MCR, SMA og ICA, og de teoretiske R1-, R2- og R3-spektrene (figur 2-4).*

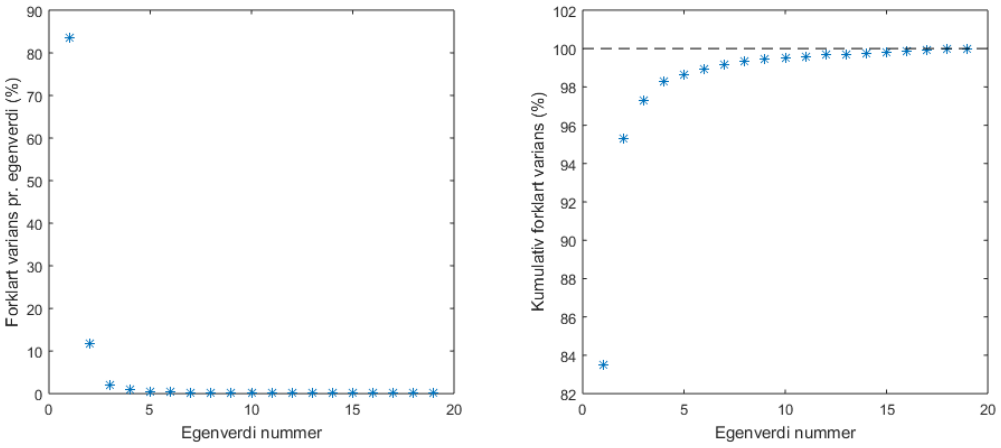
Metode	Preprosessering/ reduksjoner	R1	R2	R3 ingen post-prosessering	R3 post-prosessering med 100 målepunkters glattefilter
Minste kvadraters metode	Utvalgte	-	-	0,91	0,97
PCA	Gjennomsnitt-sentrering	0,99	0,75	0,096	0,27
MLCFA		0,97	0,80	0,86	0,96
MCR	Ingen kjente, MSC	0,98	0,91	0,79	0,90
SMA	Negative og positive verdier	0,95	0,86	0,78	0,35
ICA		0,98	0,81	0,066	0,20



Figur 4-34, sammenligning mellom de ulike estimatene av R1, R2* og R3*, for metodene ICA, MCR, PCA, MLCFA og SMA.*

4.1.10 Egenverdier

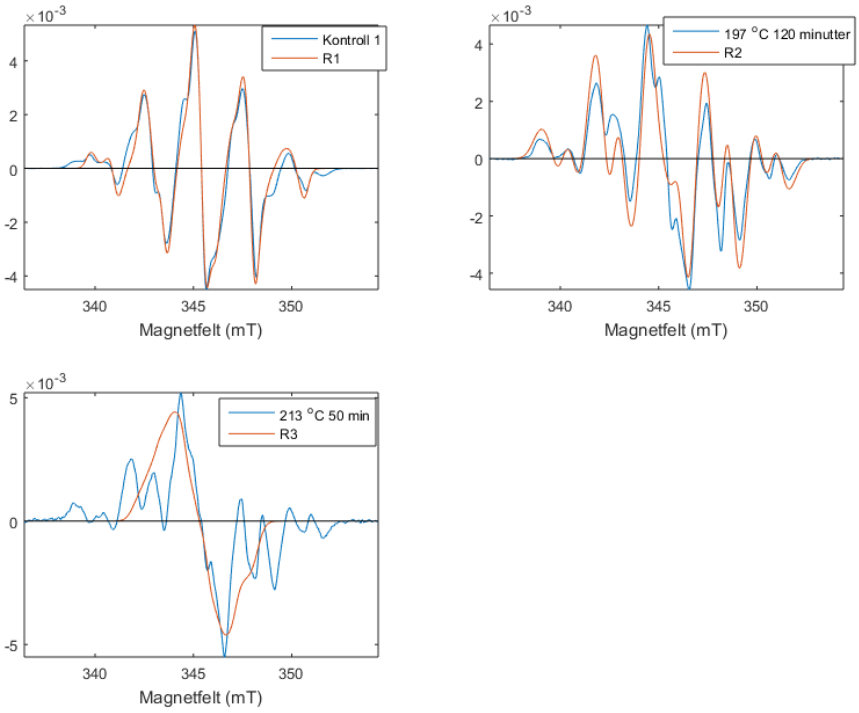
Figur 4-35 viser andelen av forklart varians som hver enkelt egenverdi står for. Det kommer fram at den første egenverdien står for ca. 83 % av den totale forklarte variansen og at ved bruk av fire egenverdier er totalt forklart varians over 98 %.



Figur 4-35, hvor mye av forklart varians hver enkelt egenverdi står for (venstre) og kumulativ forklart varians (høyre). Det kommer fram at den første egenverdien står for mesteparten av den forklarte variansen.

4.1.11 Sammenligning mellom målte og teoretiske EPR-spektre

Figur 4-36 viser en sammenligning mellom en kontroll og det teoretiske basisspekteret til R1 (figur 2-4a), prøve 197 ºC 120 minutter og det teoretiske basisspekteret til R2 (figur 2-4b) og prøve 213 ºC 50 minutter og det teoretiske basisspekteret til R3 (figur 2-4c). Basisspektrene til R1 og R2 ligner på prøve 197 ºC 120 minutter og prøve 213 ºC 50 minutter (korrelasjon 0,987 og 0,887), mens det er ingen av prøvene i datasettet som ligner på R3 (høyest korrelasjon 0,763).

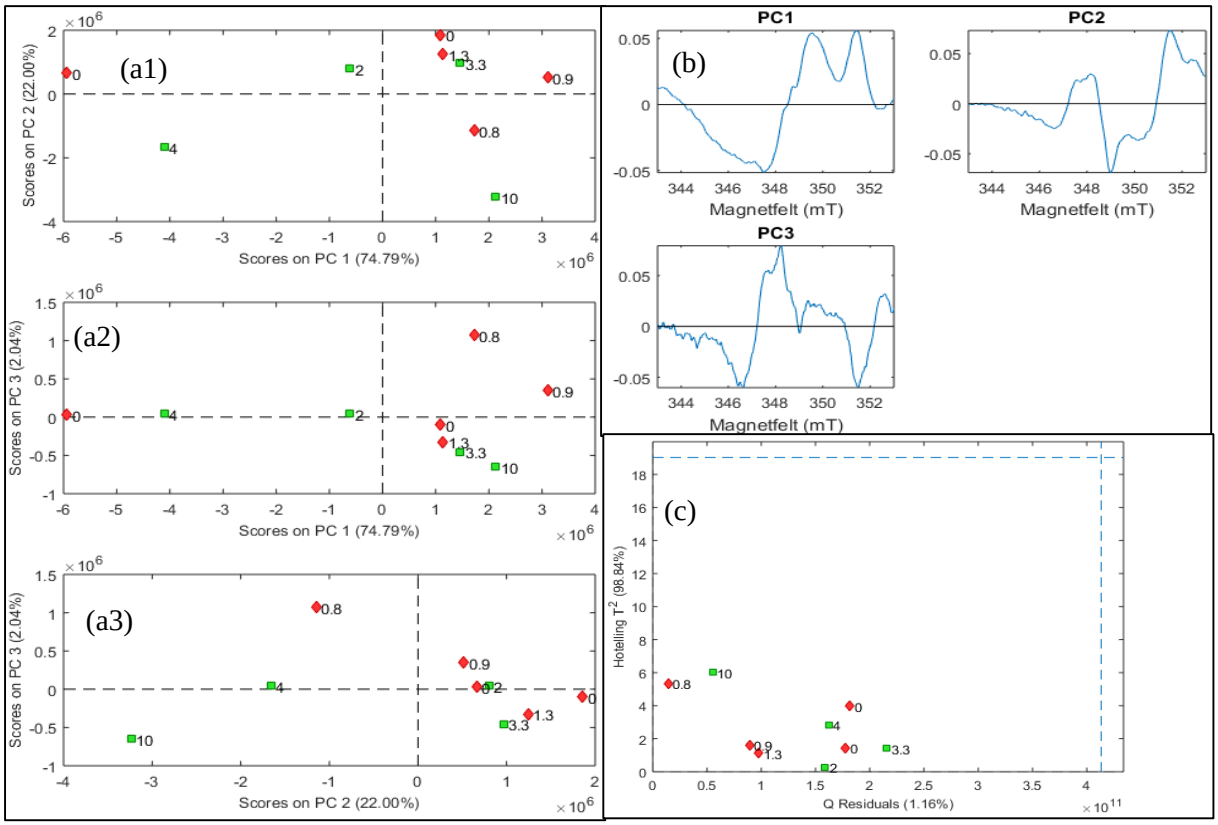


Figur 4-36, sammenligning mellom kontroll 1 og R1 (korrelasjon 0,978), prøve 197 ºC 120 minutter og R2 (korrelasjon 0,887) og prøve 213 ºC 50 minutter og R3 (korrelasjon 0,763).

4.2 Analyser av Gorilla® Glass datasettene

4.2.1 Prinsipalkomponent analyse (PCA)

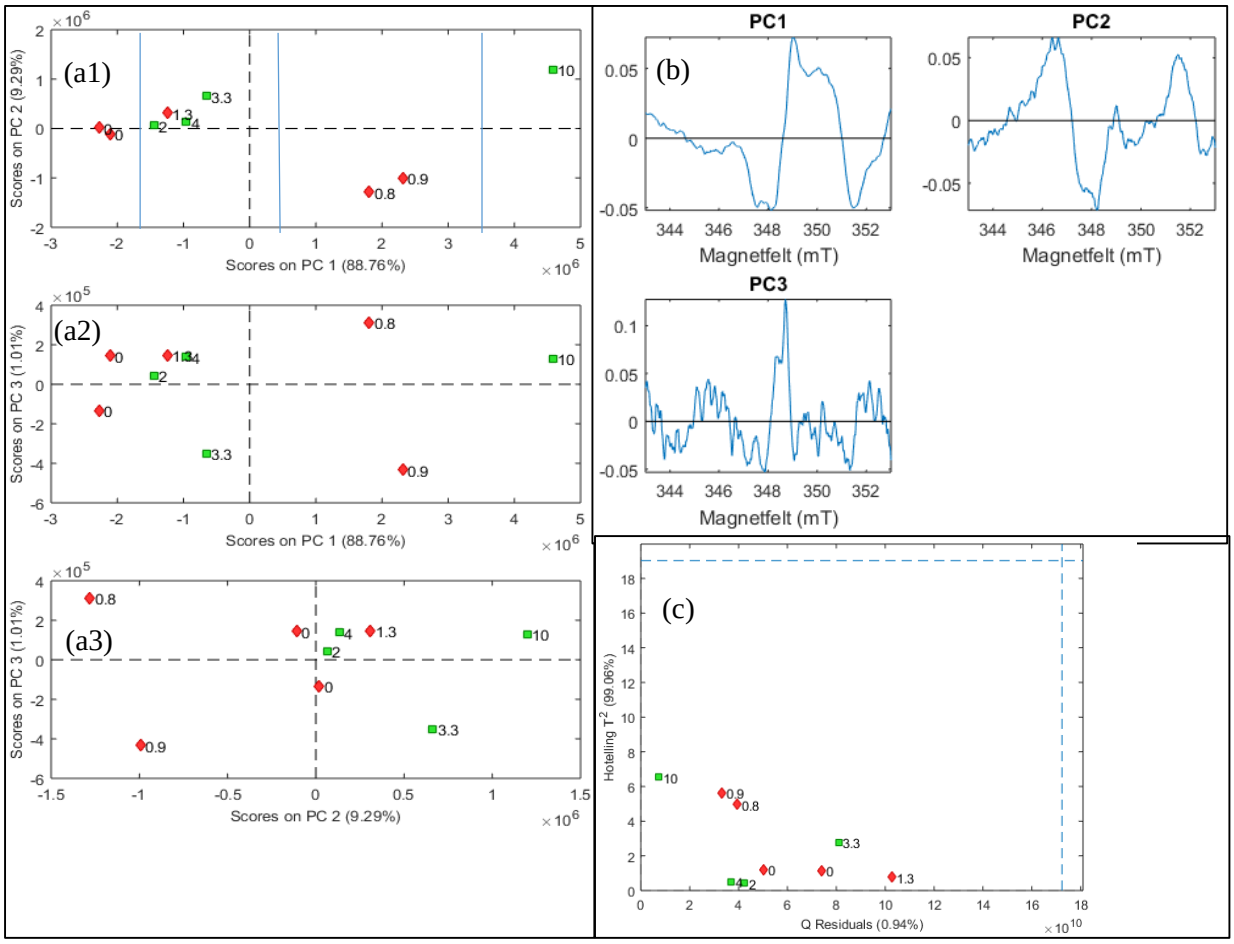
PCA av EPR-spektre av bestrålt Gorilla® Glass tatt opp i enten en SuperX eller en rektangulær kavitet gav tilnærmet likt resultat. For spektre preprosessert med glatting og sentrering, samlet ikke prøvene seg i grupper med lignende doser (se figur 4-37a for SuperX kavitet og vedlegg 8.7, figur 8-9a for rektangulær kavitet). Ladningene er vist i figur 4-37b og vedlegg 8.7, figur 8-9b. PC1 står for omtrent 70 %, PC2 for omtrent 20 % og PC3 for omtrent 2-5 % av den totale variansen. Figur 4-37c og vedlegg 8.7, figur 8-9c viser Q-residualer mot Hotelling T^2. Ingen av prøver ligger utenfor 95 % konfidensgrensen, noe som indikerer at det ikke er uteliggere i datasettet.



Figur 4-37, resultater fra PCA analysen med glattede og sentrerte EPR-spektre for SuperX kavitet datasettet, (a) PC1 mot PC2 (a1), PC1 mot PC3 (a2) og PC2 mot PC3 (a3), (b) ladningene og (c) Q-residualer mot T^2 med 95% konfidensintervall grenser (stiplede linjer). Fargene angir dosen prøvene har mottatt: rød betyr lavdose (<2 Gy), grønn betyr høydose (≥2 Gy), tallene er eksakt dose prøvene har absorbert. PC1 står for 74,79 %, PC2 for 22,00 % og PC3 for 2,04 % av total varians.

Preprosessering av EPR-spektrene med EMSC, glatting og sentrering gav lignende resultater (se figur 4-38 og vedlegg 8.7, figur 8-10). Med EMSC som preprosessering forklarer PC1 rundt 90 %, PC2 rundt 7-9 % og PC3 rundt 1 % av total varians, noe som betyr at PC1 står for mer av den totale variansen ved EMSC preprosesseringen. I skårplottet PC1 mot PC2 (figur 4-38a1)

for SuperX kavitet, er det en svak klyngedannelse langs PC1, hvor henholdsvis kontrollprøvene, prøvene 1,3-2 Gy og 3,3-4 Gy, prøvene 0,8-0,9 Gy, og 10 Gy prøvene danner klynger.



Figur 4-38, resultater fra PCA analysen med EMSC, glatting og sentrering som preprosessering på SuperX kavitet datasettet, (a) PC1 mot PC2, med forslag til klyngedannelse (vertikale blå streker) (a1), PC1 mot PC3 (a2) og PC2 mot PC3 (a3), (b) ladningene og (c) Q-residualer mot T^2 med 95% konfidensintervall grenser (stiplede linjer). Fargene angir dose prøvene har mottatt: rød betyr lavdose (<2 Gy), grønn betyr høydose (≥2 Gy), tallene er eksakt dose prøvene har absorbert. PC1 står for 88,76 %, PC2 9,29 % PC3 1,01 % av total forklart varians.

4.2.2 Regresjonsmodeller for dosebestemmelse

Minste kvadraters metode

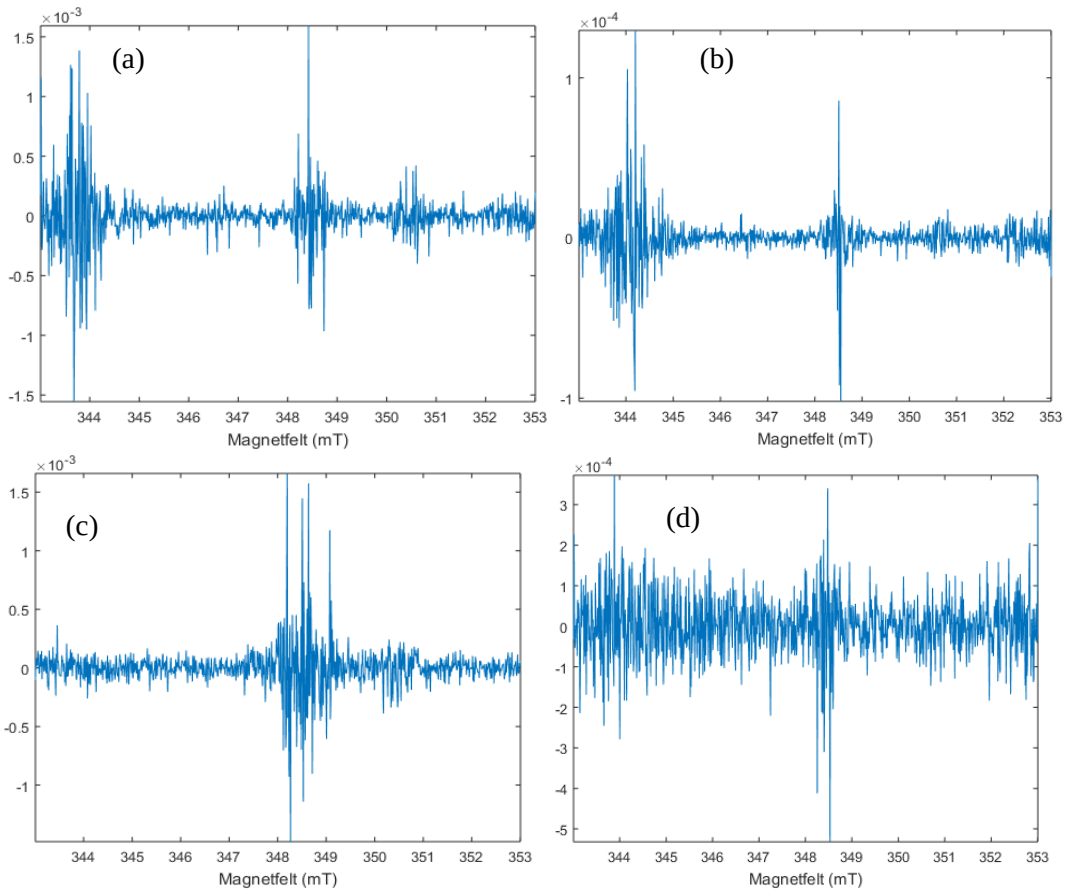
Minste kvadraters metode ble utført på begge datasettene (SuperX og rektangulær kavitet) to ganger. Første gang er det antatt at det finnes et konstantledd og andre gang med et sentrert datasett. Regresjonskoeffisientene er vist i figur 4-39.

Regresjonsmodellene ble brukt til å estimere dosene med og uten LOOCV. Sammenhengen mellom estimert dose og sann dose ble dårligere med kryssvalidering. Dette gjenspeiles i RMSEC og RMSECV vist i tabell 4-14. RMSEC er null for alle modellene uten kryssvalidering

og RMSECV er høy (7,4-39) for alle modellene med kryssvalidering. Dette viser også i regresjonsresultatene for doseestimering i figur 4-40. Dette betyr at modellene overtilpasser prøvene, som betyr at regresjonsmodellen ikke er egnet til å estimer doser til nye prøver. RMSECV er noe lavere for rektangulær kavitet enn for SuperX, og er lavest for modellene med konstantledd.

Tabell 4-14, RMSECV og RMSECV for SuperX og rektangulær kavitet med konstantledd og med sentrering.

Datasett	Metode	RMSEC	RMSECV
SuperX	Konstantledd	0	23,8
	Sentrert	0	38,8
Rektangulær	Kontantledd	0	7,37
	Sentrert	0	30,0



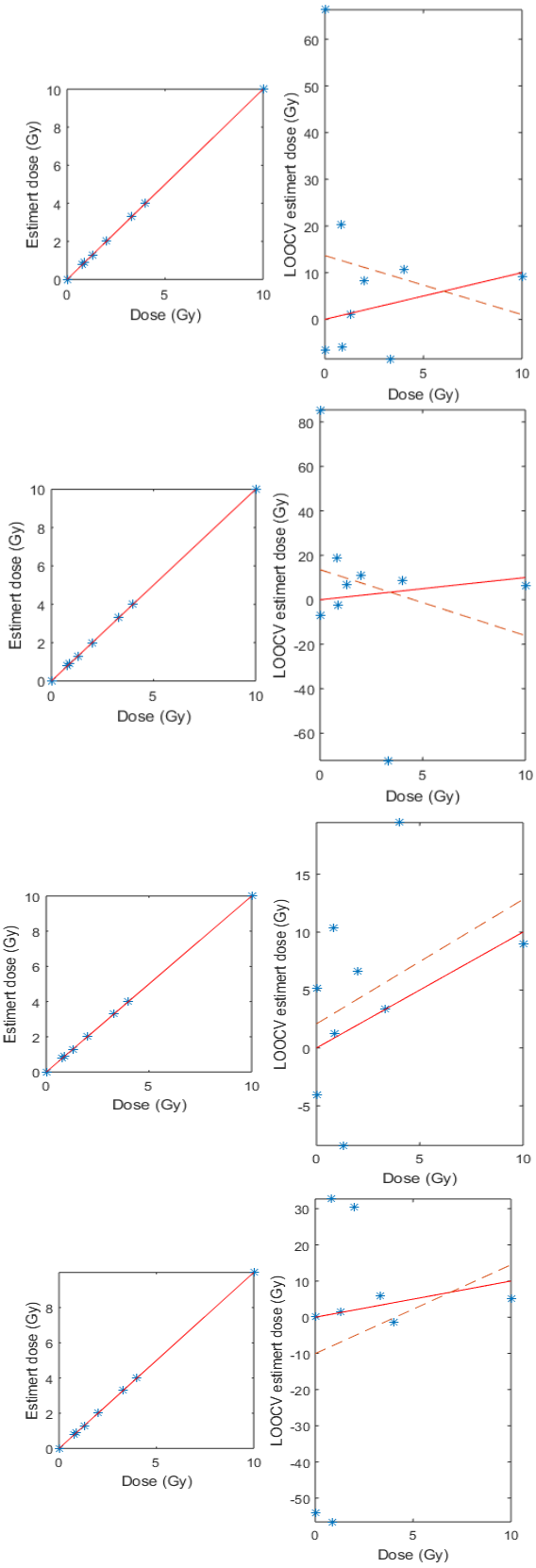
Figur 4-39, regresjonskoeffisienter fra minste kvadraters metode for: (a) SuperX kavitet med konstantledd, (b) SuperX kavitet med sentrering, (c) rektangulær kavitet med konstantledd og (d) rektangulær kavitet med sentrering.

(a)

(b)

(c)

(d)



Figur 4-40, estimering av dose ved regresjon uten (kolonne 1) og med LOOCV (kolonne 2): (a) SuperX kavitet datasettet med konstantledd, (b) sentrert SuperX kavitet datasettet, (c) rektangulær kavitet datasettet med konstantledd, (d) sentrert rektangulær kavitet datasettet. Heltrukken linje er prøver estimert riktig og stiplet linje er beste tilpasning til estimeringen. RMSECV for denne linjen er: (a) 21,35, (b) 22,35, (c) 7,01 og (d) 11,92.

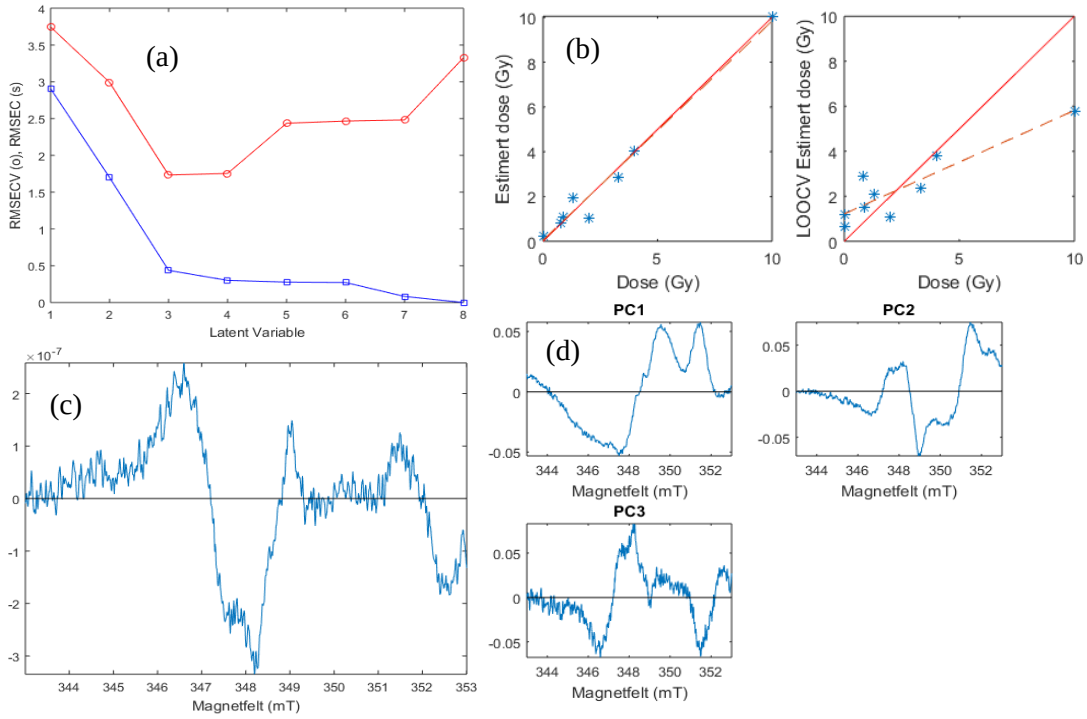
PCR

PCR analysene ble utført på begge kavitet datasettene, med enten sentrering eller MSC og sentrering som preprosessering. Resultatene vises i figur 4-41 og figur 4-42 og vedlegg 8.8, figur 8-11 og figur 8-12. To til fire prinsipalkomponenter ble valgt på bakgrunn av lavest RMSECV verdi (tabell 4-15 og figur 4-41a og figur 4-42a og vedlegg 8.8, figur 8-11a og figur 8-12a). De valgte prinsipalkomponentene forklarer i gjennomsnitt ca. 95 % av variansen og responsene. De estimerte dosene uten og med kryssvalidering er vist i figur 4-41b og figur 4-42b og vedlegg 8.8, figur 8-11b og figur 8-12b. RMSECV er høyere enn RMSEC (tabell 4-15) i alle tilfellene og LOOCV tilpasningene er også dårligere. Den beste tilpasningen er for SuperX kavitet datasettet med sentrering (lavest RMSECV).

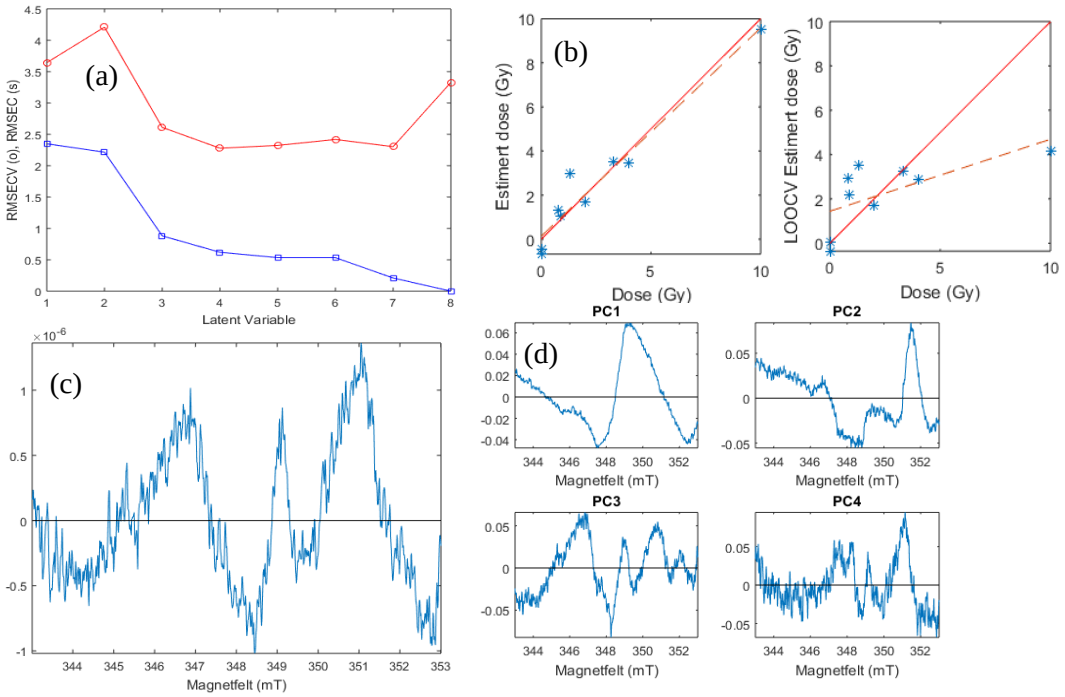
Figur 4-41c, figur 4-42c og vedlegg 8.8 viser at regresjonskoeffisientene er relativt like for de fire PCR modellene. Regresjonskoeffisientene er en lineærkombinasjon mellom prinsipalkomponentene som er med i analysen og for disse modellene er prinsipalkomponentene relativt like uavhengig av preprosessering. Ingen av LOOCV estimeringene gjort er nærme de eksakte verdiene, derfor er det lite sannsynlig at PCR er en god analytisk metode for å estimere doser for disse to datasettene. Spesielt 10 Gy prøvene underestimeres.

Tabell 4-15, RMSEC og RMSECV for PCR på SuperX og rektangulær kavitet datasettene med sentrering og MSC og sentrering som preprosessering.

	Sentrering			MSC og sentrering		
	# PC	RMSEC	RMSECV	# PC	RMSEC	RMSECV
SuperX	3	0,44	1,73	2	1,08	2,38
Rektangulær	4	0,74	2,68	4	0,62	2,28



Figur 4-41, PCR analyse av sentrert SuperX kavitet datasettet, (a) RMSEC (blå) og RMSECV (rød), (b) estimerte doser uten kryssvalidering (venstre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose. Stiplet linje er beste tilpasning til estimeringen RMSEC er 0,43 og RMSECV 0,66, (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for 74,5 % (PC1), 22,0 % (PC2) og 2,1 % (PC3) av den totale forklarte variansen for variablene, mens for responsene står PC1 for 3,4 %, PC2 63,7 % og 30,8 % av den forklart varians.



Figur 4-42, PCR analyse av rektangulær kavitet datasettet, med preprosessering sentrering og MSC, (a) RMSEC (blå) og RMSECV (rød), (b) estimerte doser uten kryssvalidering (venstre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir RMSEC 0,61 og RMSECV 1,07, (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for 81,7 % (PC1), 10,2 % (PC2), 5,7 % (PC3) og 1,4 % (PC4) av forklart varians for variablene og står for 36,9 % (PC1), 7,0 % (PC2), 47,2 % (PC3) og 4,5 % (PC4) av forklart varians hos responsen.

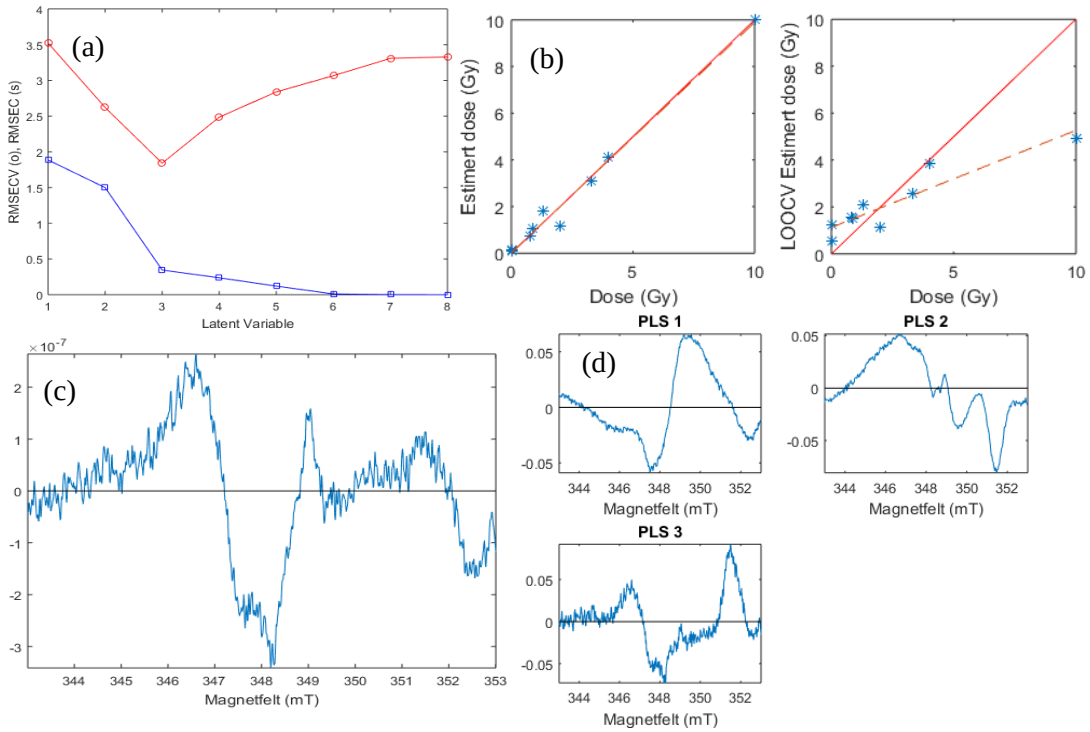
4.2.3 Delvis minste kvadraters metode (PLS)

PLS analysene ble utført på begge kavitet datasettene, med enten sentrering eller MSC og sentrering som preprosessering. Resultatene vises i figur 4-43, figur 4-44 og vedlegg 8.8, figur 8-13 og figur 8-14. To til fire prinsipalkomponenter ble valgt (tabell 4-16, figur 4-43a, figur 4-44a og vedlegg 8.8, figur 8-13a og figur 8-14a). De valgte PLS komponentene forklarer variansen i varierende mengde. PLS komponentene forklarer 97-99 % av den totale variansen hos variablene og responsene for SuperX kavitet, mens PLS komponentene forklarer 91-95 % av den totale variansen for variablene og 69-89 % for responsen til rektangulær kavitet. De estimerte dosene uten og med kryssvalidering er vist i figur 4-43b, figur 4-44b og vedlegg 8.8, figur 8-13b og figur 8-14b. RMSECV er høyere enn RMSEC (tabell 4-16) i alle tilfellene og LOOCV tilpasningene er dårligere. Den beste tilpasningen er for SuperX kavitet datasettet med sentrering (lavest RMSECV).

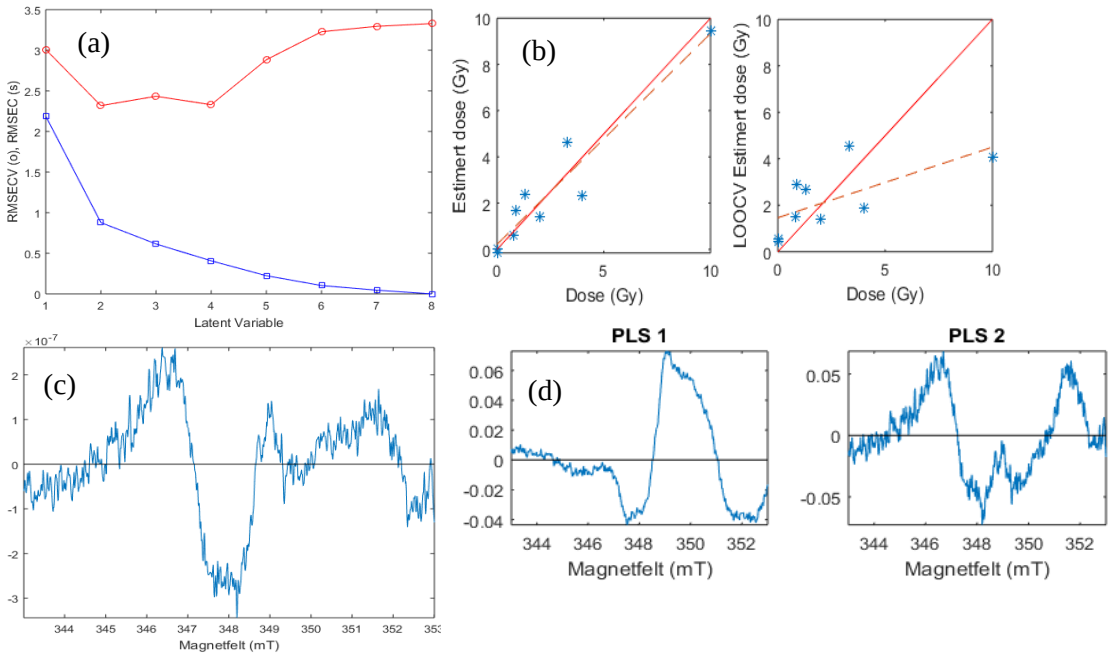
Figur 4-43, figur 4-44 og vedlegg 8.8 viser at også for PLS er regresjonskoeffisientene ganske like for alle fire modellene. Ingen av LOOCV estimeringene gjort er nærme de eksakte verdiene, spesielt 10 Gy prøven underestimeres.

Tabell 4-16, RMSEC og RMSECV for PLS på SuperX og rektangulær kavitet med sentrering og MSC og sentrering som preprosessering.

	Sentrering			MSC og sentrering		
	# PLS	RMSEC	RMSECV	# PLS	RMSEC	RMSECV
SuperX	3	0,34	1,84	2	0,88	2,32
Rektangulær	4	0,66	2,45	3	0,70	2,29



Figur 4-43, PLS regresjon av sentrert SuperX kavitet datasett, (a) RMSEC (blå) og RMSECV (rød) (b) estimerte doser uten kryssvalidering (vestre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir RMSEC 0,34 og RMSECV 0,52, (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for 41,7 % (PLS 1), 54,5 % (PLS 3) og 2,3 % (PLS 3) av den forklarte variansen til variablene og av den forklarte variansen hos responsene 59,4 % (PLS 1), 15,0 % (PLS 3) og 24,3 % (PLS 3).



Figur 4-44, PLS regresjon av SuperX kavitet datasett med preprosessering MSC og sentrering, (a) RMSEC (blå) og RMSECV (rød), (b) estimerte doser uten kryssvalidering (vestre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir RMSEC 0,83 og RMSECV 1,02, (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for 79,8 % (PLS 1) og 11,0 % (PLS 2) av forklart varians for variablene og 40,0 % (PLS 1) og 48,7 % (PLS 2) for responsene.

4.2.4 Regresjon med IPLS-variabler

IPLS på SuperX kavitet datasettet

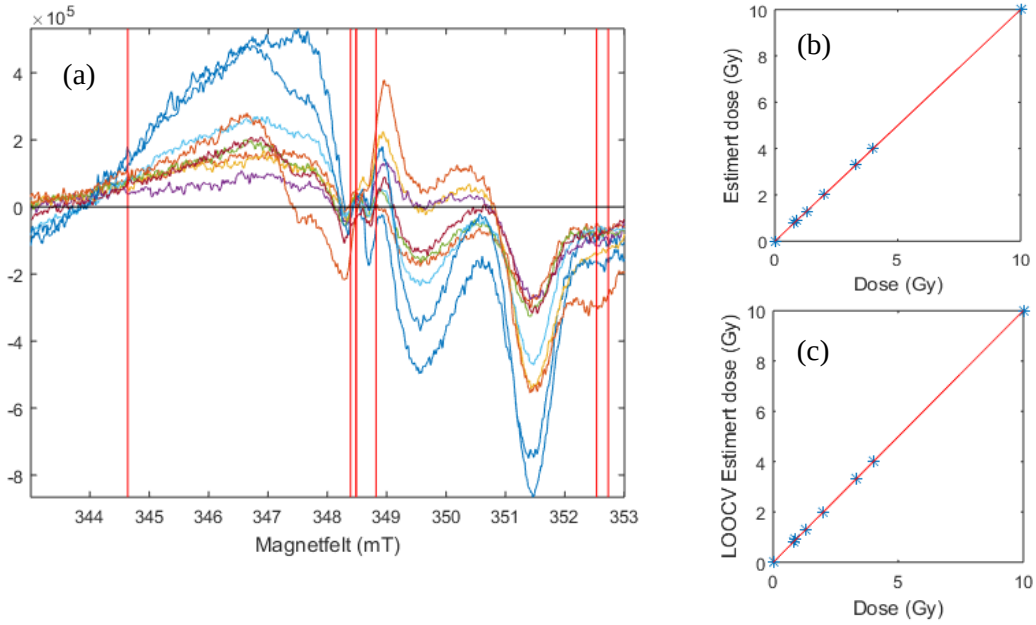
Syv variabler er optimalt for intervallbredde 1 (figur 4-45a), to variabelgrupper for intervallbredde 10 (figur 4-46a) og 100 (figur 4-47a). Variablene funnet som viktige er vist i tabell 4-18. Disse variablene ble brukt til å finne minste kvadraters løsning med konstantledd. For alle intervallbreddene var RMSEC null (tabell 4-17) og for RMSECV med intervallbredde 1 (figur 4-45c). For intervallbredde 10 og 100 ble RMSECV høy (tabell 4-17). Dette betyr at det er perfekt estimering for intervallbredde 1 og ganske dårlig tilpasning for intervallbreddene 10 og 100.

Tabell 4-17, RMSEC og RMSECV funnet ved IPLS på SuperX kavitet datasettet i figur 4-45bc, figur 4-46bc og figur 4-47bc.

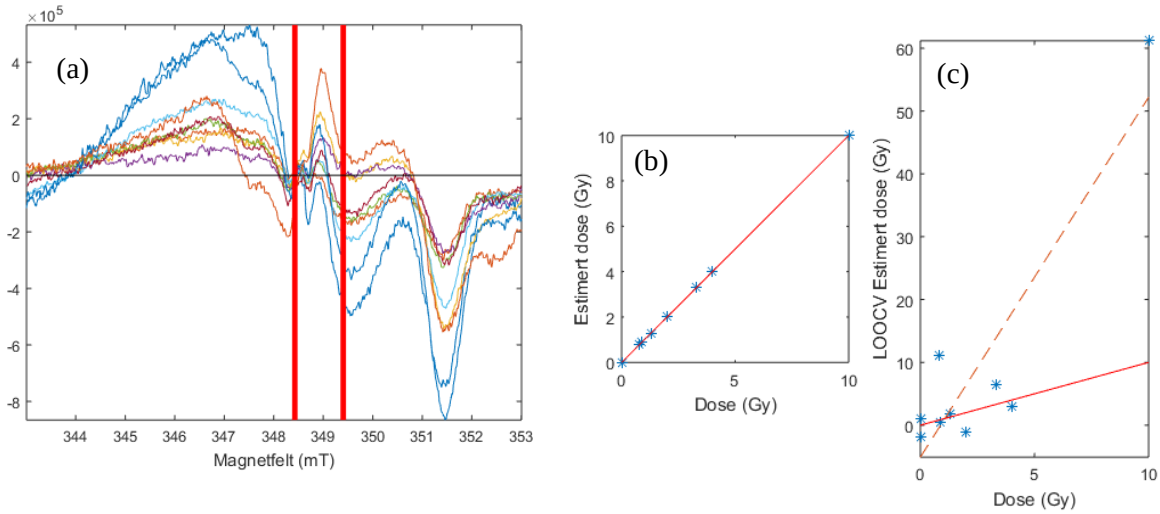
Intervallbredde	Antallet variabelgrupper	RMSEC	RMSECV
1	7	0	0
10	2	0	17,5
100	2	0	39,2

Tabell 4-18, de viktigste variablene funnet ved IPLS på SuperX kavitet datasettet i figur 4-45a, figur 4-46a og figur 4-47a.

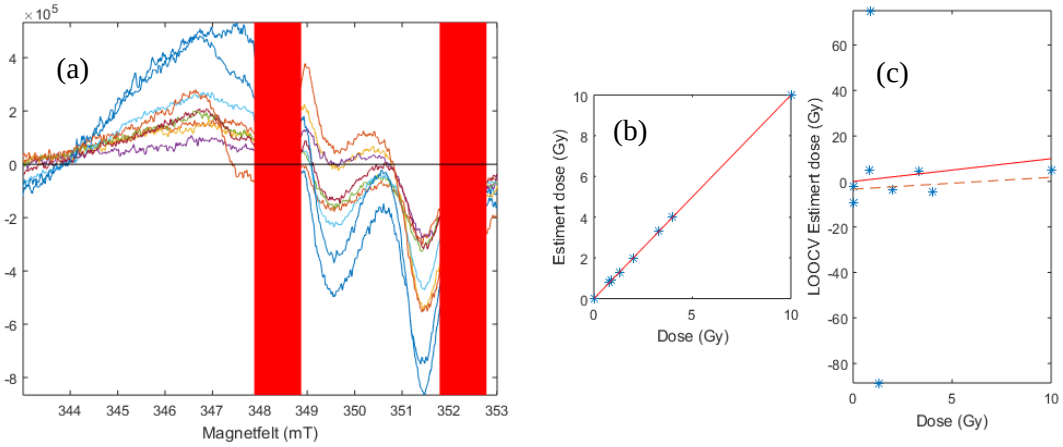
Intervallbredde	IPLS-variablene
1	344,637 mT, 348,391 mT, 348,479 mT, 348,489 mT, 348,821 mT, 352,536 mT og 352,731 mT
10	348,381-348,469 mT og 349,359-349,447 mT
100	347,893-348,860 mT



Figur 4-45, de syv beste frittstående IPLS-variablene for det SuperX kavitet datasettet, med intervallbredde på en variabel. (a) De syv viktigste frittstående variablene (rød vertikale streker). (b) Resultatene fra minste kvadraters metode, estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose. Alle punktene som ligger på den røde linjen har blitt estimer riktig.



Figur 4-46, de to viktigste frittstående IPLS gruppevariablene for det SuperX kavitet datasettet, med intervallbredde på ti variabel. (a) De to viktigste gruppene av variabler (rød vertikale streker). Resultatene fra minste kvadraters metode, (b) estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose. Beste tilpasning (stiplet linje) gir RMSECV 8,13.



Figur 4-47, de to viktigste frittstående IPLS gruppevariablene for det SuperX kavitet datasettet, med intervallbredde på 100 variabel. (a) De to viktigste gruppene av variabler (rød vertikale streker). Resultatene fra minste kvadraters metode, (b) estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose. Beste tilpasning (stiplet linje) gir RMSECV 38,89.

IPLS på rektangulær kavitet datasettet

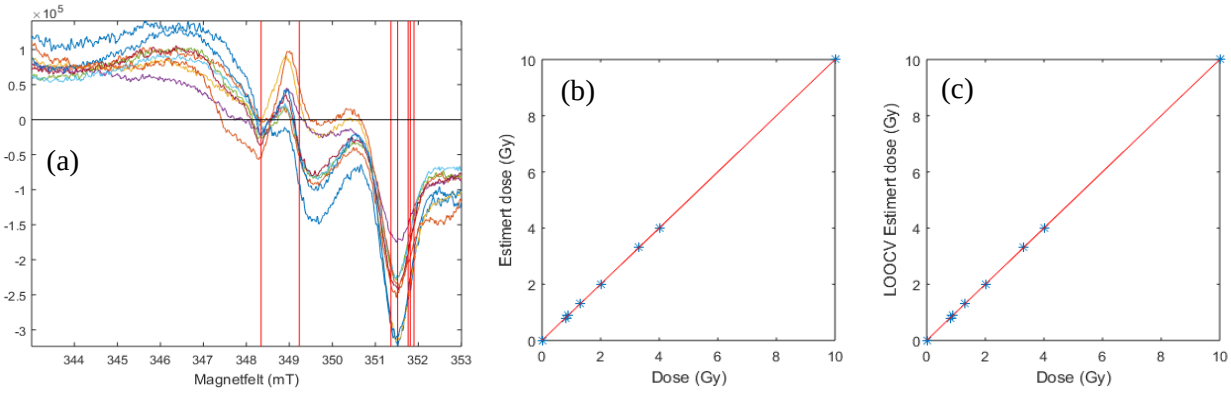
Syv variabler er optimalt for intervallbredde 1 (figur 4-48a), en variabelgruppe for intervallbredde 10 (figur 4-49a) og 100 (figur 4-50a). Disse magnetfeltene (tabell 4-20) ble brukt til å lage en regresjonsmodell ved bruk av minste kvadraters metode. Alle modellene hadde RMSEC på null (tabell 4-19) og RMSECV for intervallbredde 1 (figur 4-48c). For intervallbrede 10 og 100 ble RMSECV høy (tabell 4-19). Dette viser at det er kun med intervallbredde 1 at regresjonsmodellen klarer å estimere prøver som ikke er med i kalibreringsdatasettet. Spesielt 0 Gy og 10 Gy prøvene blir feilestimert ved LOOCV.

Tabell 4-19, RMSEC og RMSECV funnet ved IPLS på rektangulær kavitet datasettet i figur 4-48bc, figur 4-49bc og figur 4-50bc.

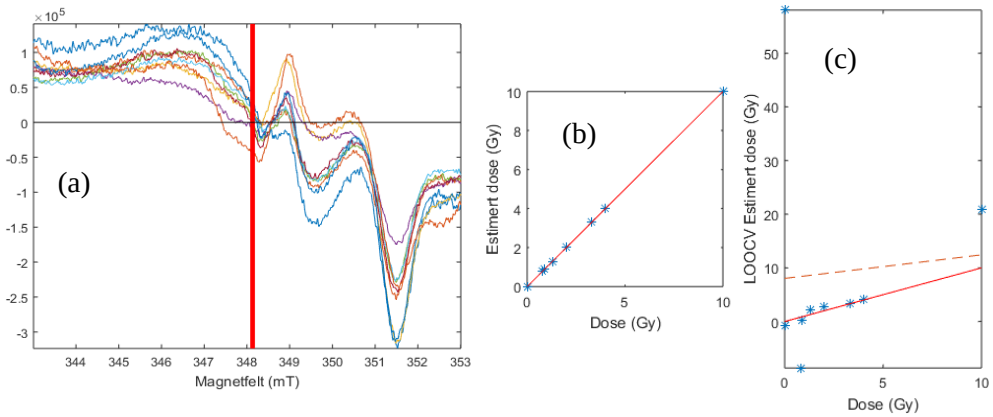
Intervallbredde	Antallet variabelgrupper	RMSEC	RMSECV
1	7	0	0
10	1	0	19,9
100	1	0	14,1

Tabell 4-20, de viktigste variablene funnet ved IPLS på rektangulær kavitet datasettet i figur 4-48a, figur 4-49a og figur 4-50a

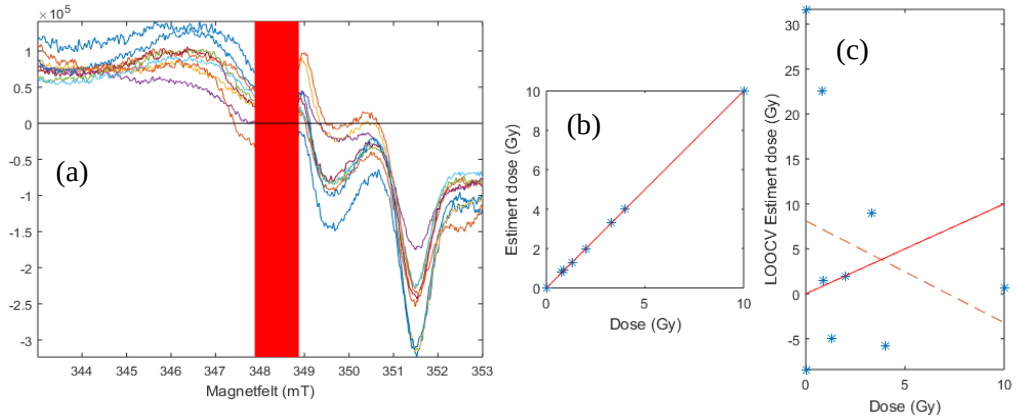
Intervallbredde	IPLS-variablene
1	348,342 mT, 349,232 mT, 351,363 mT, 351,519 mT, 351,764 mT, 351,812 mT og 351,900 mT
10	348,088-348,176 mT
100	347,893-348,860 mT.



Figur 4-48, de syv beste frittstående IPLS-variablene for det rektangulær kavitet datasettet, med intervallbredde på en variabel. (a) De syv viktigste variablene (rød vertikale streker). Resultatene fra minste kvadraters metode, (b) estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose.



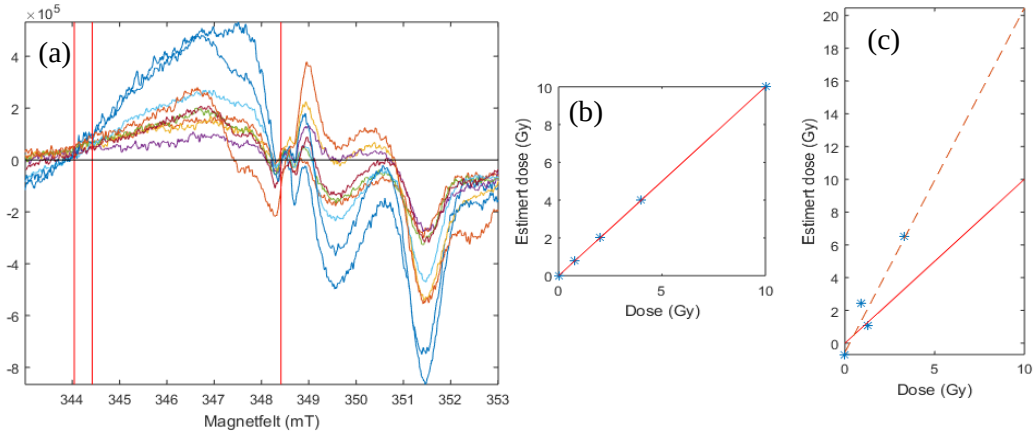
Figur 4-49, den viktigste frittstående IPLS gruppevariablene for det rektangulær kavitet datasettet, med intervallbredde på ti variabel. (a) Den viktigste gruppen av variabler (rød vertikale streker). Resultatene fra minste kvadraters metode, (b) estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose. Beste tilpasning (stiplet linje) gir RMSECV 18,72.



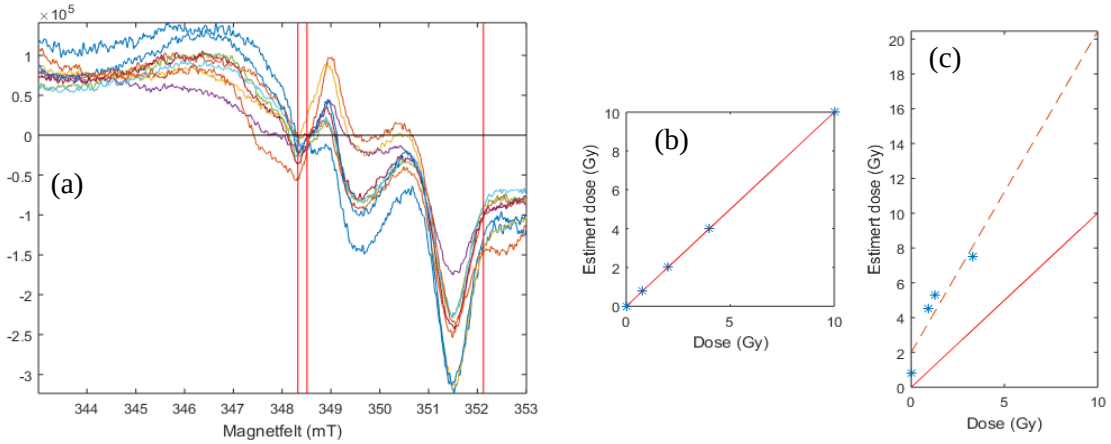
Figur 4-50, den viktigste frittstående IPLS gruppevariablene for det rektangulær kavitet datasettet, med intervallbredde på 100 variabel. (a) Den viktigste gruppen av variabler (rød vertikale streker). Resultatene fra minste kvadraters metode, (b) estimert dose mot gitt dose uten kryssvalidering og (c) LOOCV estimert dose mot gitt dose. Beste tilpasning (stiplet linje) gir RMSECV 12,33.

IPLS på kalibrerings- og valideringsdatasett

IPLS analyse ble utført på kalibreringsdatasettet av SuperX kavitet og rektangulær kavitet datasettene. Variablene 344,041 mT, 344,422 mT og 348,411 mT (figur 4-51a) ble funnet som de viktigste for SuperX kavitet og 348,323 mT, 348,508 mT og 352,125 mT (figur 4-52a) for rektangulær kavitet. Deretter ble en minste kvadraters metodes regresjonsmodell for kalibreringsdatasettet laget ut fra disse variablene for begge kavitetene. Regresjonsmodellen ble testet på valideringsdatasettet og resultatene er vist i figur 4-51c og figur 4-52c. Regresjonsmodellene estimerer riktige doser for kalibreringsdatasettet (RMSECV = 0) for begge kavitetene (figur 4-51b og figur 4-52b), mens modellen har lavere nøyaktighet på estimeringen til valideringsdatasettene, RMSEP er 1,82 for SuperX kavitet (figur 4-51c) og 3,44 for rektangulær kavitet (figur 4-52).



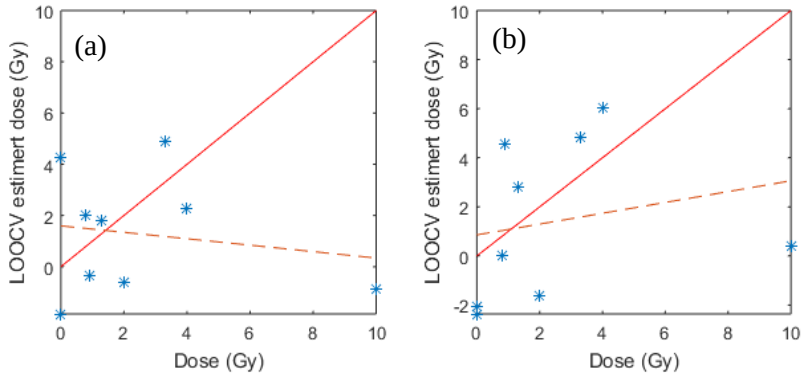
Figur 4-51, (a) de viktigste variablene markert med røde streker for SuperX kavitet datasettets kalibreringsprøver. De estimerte dosene mot de gitte dosene for kalibreringsdatasettet (b) og for valideringsdatasettet (c). Den røde linjen viser korrekt estimerte prøver. Beste tilpasning til beste linje (stiplet) gir RMSEP 0,52.



Figur 4-52, (a) de viktigste variablene markert med røde streker for rektangulær kavitet datasettets kalibreringsprøver. De estimerte dosene mot de gitte dosene for kalibreringsdatasettet (b) og valideringsdatasettet (c). Den røde linjen viser korrekt estimerte prøver. Beste tilpasning til beste linje (stiplet) gir RMSEP 0,61.

IPLS med kryssvalidering

En IPLS analyse ble gjort der en prøve ble holdt utenfor datasettet. IPLS analysen ble kjørt med intervallbredde 1. Deretter ble en minste kvadraters regresjonsmodell laget, og deretter brukt til å estimere dosen til den utelatte prøven. For hver prøve i datasettet ble en ny samling av viktige variabler funnet. Resultatene gitt i figur 4-53, viser at ingen prøver blir estimert riktig hverken for SuperX kavitet (RMSECV = 4,16) eller for det rektangulær kavitet datasettet (RMSECV = 3,92). 0 Gy prøvene ble for SuperX kavitet feilestimert til 4 Gy og -2 Gy, mens begge ble estimert til -2 Gy for rektangulær kavitet. Begge kavitetene feilestimerer 10 Gy prøvene til rundt null. Dette viser at det må utvises stor forsiktighet ved bruk av IPLS, siden variablene som blir funnet kun er viktige for kalibreringsdatasettet og ikke for nye prøver.



Figur 4-53, de LOOCV estimerte dosene mot de gitte dosene for (a) SuperX kavitet datasettet, tilpasning til beste linje (stiplet) gir RMSECV 2,19 og (b) rektangulær kavitet datasettet, tilpasning til beste linje (stiplet) gir RMSECV 2,98. Den røde linjen viser korrekt estimerte prøver.

4.2.5 Variabelreduksjon

Lasso

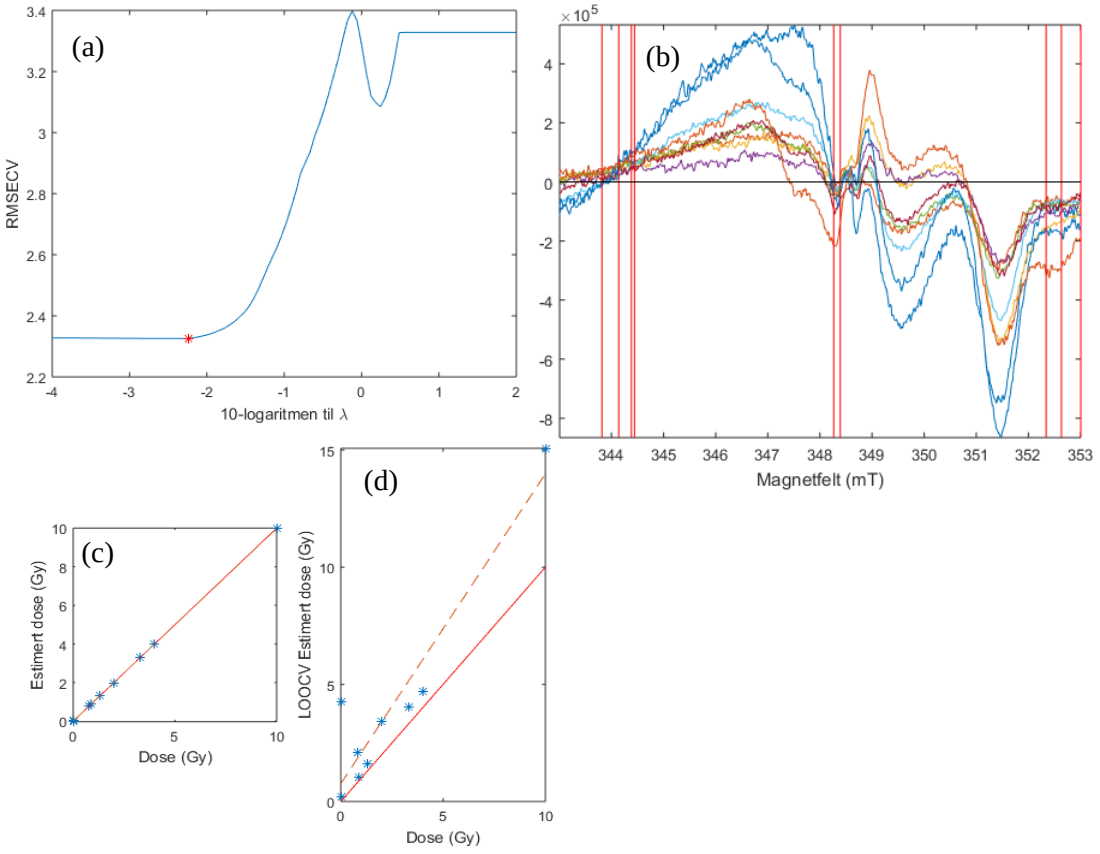
Lasso analysen ble testet med flere ulike reguleringsparametere (λ) på SuperX kavitet datasettet. Figur 4-54a viser RMSECV verdiene for de ulike λ. Den optimale λ-verdien med henblikk på RMSECV er 0,0057. De variablene som lasso algoritmen setter som forskjellig fra null er vist i figur 4-54b. Disse er 343,816 mT, 344,139 mT, 344,383 mT, 344,442 mT, 348,264 mT, 348,391 mT, 352,340 mT, 352,634 mT og 353,005 mT. Resultatet fra lasso analysen med disse ni variablene ble brukt til å estimere dosene uten (figur 4-54c) og med kryssvalidering (figur 4-54d). Spesielt den ene 0 Gy og 10 Gy prøven (figur 4-54d) blir feilestimert. RMSEC ble 0 og RMSECV ble 2,33, som antyder at modellen overtilpasser prøvene. Siden ni variabler ble plukket ut fra ni prøver, kan en regresjonsmodell lages som gir nøyaktig estimering av

prøvene, men denne modellen vil ikke nødvendigvis fange opp egenskaper til prøver som ikke inngår i kalibrering datasettet.

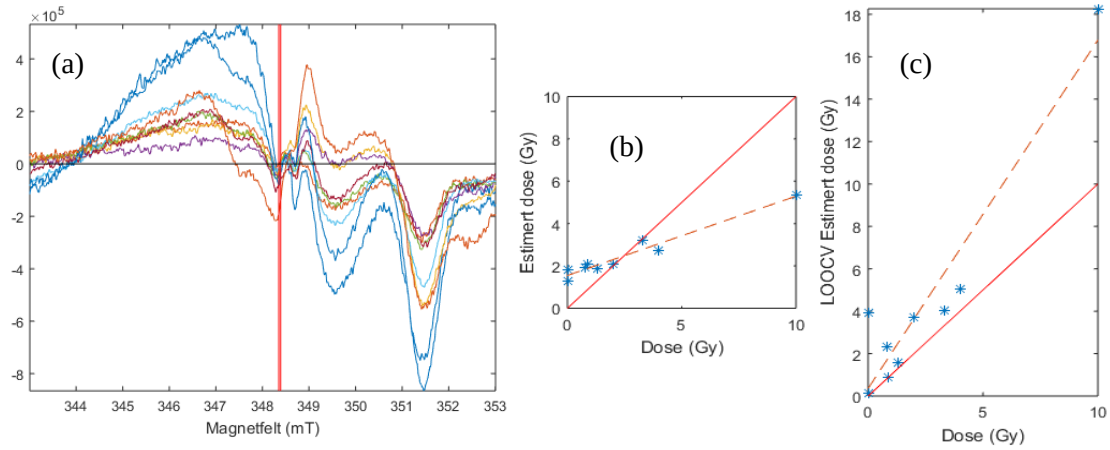
Lasso algoritmen ble også testet med sikte på å finne maksimalt fem variabler, for å se hvor få variabler det trengs før modellen blir vesentlig dårligere. Modellen blir best med kun to variabler (348,352 mT og 348,391 mT, figur 4-55a). Disse variablene ble brukt til å estimere prøvene uten (figur 4-55b) og med kryssvalidering (figur 4-55c). RMSEC ble 1,87 og RMSECV ble 3,17. LOOCV estimeringene ble dårligere enn når ni variabler ble brukt i modellen. Beste tilpasning til de estimerte dosene (stiplet linje figur 4-55b) indikerer at dosestimeringen følger en annen trend enn den forventede.

Tilsvarende lasso analyse ble utført for det rektangulær kavitet datasettet. Figur 4-56a viser at det ikke finnes en globalt minimumsverdi for RMSECV for dette datasettet. Lasso reguleringsparameteren λ ble valgt til 0,0087, dette er i knekkpunktet til RMSECV, som ble valgt for at RMSECV skulle være lav og λ så høy som mulig, siden det gjør modellen enklere. De ni selekterte variablene vist i figur 4-56b (magnetfeltene 344,804 mT, 345,312 mT, 346,495 mT, 348,430 mT, 350,679 mT, 350,845 mT, 350,854 mT, 352,702 mT og 352,761 mT). Dose estimeringen av denne lasso modellen er vist i figur 4-56cd, hvor RMSEC ble 0 og RMSECV ble 2,78.

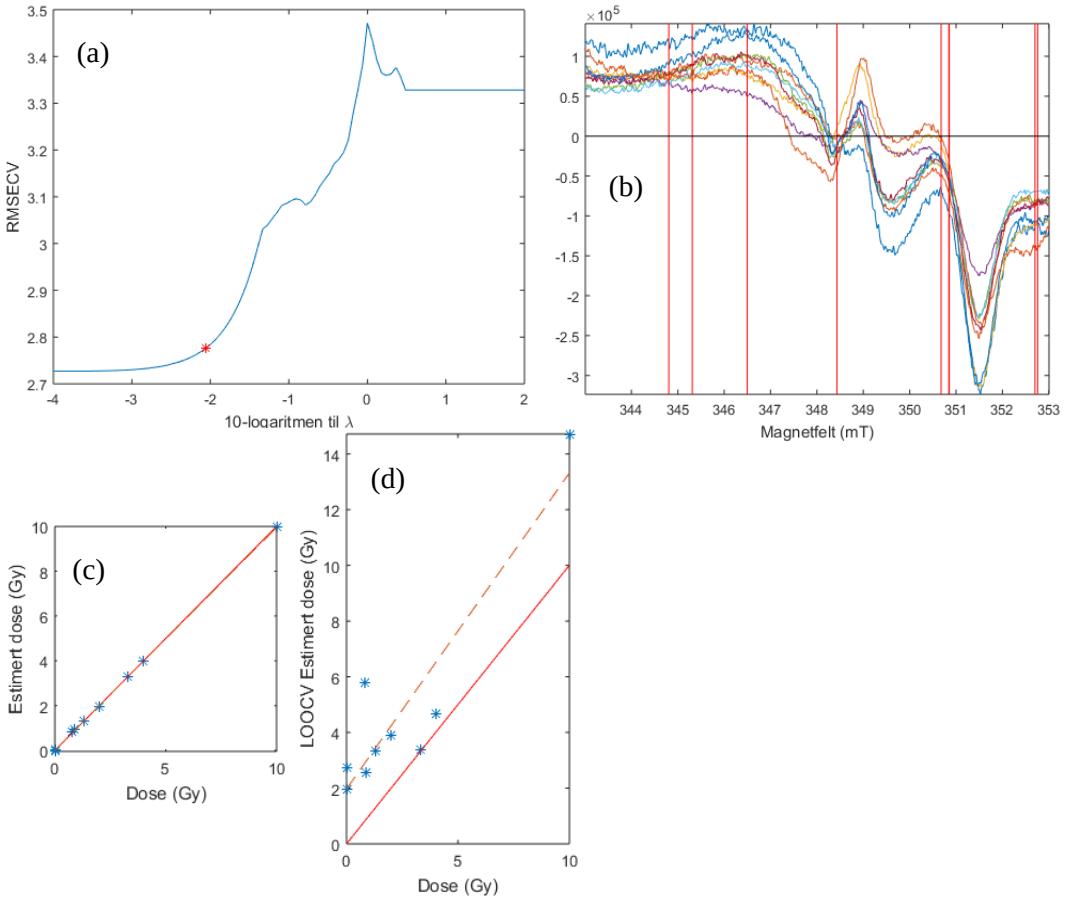
Lasso algoritmen ble også brukt for å finne den beste modellen med maksimalt tre variabler. De tre selekterte magnetfeltene er: 348,342 mT, 350,679 mT og 352,751 mT, vist i figur 4-57a. Dosene estimert med og uten kryssvalidering er vist i figur 4-57bc. LOOCV estimeringen er noe bedre enn når ni variabler var inkludert i modellen (RMSEC er 1,34 og RMSECV er 2,62).



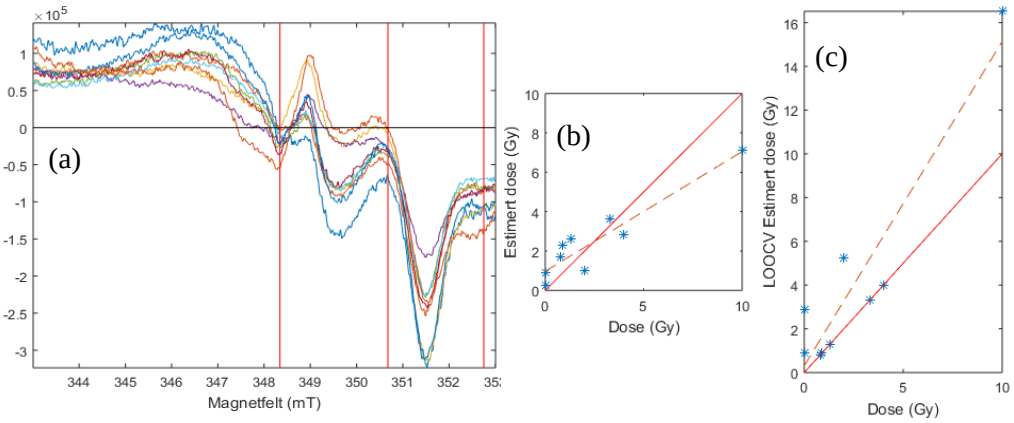
Figur 4-54, resultatene av lasso på SuperX kavitet datasettet. (a) RMSECV verdier for ulike λ-verdier. Den laveste RMSECV verdien er markert med stjerne. (b) De viktigste variablene (rød vertikale streker) funnet ved optimal λ-verdi. Estimert dose mot gitt dose uten kryssvalidering (c) og med LOOCV (d) for den optimal lasso modell. Tilpasning til beste linje (stiplet) gir RMSECV 1,43.



Figur 4-55, lasso modellen av SuperX kavitet datasettet, med maksimalt fem variabler. (a) De viktigste variablene (rød vertikale streker). Estimert dose mot gitt dose uten (b) og med kryssvalidering (c). Tilpasning til beste linje (stiplet) gir RMSEC 0,24 og RMSECV 1,62.



Figur 4-56, resultatene av lasso på rektangulær kavitet datasettet. (a) RMSECV verdier for ulike λ-verdier, av mangel på et globalt bunnpunkt for RMSECV, er λ-verdien valgt til å være 0,0087 (markert med stjerne). (b) De viktigste variablene (rød vertikale streker). Estimert dose mot gitt dose uten (c) og med LOO kryssvalidering (d). Tilpasning til beste linje (stiplet) gir RMSECV 1,49.

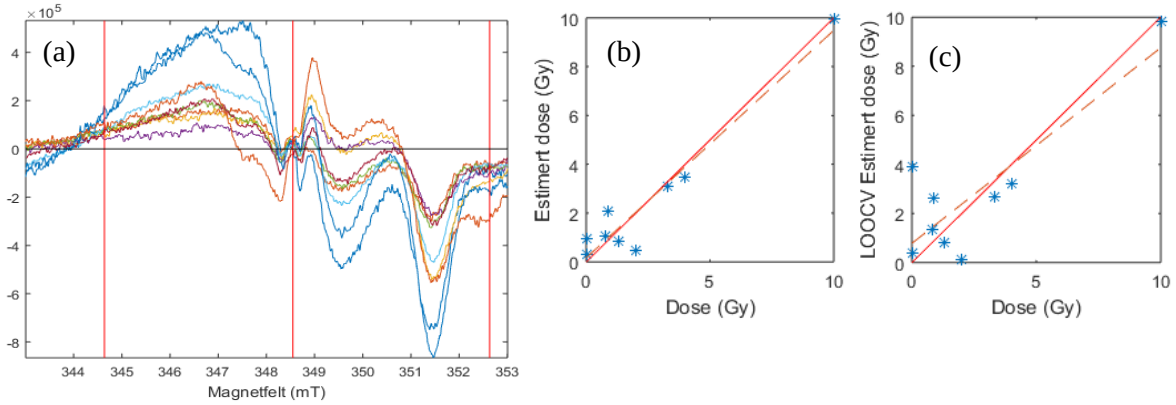


Figur 4-57, lasso modellen av rektangulær kavitet datasettet, med maksimalt tre variabler. (a) De viktigste variablene (rød vertikale streker). Estimert dose mot gitt dose uten (b) og med LOOCV (c). Tilpasning til beste linje (stiplet) gir RMSEC 0,67 RMSECV 1,61.

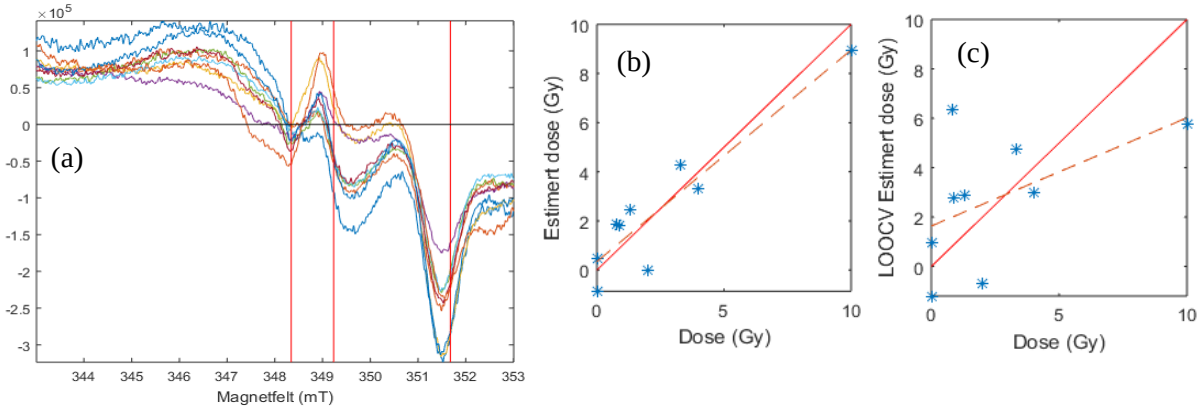
Ved reduksjon av IPLS

I analysen av IPLS-variable med intervallbredde 1, kommer det fram at det er tre grupper med viktige variabler for SuperX kavitet, vist i figur 4-45a. Istedenfor at det er syv variabler fordelt over tre grupper ble gjennomsnittsvariabelen i disse gruppene funnet og en lineær regresjonsmodell ble lagt basert på disse variablene. De variablene som ble plukket ut er 344,637 mT, 348,548 mT og 352,634 mT, vist i figur 4-58a. Denne modellen ble brukt til å estimere doser uten (figur 4-58b) og med LOOCV (figur 4-58c), og RMSEC er 0,77 og RMSECV er 1,62.

Det samme ble gjort for det rektangulær kavitet datasettet. Her er det tydelig tre grupper med variable, se figur 4-48a, og 348,342 mT, 349,232 mT og 351,626 mT (figur 4-59a) ble brukt i regresjonsmodellen, vist i figur 4-59bc. RMSEC og RMSECV ble henholdsvis 1,1 og 2,73.



Figur 4-58, (a) tre variable plukket ut til å lage en regresjonsmodell, fra IPLS analysen av SuperX datasettet (figur 4-45). Estimert dose mot gitt dose uten (b) og med LOOCV (c), Tilpasning til beste linje (stiplet) gir RMSEC 0,74 og RMSECV 1,48.

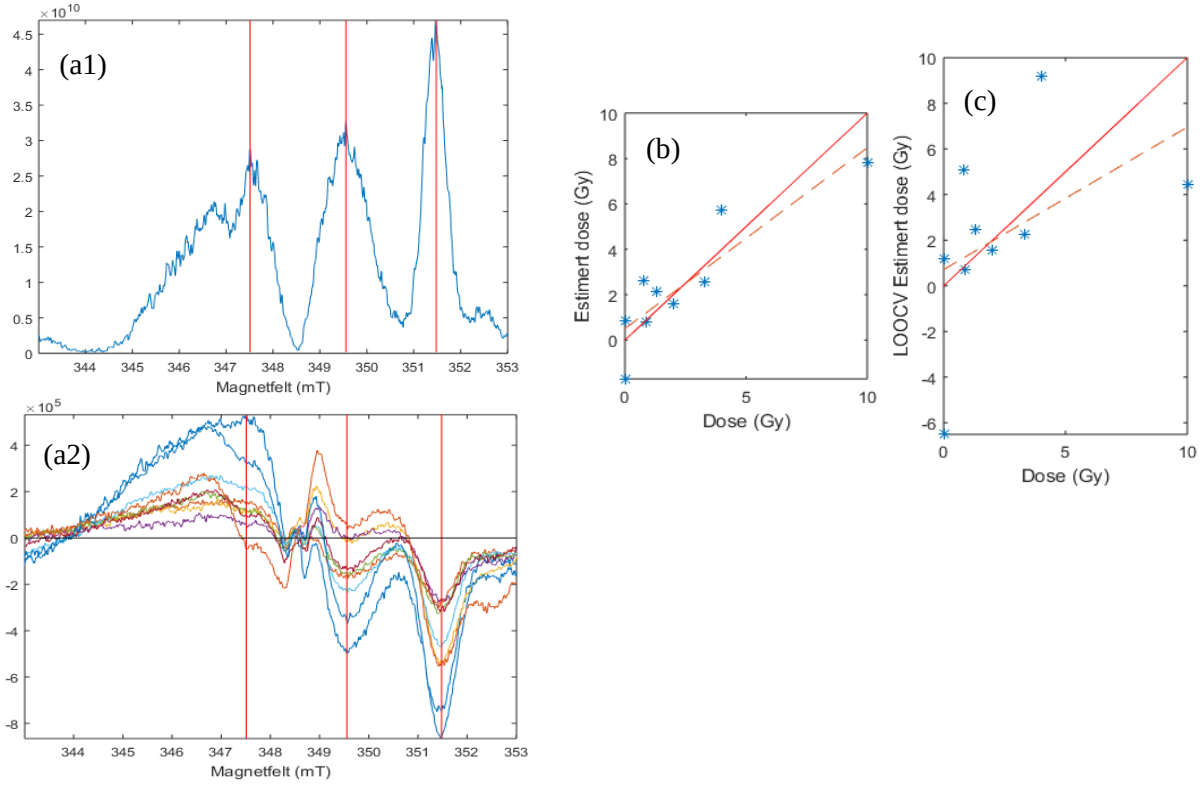


Figur 4-59, (a) tre variablene plukket ut til å lage en regresjonsmodell, fra IPLS analysen av rektangulær datasettet (figur 4-48). Estimert dose mot gitt dose uten (b) og med LOOCV (c). Tilpasning til beste linje (stiplet) gir RMSEC 1,02 og RMSECV 2,15.

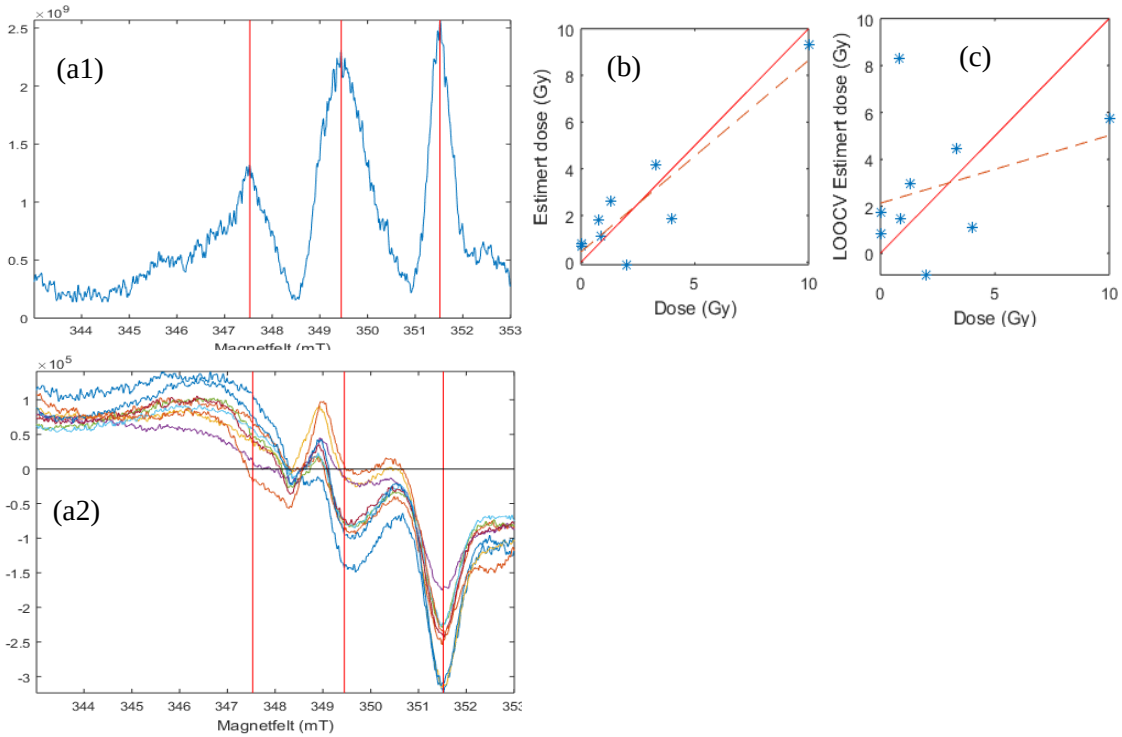
Varians analyse

En regresjonsmodell ble lagd basert på variablene med høyest varians. Figur 4-60a1 viser hvor mye varians hver enkelt variabel beskriver. Det er tre topper ved 347,511 mT, 349,554 mT og 351,480 mT med høy varians (figur 4-60a). Regresjonsmodellen (figur 4-60bc) basert på disse tre variablene, ga stort avvik mellom de estimerte dosene og de gitte dosene, spesielt for prøvene 0 Gy, 4 Gy og 10 Gy. RMSEC og RMSECV er henholdsvis 1,34 og 3,69.

Samme analyse ble også gjort på rektangulær kavitet datasettet. Magnetfeltene med høyest varians (figur 4-61a) var 347,531 mT, 349,447 mT og 351,519 mT. Regresjonsmodellen (figur 4-61bc) bygget på disse variablene og ga RMSEC på 1,25 og RMSECV på 3,33, som indikerer relativt store avvik mellom de gitt og de estimerte dosene. Dette var spesielt tilfelle for 0,8 Gy, 2 Gy, 4 Gy og 10 Gy prøvene.



Figur 4-60, variablenes varians (a1, a2) for SuperX kavitet datasettet. De røde strekene representerer de tre variablene med største varians. Estimert dose mot gitt dose uten (b) og med LOOCV (c). Tilpasning til beste linje (stiplet) gir RMSEC 1,19 og RMSECV 3,51.

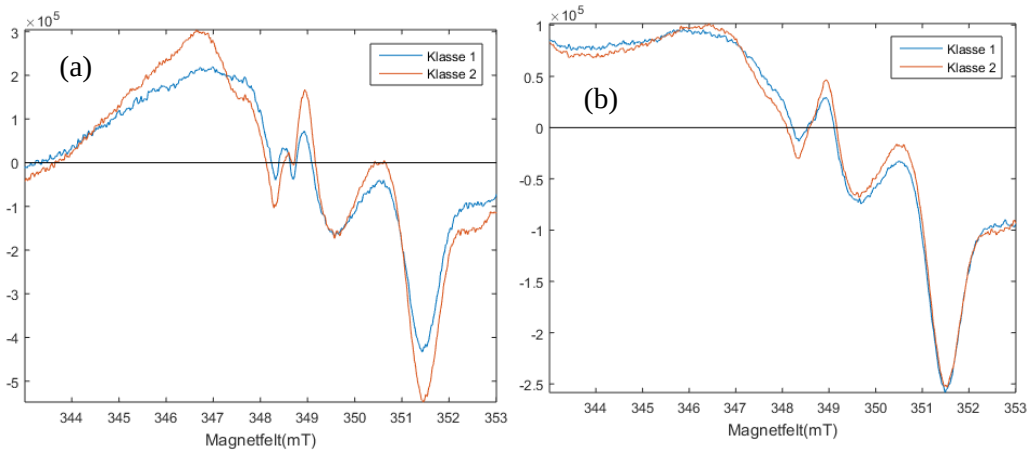


Figur 4-61, variablenes varians (a1, a2) for rektangulær kavitet datasettet. De røde strekene representerer de tre variablene med største varians. Estimert dose mot gitt dose uten (b) og med LOOCV (c). Tilpasning til beste linje gir RMSEC 1,13 og RMSECV 2,55.

4.2.6 Klassifiseringsmodeller

Lineær diskriminant analyse (LDA)

En LDA analyse ble utført med Mahalanobis avstand og med euklidsk avstand som avstandsmål på SuperX og rektangulær kavitet datasettene. Gruppesenterspektrene brukt som grunnlag i LDA analysen er vist i figur 4-62, og viser tydelig forskjeller imellom klassene for SuperX kavitet datasettet. Klassifiseringen med Mahalanobis avstand på begge datasettene gav riktig klassetildeling ved LOO kryssvalidering (tabell 4-21a og tabell 4-22a). Derimot gav klassifiseringen med euklidsk avstand flere feilklassifiserte prøver i begge datasettene (tabell 4-21b og tabell 4-22b), og tyder på at et euklidsk avstandsmåle er uegnet for klassifisering av disse datasettene.



Figur 4-62, gruppesenterspektrene som er grunnlaget for LDA analysen, av (a) SuperX kavitet datasettet og (b) rektangulær kavitet datasettet.

Tabell 4-21, LDA med (a) Mahalanobis og (b) euklidsk avstand, for SuperX datasettet, hvor klasse 1 er lavdose, klasse 2 er høydose.

(a) Mahalanobis distanse					(b) Euklidsk distanse			
Gitt dose [Gy]	Klasse	Uten CV	LOOCV		Gitt dose [Gy]	Klasse	Uten CV	LOOCV
0	1	1	1		0	1	2	2
0	1	1	1		0	1	1	1
0,8	1	1	1		0,8	1	2	2
0,9	1	1	1		0,9	1	1	1
1,3	1	1	1		1,3	1	1	1
2	2	2	2		2	2	1	1
3,3	2	2	2		3,3	2	1	1
4	2	2	2		4	2	2	2
10	2	2	2		10	2	2	2
Nøyaktighet		100 %	100 %				56 %	56 %

Tabell 4-22, LDA med (a) Mahalanobis og (b) euklidsk avstand, for rektangulær kavitet datasettet, hvor klasse 1 er lavdose, klasse 2 er høydose.

(a) Mahalanobis distanse					(b) Euklidsk distanse			
Gitt dose [Gy]	Klasse	Uten CV	LOOCV		Gitt dose [Gy]	Klasse	Uten CV	LOOCV
0	1	1	1		0	1	1	2
0	1	1	1		0	1	1	1
0,8	1	1	1		0,8	1	2	2
0,9	1	1	1		0,9	1	2	2
1,3	1	1	1		1,3	1	1	2
2	2	2	2		2	2	1	1
3,3	2	2	2		3,3	2	2	1
4	2	2	2		4	2	1	1
10	2	2	2		10	2	2	1
Nøyaktighet		100 %	100 %				56 %	33 %

PLSDA

PLS for regresjon ble brukt som utgangspunkt for klassifisering av prøvene med PLSDA. Tre PLS komponenter ble benyttet for det sentrerte SuperX kavitet datasettet. Resultatene i tabell 4-23a viser at ved LOOCV ble tre av prøvene klassifisert feil, mens alle prøvene ble klassifisert riktig uten bruk av kryssvalidering.

Tilsvarende analyse ble gjort med preprosessering MSC og sentrering, ved bruk av to PLS komponenter. Tabell 4-23b viser at henholdsvis tre og fire prøver ble klassifisert feil uten og med LOOCV. Dette viser at MSC ikke er optimal preprosesseringsteknikk for dette datasettet.

PLSDA ble også utført på det rektangulær kavitet datasettet med enten sentrering eller MSC og sentrering som preprosessering. I begge tilfellene (tabell 4-23cd) ble 1,3 Gy prøven klassifisert feil ved kalibrering. Ved LOOCV ble tre prøver feilestimert.

Tabell 4-23, de estimerte klassene med og uten LOOCV av (a,b) SuperX kavitet og (c,d) rektangulær kavitet datasettene, med preprosessering (a) sentrering med tre PLS komponenter, (b) MSC og sentrering med to PLS komponenter, (c) sentrering med fire PLS komponenter og (d) MSC og sentrering med tre PLS komponenter, der klasse 1 er lavdose og klasse 2 er høydose (tabell 3-2).

SuperX kavitet								
(a) Sentrert					(b) MSC og sentrert			
Prøve [Gy]	Klasse	Uten CV	LOOCV		Prøve [Gy]	Klasse	Uten CV	LOOCV
0	1	1	2		0	1	1	2
0	1	1	1		0	1	1	1
0,8	1	1	1		0,8	1	2	1
0,9	1	1	1		0,9	1	1	1
1,3	1	1	2		1,3	1	1	2
2	2	2	1		2	2	1	1
3,3	2	2	2		3,3	2	1	2
4	2	2	2		4	2	1	2
10	2	2	2		10	2	2	2
Nøyaktighet		100 %	67 %				56%	67 %
Rektangulær kavitet								
(c) Sentrert					(d) MSC og sentrert			
Prøve [Gy]	Klasse	Uten CV	LOOCV		Prøve [Gy]	Klasse	Uten CV	LOOCV
0	1	1	1		0	1	1	1
0	1	1	1		0	1	1	1
0,8	1	1	1		0,8	1	1	2
0,9	1	1	1		0,9	1	1	1
1,3	1	2	2		1,3	1	2	2
2	2	2	1		2	2	2	1
3,3	2	2	1		3,3	2	2	2
4	2	2	2		4	2	2	2
10	2	2	2		10	2	2	2
Nøyaktighet		89 %	67 %				89 %	67 %

K-gjennomsnitt klassifisering (K-means)

Det ble utført en k-gjennomsnitt (K-means) analyse med to klasser, høy og lavdose. K-gjennomsnitt analyse ble utført på datasettene med både Mahalanobis og euklidsk avstand som avstandsmål, med henholdsvis ingen preprosessering, gjennomsnittsentrering og MSC som preprosesseringsteknikker.

Resultatene gitt i tabell 4-24 viser at det ble feilklassifisert i begge datasettene og for begge avstandsmål. I flere tilfeller ble 0 Gy prøver (tabell 4-24a,b,c) klassifisert i grupper høydose og 4 Gy og 10 Gy prøvene ble klassifisert i gruppen lavdose (tabell 4-24). Forskjellige preprosesseringer av spektrene påvirker i noen grad hvilke prøver som ble feilklassifisert, men økte stort sett ikke klassifiserings nøyaktigheten. K-gjennomsnitts metoden er ikke en god klassifiseringsmetode for disse datasettene.

Vedlegg 8.10 viser resultatene fra hver enkelt analyse beskrevet over med estimerte doser med og uten LOOCV. For alle modellene er det forskjell mellom klassifiseringen med og uten LOOCV. I vedlegg 8.10 vises også klassegjennomsnittspektrene som er benyttet.

Tabell 4-24, klassifisering av (a) SuperX kavitet datasettet med Mahalanobis avstand, (b) SuperX kavitet datasettet med euklidsk avstand, (c) rektangulær kavitet datasett med Mahalanobis avstand og (d) rektangulær kavitet datasett med euklidsk avstand. Klassifisering ut fra høy (2-10 Gy) og lavdose (0-1,3 Gy), med preprosessering: Ingen preprosessering, sentrering og MSC.

	Gitt dose [Gy]	Gitt klasse	Ingen	Sentrering	MSC
(a)	SuperX, Mahalanobis avstand				
	0	Lav	Lav	Lav	Høy
	0	Lav	Høy	Lav	Lav
	0,8	Lav	Høy	Lav	Høy
	0,9	Lav	Lav	Høy	Høy
	1,3	Lav	Lav	Høy	Lav
	2	Høy	Lav	Lav	Lav
	3,3	Høy	Høy	Lav	Lav
	4	Høy	Høy	Høy	Lav
	10	Høy	Høy	Lav	Høy
	Nøyaktighet		67 %	44 %	33 %
(b)	SuperX, euklidsk avstand				
	0	Lav	Lav	Lav	Lav
	0	Lav	Høy	Høy	Lav
	0,8	Lav	Høy	Høy	Høy
	0,9	Lav	Høy	Høy	Høy
	1,3	Lav	Høy	Høy	Lav
	2	Høy	Høy	Høy	Lav
	3,3	Høy	Høy	Høy	Lav
	4	Høy	Lav	Lav	Lav
	10	Høy	Høy	Høy	Høy
	Nøyaktighet		44 %	44 %	44%
(c)	Rektangulær, Mahalanobis avstand				
	0	Lav	Lav	Lav	Lav
	0	Lav	Lav	Høy	Lav
	0,8	Lav	Høy	Lav	Lav
	0,9	Lav	Høy	Høy	Høy
	1,3	Lav	Lav	Høy	Lav
	2	Høy	Høy	Høy	Høy
	3,3	Høy	Lav	Lav	Høy
	4	Høy	Lav	Lav	Lav
	10	Høy	Lav	Høy	Lav
	Nøyaktighet		44 %	44 %	67 %
(d)	Rektangulær, euklidsk avstand				
	0	Lav	Lav	Lav	Lav
	0	Lav	Lav	Lav	Lav
	0,8	Lav	Høy	Høy	Høy
	0,9	Lav	Høy	Høy	Høy
	1,3	Lav	Lav	Lav	Lav
	2	Høy	Lav	Lav	Lav
	3,3	Høy	Lav	Lav	Lav
	4	Høy	Lav	Lav	Lav
	10	Høy	Høy	Høy	Høy
	Nøyaktighet		44 %	44 %	44 %

5 Diskusjon

5.1 Formål

Denne oppgaven har fokusert på to forskjellige materialer som kan bli brukt i EPR-dosimetri. Det ene materialet er aminosyren alanin som tidligere har blitt brukt i bl.a. Vanhaelewyn et al. [1], Heydari et al. [9], Malinen et al. [11] og Callens et al. [24]. Fokuset i denne oppgaven har vært å karakterisere de underliggende EPR-spektrene av bestrålt alanin, ved å utforske flere kjente statistiske modeller for dekomponering av datasett i ladninger og skårer. Det andre materialet som har blitt undersøkt er Gorilla® Glass, brukt i Fattibene et al. [10]. Målet har vært å finne ut om det er mulig å estimere stråledoser eller klassifisere prøver i en høydose klasse eller lavdose klasse fra EPR-spektrene av bestrålt Gorilla® Glass.

5.2 Validering

Flere mål har blitt brukt til å vurdere modellene presentert i oppgaven. Disse målene har imidlertid noen begrensinger.

For predikerte målinger som passer veldig dårlig med den antatte modellen vil determinanskoeffisienten, R^2 kunne bli negativ. Det er i midlertid ikke normalt å operere med negativ R^2. Derfor har negative R^2 blitt satt lik null i denne oppgaven. Alle modellene med R^2-verdi null, kan derfor ikke sammenlignes direkte. En negativ R^2 betyr at den foreslåtte modellen er dårligere enn en horisontal linje [68], og at den estimerte modellen forklare mer av variansen i dataene enn det som finnes i den sanne modellen. Dette kommer av at gjennomsnittet av de estimerte $\hat{y}$ er større enn gjennomsnittet av de målte y.

Manglende tilpasning (LOF) verdien blir i noen tilfeller >1. Dette betyr at verdien til residualene er større enn de målte verdiene og at modellen ikke passer med måledataene. Dette skjer når fortegnet til det estimerte målepunktet og det målte punktet er forskjellig. Dette kan skjer når korrelasjonen mellom spektrene er lav eller negativ. I tilfeller med negativ korrelasjonen vil ikke LOF være et godt mål på hvor like spektrene er, siden LOF verdien da vil bli under 1 om spektrene blir flippet (skalert med -1, da blir korrelasjonen positiv). Derfor er LOF verdien mest interessant når korrelasjonen er høy.

RMSEC verdien øker alltid når flere komponenter blir tatt med i modellen, siden modellen klarer å tilpasse dataene bedre [46]. Derfor er RMSECV et bedre mål for hvor godt modellene

klarer å estimere prøvene. RMSECV verdien blir bare lavere dersom modellen blir bedre til å estimere nye prøver når flere komponenter legges til modellen [46].

5.3 Alanin datasettet

5.3.1 Om datasettet

Figur 4-36 viser at de teoretiske basisspektrene til R1 og R2 (figur 2-4) er mulig å identifisere ved å plukke ut enkelte av målingene i alanin datasettet. Dette gjør det vanskelig å konkludere om de statistiske modellene er egnet til å estimere de teoretiske spektrene til alanin eller ikke. Ingen av prøvene har høy korrelasjon med det teoretiske spekteret til R3 (figur 2-4c). Dette gjør at modellen som klarer å estimere komponenter som har høy korrelasjon med R1-, R2- og R3-spektrene, vil gi den mest treffsikre statistiske modellen.

I alanin datasettet er det stor avhengighet mellom de ulike spektre. De to største egenverdiene forklarer 95 % av den observerte variansen i datasettet (figur 4-35). Dette fører til at et fåtall komponenter er nok til forklare mesteparten av variansen som finnes mellom spektrene. Dette har en direkte betydning for PCA, MCR og ICA. I disse metodene forklarer den første komponenten mesteparten av den observerte variansen og de resterende forklarer mindre. Dette betyr ikke at informasjonen som komponent to og tre beskriver ikke er viktig informasjon, men at det er vanskelig å vite om det er nyttig informasjon eller kun støy som blir beskrevet. Dette er i overenstemmelse med Callens et al. [24] som skriver at hvis den totale variansen i datasettet primært blir forklart med den første faktoren fra MLCFA analyse, betyr det ikke at den andre faktoren er av mindre betydning. Callens et al. [24] skriver også at de estimerte felles faktorene ikke nødvendigvis forklarer noe om de sanne felles faktorene. Disse to påstandene har dette arbeidet vist at også stemmer for andre statistiske metoder, som PCA og ICA.

5.3.2 Antall radikalkomponenter

Ved å estimere andeler av R1 og R2 i de målte EPR-spektrene med minste kvadraters metode (tabell 4-1), kommer det fram at LOF verdien øker ved økende oppvarmingstemperatur og tid. Dette indikerer at andelen som ikke kan forklares med de teoretiske R1- og R2-spektrene, ikke er konstant igjennom hele datasettet og øker med økende temperatur. Det vil derfor være behov for en komponent som reagerer annerledes enn R1- og R2-radikalene ved oppvarming over

lengre tid. Behovet for denne tredje komponenten samsvarer med Vanhaelewyn et al. [1], Heydari et al. [9] og Malinen et al. [11].

PCA med MSC og sentrering som preprosessering av spektrene dannet prøvene en parabel sortert etter oppvarmingstid (figur 4-4b). Dette viser at det finnes sammenheng mellom antatt andel av R3 radikalet og PC1/2 verdi, ved at prøvene får lavere PC1 verdi med økende oppvarmingstid og temperatur.

Ved å innføre en komponent R3* estimert av langtidsoppvarmede prøver (figur 4-1b) går LOF verdien ned (vedlegg 8.2). Dette indikerer behovet for at en komponent R3 er nødvendig, for å kunne reprodusere måledataene som har blitt oppvarmet lenge. Alle estimatene på R1*-, R2*- og R3*-spektrene funnet i denne oppgaven gir en bedre LOF verdi (vedlegg 8.5) enn det som var tilfellet for de teoretiske R1- og R2-spektrene (tabell 4-1). Dette viser at uavhengig av formen til R1-, R2- og R3-spektrene vil tre komponenter gi en lavere LOF verdi enn kun de teoretiske R1- og R2-spektrene gjør.

For MCR analysen i kapittel 4.1.4, blir omtrent det samme R3*-spekteret identifisert uavhengig om 0-2 radikalspektre er antatt kjent. Dette viser at uavhengig av antagelse om R1- og R2-spekteret vil det samme spekteret dukke opp. Dette viser behovet for en komponent R3 som ikke bare er et støyspekter. Dette R3-spekteret kan bli brukt for å forklare mistilpasningen i LOF verdi (tabell 4-1). Når MCR blir gjort med fire komponenter dukker det opp et fjerde spekter som ser ut som støy. Dette viser at det kun er tre spektre som lar seg identifisere ved MCR i dette datasettet.

Resultatene i denne oppgaven tyder derfor på at det finnes et tredje radikal i bestrålt alanin.

5.3.3 Estimerte radikalspektre

Heydari et al. [9] foreslo hvordan spektrene til radikalene R1-, R2- og R3-spektrene ser ut basert på kvantemekaniske simuleringer. Disse R1- og R2- spektrene er vist i figur 2-4ab og er i denne oppgaven kalt teoretiske basisspektere. Tabell 4-13 viser at alle modellene som er testet klarer å estimere en komponent som ligner på det teoretiske R1-spekteret (korrelasjon 0,95-0,99). Dette bekrefter at det teoretiske R1-spekteret som Heydari et al. [9] fant, er mulig å identifisere ved hjelp av statistikk. En grunn for at det er mulig å identifisere det teoretiske basisspekteret til R1, kan være at R1-spekteret korrelerer (0,978) med spekteret til kontrollene (figur 4-36).

Det kan også antyde at bidraget fra R1-radikalet til EPR-spekteret er så sterkt at alle metodene klarer å gjengi R1-spekteret.

Tabell 4-13 viser at alle metodene klarer å finne estimater som ligner på det teoretiske R2-spekteret (korrelasjon >0,75). PCA er metoden som estimerer R2*-spekteret med lavest korrelasjon (korrelasjon 0,75). Modellen som gir best korrelasjon mellom R2* estimatet og det teoretiske R2-spekteret er MCR, med ingen kjente spektre og MSC som preprosessering (korrelasjon 0,91). Dette viser at R2-spekteret til Heydari et al. [9] er mulig å gjenkjenne ved statistiske metoder.

R3-spekteret vist i figur 2-4c, er en glattet variant av R3-spekteret vist i Heydari et al. [9]. Tabell 4-13 viser at det er veldig varierende korrelasjon mellom det teoretiske R3-spekteret og de estimerte R3*-spektrene (0,07-0,91). Dette viser at utseende på det estimerte R3*-spekteret er veldig avhengig av metoden som har blitt brukt for å estimere R3*-spekteret. Korrelasjonen er lavest for PCA og ICA, mens korrelasjonen blir høyest for minste kvadraters residual analyse og for MLCFA estimatet. Disse metodene tar utgangspunkt i utvalgte prøver noe som fører til en overtilpasning for disse prøvene, men som likevel funger godt til å gjengi et estimat på R3*-spekteret.

MCR, MLCFA og SMA gir gode estimater på R1-, R2- og R3- spektrene. Disse teknikkene egner seg til å finne reelle komponenter i spektre, siden ingen av teknikkene prøver å maksimerer variansen eller forutsetter at komponentene er lineært uavhengige, sånn som PCA og ICA gjør. Både MCR og SMA teknikkene er utviklet for å indentifisere de kjemiske komponentene i målinger. Modellene ICA og PCA er ikke egnet til å finne estimater for R3-spektret, men klarer til en viss grad å estimere R1- og R2- spektrene.

Vanhaelewyn et al. [1] presenterer et estimat på hvordan R1- og R2-spektrene skal se ut med utgangspunkt i MLCFA beregninger. Disse to spektrene ligner på faktor 1 og 2 (figur 4-11ab) fra MLCFA analysen (kapittel 4.1.3). Vanhaelewyn et al. [1] viser også en tredje komponent, men den blir ikke kalt R3. Denne tredje komponenten ligner på faktor 3 vist i figur 4-11c. Også denne oppgaven viser at faktor 3 fra MLCFA mest sannsynlig ikke er et godt estimat på R3-spekteret. En grunn er at ingen andre av modellene som er testet finner en komponent som ligner på faktor 3 (figur 4-11c).

Resultatene i denne oppgaven tyder derfor på spektrene til R1-, R2- og R3-radikalene, er mulig å estimere ut fra statistiske analyser med teknikkene MCR og SMA. MLCFA klarer også å finne R1-, R2- og R3-spektrene, men ikke i samme analyse.

Preprosessering av spektrene

Preprosessering av alanin datasettet ga ikke store utslag på de identifiserte komponentene. De fleste analysene forblir relativt like uavhengig av om preprosessering er benyttet eller ikke. Preprosesseringsteknikkene MSC og EMSC gir omtrent det samme resultatet på spektrene (vedlegg 8.1). Dette kommer av at EMSC bygger på MSC [40]. MSC og normalisering til enhetsareal gjør at de preprosesserte spektrene blir nesten helt identiske (vedlegg 8.1), siden MSC er en form for normalisering, som skalerer vek forskjellene i datasettet. For PCA er det stor forskjell i resultatet om spektrene blir sentrerte (figur 4-3) eller MSC og sentrerte (figur 4-4). Dette viser at MSC er en viktig preprosesseringsteknikk hvis målet med PCA er å finne sammenhenger i skårplottet, men hvis målet er å estimere R1-, R2- og R3-spektrene er kun sentrering bedre.

Dersom alanin spektrene skal preprosesseres er det tilstrekkelig å bruke sentrering eller normalisering. Denne oppgaven viser at mer avanserte preprosesseringer, som MSC, ikke nødvendigvis påvirker de identifiserte spektrene.

En vanlig preprosesseringsteknikk som ikke er benyttet på alanin datasettet er autoskalering [50]. I autoskalering sentreres dataene og deles på standardavviket. Dette er ikke en fornuftig preprosesseringsteknikk for dette datasettet, siden standardavviket for noen av variablene er veldig lite, noe som medføre oppblåsing av disse variablene. Variablene som har lavt standardavvik er primært variablene med høyt og lavt magnetfelt.

Vurdering av metodene

PCA, MLCFA, MCR, SMA og ICA er alle eksempler på bilineære modeller. Metodene dekomponerer datasettet **X** inn i to komponenter, på formen $\mathbf{X} = \mathbf{AB}^T$. Forskjellen mellom de ulike metodene ligger i hvordan **A** og **B** blir funnet, spesielt initialisering av algoritmene er forskjellig.

Metoden som skiller seg mest ut fra de andre metodene, er bruken av residualene som estimat til R3*-spekteret (kapitel 4.1.1). Denne metoden gir bare god tilpasning dersom R3*-spekteret estimeres fra langtidsoppvarmede prøver, hvor andelen av R3 kan antas å være høy (figur 4-1b). Dette R3*-spekteret vil ha en korrelasjon på null med R1 og R2. Residual metoden egner seg derfor ikke for å finne estimater på R3-spekteret, siden modellen tar utgangspunkt i den informasjonen som ikke lar seg beskrive av de teoretiske spektrene til R1 og R2, må tilfalle R3-spekteret.

PCA (kapittel 4.1.2) skiller seg fra de andre metodene ved at PCA ladningene blir lineært uavhengige. Dette er en ulempe når målet er å estimere R1-, R2- og R3-spektrene, siden det ikke foreligger noen grunn til å anta at spektrene til R1, R2 og R3 er lineært uavhengige av hverandre. Dette gjøre at prinsipalkomponentene ikke fungerer til å finne estimater på de underliggende komponentene til alanin.

MLCFA (kapittel 4.1.3) leter etter felles faktorer som forklarer korrelasjon mellom de opprinnelige variablene. MLCFA resultatene er uavhengige av preprosessering, siden korrelasjonene mellom variablene er omtrent den samme uavhengig av prosesseringsteknikk. Dette gjør MLCFA til en robust metode for å finne underliggende faktorer. Vanhelewyn et al. [1] gjorde en lignende analyse som denne oppgaven med MLCFA, og viste at det var mulig å finne R1- og R2-spektre ved hjelp av MLCFA, men ikke R3-spektret. Dette er i overenstemmelse med denne oppgaven (figur 4-11). Det beste estimatet på R3 som er identifisert med MLCFA kommer fra et redusert datasett bestående av langtidsoppvarmede prøver. For dette utvalget faller antallet identifiserte komponenter til to, i overenstemmelse med Vanhelewyn et al. [1]. Dette utvalget klarer å estimere R3*-spekteret og en blanding av R1*- og R2*-spekteret (figur 4-12). MLCFA klarer å finne estimater på R1-, R2- og R3-spektrene, men MLCFA klarer ikke å estimere dem i samme analyse.

Flere ulike betingelser kan brukes i en MCR analyse, og disse vil kunne påvirke skårene og ladningene. Dette gjør det vanskelig å vite med sikkerhet om den beste løsningen er funnet. Den MCR estimerte komponenten for R3-spekteret (figur 4-25), er relativ lik uavhengig av preprosessering og antallet kjente spektre. Dette viser at MCR er relativt robust med hensyn på de estimerte komponentene.

SMA bygger på at det finnes noen rene variabler som bare påvirker den ene av de underliggende spektrene. I tilfellet med alanin spektre er dette en antagelse som ikke er ideell. For eksempel, er det ingen grunn til å anta at for ren variabelen 1 skal komponent 1 ha en verdi, mens komponent 2 og 3 må krysse *x*-aksen i det samme punktet (figur 4-27). Dette gir en kraftig begrensning i hvordan spektrene ser ut. En annen svakhet med SMA (kapittel 4.1.5) er at om ikke α-verdien i ligning (14) velges riktig risikeres det deling på null eller at α-veriden blir dominerende. I begge tilfeller blir renhetsvektoren vanskelig å tolke eller feil.

De R1-, R2- og R3-spektrene som blir estimert i forskjellige kjøringer av ICA (figur 4-29) har forskjellig rekkefølge og er skalert med -1, siden ulike rotasjonsmatriser blir funnet når ulike startpunkter benyttes. Ifølge Hyvärinen [55] er dette en måte å se at komponentene som er

identifisert er de riktige. Komponentene funnet ved PCA og ICA er like. Dette kommer av at både PCA og ICA går ut fra en SVD transformasjon.

Resultatene fra denne oppgaven viser at SMA og MCR kan benyttes for å estimere alle tre radikalspektrene, mens ICA og PCA ikke er egnet til å estimere alle tre radikalspektrene.

5.3.4 Mengder og andeler av radikalkomponentene

Vanhaelewyn et al. [1] og Heydari et al. [9] viser ulike andeler av de ulike radikalkomponentene. Vanhaelewyn et al. [1] baserer seg på temperatur behandlete prøver og konkluderer med at mengden R3-radikal i prøvene går raskt nedover med økende temperatur, sammenlignet med R1- og R2-radikalene. Dette er motstridene til Heydari et al. [9] som baserer seg på kvantemekaniske simuleringer og finner at mengden R3-radikal går sakte ned med økende temperatur, sammenlignet med R1- og R2-radikalene. Funnene i denne oppgaven følger trenden til Heydari et al. [9] (figur 4-31).

Heydari et al. [9] finner at andelene ved romtemperatur av de teoretiske R1/R2/R3-spektrene er 59/33/8 %. Fra figur 4-31 kommer det at andelsestimeringen med residualen fra minste kvadraters metode stemmer bra (59/39/2 %) for alle tre komponentene ved romtemperatur. De andre modellen overestimerer andelen R1* i forhold til Heydari et al. [9]. SMA er den analysen som er lengst unna de andelene som er beskrevet i Heydari et al. [9]. SMA estimerer andelene ved romtemperatur til å være 22/52/26 %.

Ved oppvarming til 207 ^{o}C i 40 minutter kommer Heydari et al. [9] fram til at andelene til de teoretiske R1/R2/R3-spektrene er 6/51/43 %. Andelsestimeringen ved temperatur 205 ^{o}C (figur 4-31b) viser at MCR og SMA klarer å estimere andelene i omtrent samme størrelse som Heydari et al. [9] etter oppvarming i 80 minutter. Det ser ut til at andelen R3* estimert ved SMA og MCR øker saktere enn det Heydari et al. [9] beskriver, siden andelen av R3* etter 40 minutter er på 10-20 % (figur 4-31b). De andre metodene klarer ikke å estimere andeler som ligner på dem presentert i Heydari et al. [9]. Alle modellene og temperaturene har samme trend som Heydari et al. [9] ved at de får en lavere andel R1* og høyere andel R2* og R3* med økende tid (figur 4-31).

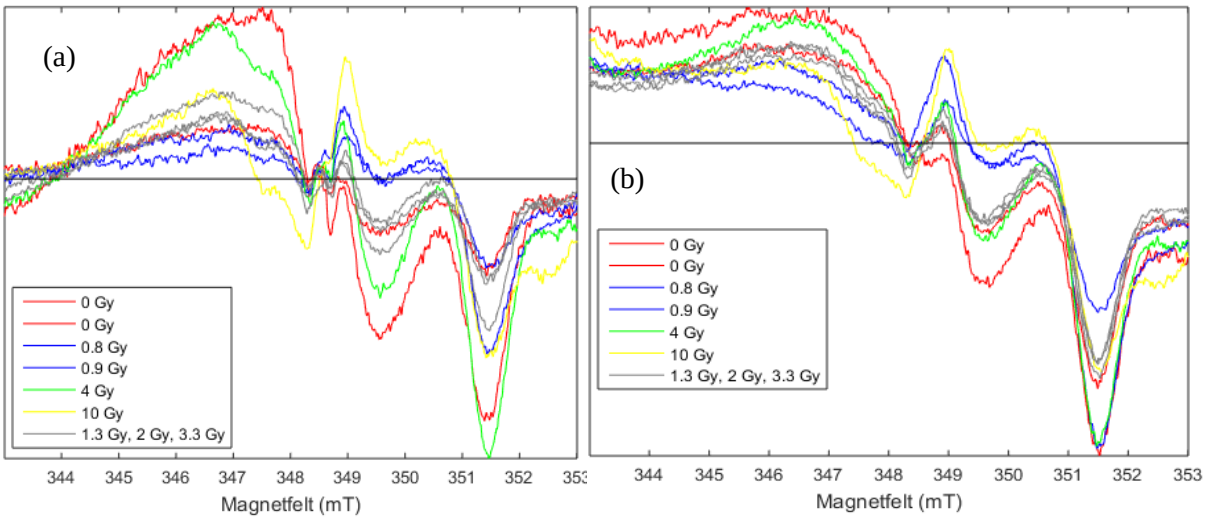
5.4 Gorilla® Glass datasettene

5.4.1 Om datasettene

I Gorilla® Glass datasettene er det totalt ni prøver med 1024 variabler. Dette er et stort antall variabler i forhold til prøver. Variabelreduksjon brukes derfor for å hindre at modellene overtilpasser prøvene.

Dosene som Gorilla® Glass prøvene absorberte er ujevnt fordelt med seks prøver ≤2 Gy og tre prøver >2 Gy. Når 10 Gy prøven skal predikeres med LOOCV, blir estimatet dårlig, siden ingen andre prøver har høyere enn 4 Gy i absorbert dose. Modellen må derfor ekstrapoleres til 10 Gy.

EPR-spektrene i SuperX kavitet datasettet er også inkonsistente. Den ene 0 Gy prøven ligner mer på 4 Gy prøven enn på den andre 0 Gy prøven (figur 5-1a). I tillegg har 4 Gy spekteret mer ekstreme verdier enn det 10 Gy spekteret (figur 5-1a). Dette er en svakhet ved datasettet og vanskeligjør presise estimeringer og klassifiseringer. Dette gjør at de variablene som er funnet til å være de viktigst i regresjonsmodellene ligger stort sett i områder hvor variansen imellom spektrene er lav.



Figur 5-1, (a) SuperX kavitet og (b) rektangulær kavitet datasettene. Rød er de to 0 Gy prøvene, blå er 0,8 Gy og 0,9 Gy, grønn er 4 Gy og gul er 10 Gy, resten av prøvene er grå.

Rektangulær kavitet datasettet er noe mer konsistent, men spekteret til 4 Gy prøven ligger imellom spektrene til de to 0 Gy prøvene (figur 5-1b). Dette er ikke logisk og vil være med på å gjøre modellene dårligere.

Hotelling T^2 mot Q residualene (figur 4-37c, figur 4-38c og vedlegg 8.7) viser at det mest sannsynlig ikke er noen uteliggere i datasettet. Dette er i overenstemmelse med konklusjonen

til Fattibene et al. [10]. I regresjonsmodellene som er testet viser det seg at spesielt 0 Gy og 10 Gy prøvene har vært vanskelig å estimere nøyaktig.

Det er ikke store forskjeller om datasettene har blitt preprosessert eller ikke, og det er små forskjeller i hvilken preprosesseringsteknikk som gir best estimering av doser for de to datasettene. For SuperX kavitet gir, for eksempel, sentrering bedre doseestimering enn MSC og sentrering for PLS og PCR (tabell 5-1). For rektangulær kavitet er det motsatt.

5.4.2 Regresjonsmodeller for doseestimering

I det opprinnelige studiet med Gorilla® Glass av Fattibene et al. [10] ble dosene estimert ved å lage et referansespekter ved å sammenligne et kontrollspekter med absorbert dose 0 Gy mot 10 Gy spekteret. Dette referansespekteret ble så brukt til sammenligne de andre EPR-spektrene for å estimere doser. Referansespekteret ble så benyttet til å lage en lineær regresjonsmodell basert på differansen mellom referansespekteret og de målte EPR-spektrene. En svakhet med denne metoden er, at hvis referansespekteret blir forandret vil det påvirke hele modellen. I denne oppgaven er det fokusert på metoder som baserer seg på å bruke alle målespektrene til å lage regresjonsmodeller og noen variabelseleksjonsmetoder.

Tabell 5-1 viser at alle modellene har lav RMSEC verdi. Dette indikerer at det er mulig å lage en modell som er tilpasset alle prøvene. Ved kryssvalidering blir RMSECV høyere for alle modellene, spesielt for modellene regresjon med alle variablene og IPLS med intervallbredde 10 og 100. Dette viser at de fleste modellene ikke er godt egnet til å estimere nye prøver, noe som indikerer overtilpasning av kalibreringsdatasettet.

Siden regresjon med alle variablene gir en veldig høy RMSECV verdi (tabell 5-1), viser dette at det er nødvendig å redusere antallet variabler/komponenter som blir brukt i regresjonsmodellen. Variabelreduksjon viser at ved å gå fra 1024 variabler og ned til 2-4 komponenter blir RMSEC verdien litt høyere (tabell 5-1) og RMSECV verdien mye lavere. Dette betyr at modellen blir bedre til å estimere doser på nye prøver og at informasjonen som trengs for å estimere doser kan fanges opp i kun noen fåtall variabler.

Grunnen til at regresjon med 1024 variabler får en RMESC verdi på null (tabell 5-1) er at, en modell med flere variabler enn prøver alltid vil klarer å lage en regresjonsmodell som forklarer prøvene perfekt [61]. Dette vil gi en modell som overtilpasser prøvene i kalibreringsdatasettet

og ikke er egnet til å estimere nye prøver. Dette er også tilfellet for lasso med ni variabler (tabell 5-1).

IPLS med bredde 1 har den laveste RMSECV verdien (tabell 5-1). Dette kommer av at de ni variablene som blir funnet er plukket ut ved å minimere RMSECV. Siden IPLS får en så lav RMSECV er det grunn til å tro at variablene som har blitt plukket ut overtilpasser datasettet, dette fordi IPLS med full LOOCV (figur 4-53) har en betydelig dårligere RMSECV verdi for begge datasettene (tabell 5-1). Dette viser at IPLS er best til å finne viktige variabler i et kalibreringsdatasett, men fungerer ikke like bra til å finne variabler som er egnet til å estimere doser til ukjente prøver.

Tabell 5-1, samling av RMSEC og RMSECV verdiene for regresjonsmodellene modellene, med SuperX kavitet og rektangulær kavitet datasettene.

	SuperX kavitet				**Rektangulær kavitet**		
	Antall komponenter	RMSEC	RMSECV		Antall komponenter	RMSEC	RMSECV
Regresjon							
Konstantledd	1024	0	23,8		1024	0	7,37
Sentrert	1024	0	38,8		1024	0	30,0
PCR							
Sentrert	3	0,44	1,73		4	0,74	2,68
MSC og sentrert	2	1,08	2,38		4	0,62	2,28
PLS							
Sentrert	3	0,34	1,84		4	0,66	2,45
MSC og sentrert	2	0,88	2,32		3	0,7	2,29
IPLS							
Bredde 1	7	3,9E-07	1,5E-05		7	3,7E-06	8,80E-05
Bredde 10	20	0	17,5		10	0	19,9
Bredde 100	200	0	39,2		100	0	14,1
Variabel reduksjon							
IPLS	3	0,77	1,62		3	1,10	2,73
Maks varians	3	1,34	3,69		3	1,25	3,33
LASSO							
Beste antallet	9	0	2,33		9	0	2,78
Redusert antall	2	1,87	3,17		3	1,34	2,62
IPLS redusert							
	3	1,9E-07	1,82		3	2,3E-05	3,44
IPLS LOOCV							
			4,16				3,92

Den modellen som har lavest RMSECV verdi, sett bort fra IPLS, er for PCR brukt på SuperX kavitet datasettet med sentrering (tabell 5-1). For rektangulær kavitet datasettet er metodene PLS med tre komponenter og PCR med fire komponenter like gode når det kommer til RMSECV (tabell 5-1). Dette viser at det finnes flere metoder som er like gode til å lage regresjonsmodeller for rektangulærkavitet datasettet, som kun bruker et fåtall komponenter. PCR bygger på prinsipalkomponentene og PCR modellen har derfor med informasjon om alle variablene. Dette gjør PCR modellen robust. Det samme gjelder for PLS.

Variabelreduksjon ved å plukke ut de variablene som har høyest varians, klarer ikke å finne de variablene som er best egnet til å estimere nye prøver. Grunnen til dette er at 0 Gy prøvene og 4 Gy prøven følger hverandre igjennom hele datasettet (figur 5-1a), dette gjør at det å plukke ut variabler kun etter hvor det er mest varians ikke vil være egnet til å finne de variablene som skiller dosene maksimalt.

Fattibene et al. [10] konkluderer med at det er mulig å estimere dosene i området >1,5 Gy med 20 % nøyaktighet og prøven 3,3 Gy med 5 % nøyaktighet. Ut fra resultatene som Fattibene et al. [10] presenterer er det vanskelig å se hvordan disse nøyaktighetene dukker opp.

Ved å regne ut R^2 på dosene 0 Gy, 0,9 Gy, 1,3 Gy og 3,3 Gy på dataene til Fattibene et al. [10] kommer det fram at R^2 blir 0,66. I denne oppgaven blir regresjonsmodellen laget med PCR på SuperX kavitet med sentrering, ved bruk av kryssvalidering (figur 4-41) estimert til å ha en R^2 på 0,60 på de samme målingene. Dette viser at estimeringen funnet i denne oppgaven er litt dårligere enn det Fattibene et al. [10] fant. En grunn til forskjellen er at Fattibene et al. [10] har delt datasettet opp i et kalibreringssett og et testsett, mens det i denne oppgaven har det blitt brukt LOOCV som validering. Dette spiller inn på hvor mange spektre som er med å kalibrere modellen. En annen grunn til forskjellen kommer av at R^2 til Fattibene et al. [10] blir regnet ut fra resultatene til seks forskjellige målinger.

Ved å sammenligne resultatene som Fattibene et al. [10] får, med PCR av sentret SuperX datasettet (figur 4-41) med kryssvalidering, kommer det frem at avviket i estimeringen av 0 Gy og 0,9 Gy prøvene har lavere avvik enn gjennomsnittsavviket for Fattibene et al. [10], mens 1,3 Gy er akkurat på gjennomsnittet og 3,3 Gy har noe større avvik enn Fattibene et al. [10]. Dette viser de kryssvaliderte verdiene funnet i denne oppgaven er omtrent like gode som det Fattibene et al. [10] fant.

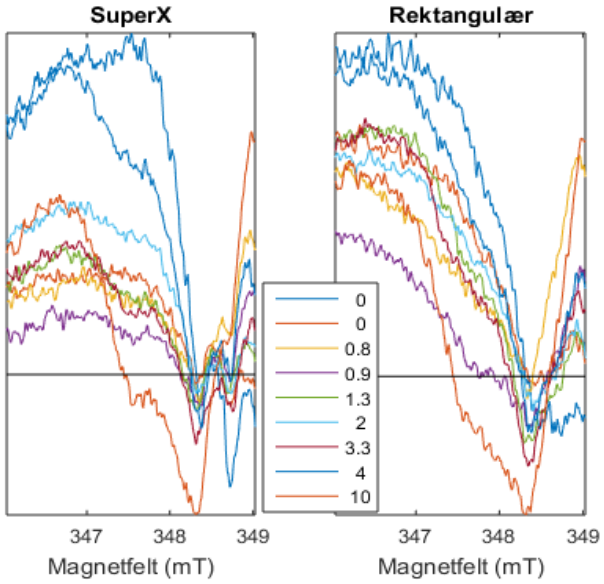
Modellen som Fattibene et al. [10] bruker til å estimere doser er basert på et referansespekter. Denne oppgaven viser at det ikke er nødvendig å benytte et referansespekter for å kunne

estimere dosene. Ved å bruke lineærkombinasjoner (PCR, PLS) for å estimere doser blir modellene mer robuste mot målefeil og lettere å bruke til å estimere nye modeller.

Alle metodene som er testet i denne oppgaven har forskjellige variabler som blir ansett som viktige for modellene (tabell 5-2). Metodene benytter forskjellige deler av spektrene, men de viktigste variablene er stort sett midt i måleområdet (346-349 mT). Dette intervallet dekker området som det er størst varians imellom spektrene (346-348 mT) og det området hvor spektrene er mest like (348-349 mT). Dette viser at enkelte av metoden som er testet ikke er avhengig av en visuell sammenheng mellom spekterutseende og estimert dose (figur 5-1).

Tabell 5-2, de viktigste variable (mT) for metodene PCR, PLS, IPLS, Regresjon og lasso for SuperX kavitet og rektangulær kavitet datasettene, funnet i kapittel 4.2.

	SuperX	Rektangulær
PCR/PLS	346-348,5	346-348,5
IPLS	348-349	351-352
Regresjon	343-344 og 348-349	348-349
Lasso	344-345	Hele måleområdet
Maksvarians	347,5, 349,5 og 351,5	347,5, 349,4 og 351,5



Figur 5-2, de viktigste variablene (346-349 mT) for SuperX og rektangulær kavitet datasettene

For å hindre overtilpasning er det nødvendig å gå ned på antallet variabler/komponenter. Fra tabell 5-1 kommer det fram at modeller med 2-4 variabler/komponenter gir en grei doseestimering. Grunnen til at flere variabler ikke lønner seg er at med flere komponenter enn prøver vil det alltid være mulig å tilpasse modellen på en sånn måte at alle prøvene i kalibreringsdatasettet blir korrekt estimert [61]. Dette er ikke ideelt siden det er ønskelig å bruke modellen til å estimer doser til nye prøver.

I denne oppgaven blir det funnet at det er bedre å bruke noen beregnede komponenter (PLS eller PCR) enn å benytte noen av de originale variablene (sånn som lasso eller IPLS) i regresjonsmodellen. Fra tabell 5-1 kommer det fram at RMSECV verdien blir omtrent lik for PCR og PLS. Derfor er det vanskelig å konkludere med hvilken av disse metodene som er best, men disse metodene gir en mer robust regresjonsmodellen enn bruk av IPLS-variabler.

5.4.3 Klassifisering ut fra lav eller høy absorbert dose

For begge datasettene blir alle prøvene klassifisert riktig med LDA, hvis Mahalanobis avstand blir benyttet (tabell 4-21a og tabell 4-22a). Når euklidsk avstand blir benyttet (tabell 4-21b og tabell 4-22b) blir noen av prøvene klassifisert feil. Dette kan skyldes at avstandene mellom prøvene blir best beskrevet med ellipser, som beskriver standardavviket vek fra klassegjennomsnittet til prøvene i alle retninger [69]. Dette viser at datasettene klassifiserer prøvene best når det er lik varians mellom variablene.

Klassifisering med euklidsk LDA (tabell 4-21b og tabell 4-22b) og k-gjennomsnittsalgoritme (tabell 4-24) klassifiserer enten de to 0 Gy prøvene i forskjellige klasser eller den ene 0 Gy i samme klasse som 10 Gy. En grunn for dette er at gruppegjennomsnittspektrene (vedlegg 8.10 og figur 4-62) er ganske like, samt at forskjellen til gruppegjennomsnittspektrene er i områdene som ved visuell undersøkelse ikke inneholder noen sammenheng mellom dose og spekterutseende (figur 5-1). Dette gjør det vanskelig å klassifisere dosene riktig. Grunnen til at LDA med de samme gruppegjennomsnittspektrene klarer å estimere riktig med Mahalanobis avstand, er at Mahalanobis avstand forvrenger gruppegjennomsnittspektrene ved å skalere med kovariansmatrisen og det gjør at de små forskjellene som er i gruppegjennomsnittspektrene blir tydeligere. K-gjennomsnittsalgoritmen klassifiserer med samme nøyaktighet uavhengig av preprosessering og avstandsmål (tabell 4-24). Dette skyldes at algoritmen startes uten noen kunnskap om hvordan det sanne gruppegjennomsnittspekteret ser ut, i motsetning til LDA som starter med å regne ut gjennomsnittspekteret for alle prøvene i en klasse i kalibreringsdatasettet.

Det kommer fram at PLSDA klassifiserer (tabell 4-23) bedre enn k-gjennomsnitts klassifisering (tabell 4-24) og euklidsk LDA (tabell 4-21b og tabell 4-22b). Dette kommer av at PLSDA gjør diskriminant analyse på PLS komponentene, og derfor bygger på informasjonen om hvilken retninger som maksimerer kovariansen mellom variablene og responsen (her klasse 0/1), noe som betyr at PLSDA har mye tilfelles med Mahalanobis avstand ved at de begge bygger på kovariansmatriser. Den største forskjellen mellom LDA med Mahalanobis avstand og PLSDA er måten kovariansmatrisen blir funnet. En annen forskjell mellom LDA og PLSDA er at LDA bygger på hele spektrene mens PLSDA bygge på et lite utvalg komponenter. Dette medfører at den viktigste informasjonen ikke klarer å bli beskrevet godt nok med PLS med 2-4 komponenter for å bli brukt i en diskriminant analyse.

Klassifisering med LDA og Mahalanobis avstand (tabell 4-21a og tabell 4-22a) gir korrekt klassifisering (over eller under 2 Gy). Dette resultatet er bedre enn noe som ble oppdaget med regresjon med LOOCV (tabell 5-1). Dette betyr at hvis målet kun er å skille mellom prøver som

har absorbert en høy/lavdose, er klassifisering best å bruke, men hvis målet er å finne mer eksakte doser, er regresjon å anbefale. Spesielt ved bruk av PCR og PLS har denne oppgaven vist at det finnes en sammenheng mellom EPR-spektrene og absorbert dose for Gorilla® Glass.

5.5 Videre arbeid

For å finne enda bedre estimater på R1-, R2- og R3-spekteret kan det være interessant å lage et tilsvarende alanin datasett med flere prøver, temperaturer og tider, for å kunne kvalitetssikre resultatene fra dette arbeidet.

For Gorilla® Glass datasettet trengs det et nytt datasett om det skal være mulig å konkludere om Gorilla® Glass egner seg som dosimeter eller ikke. Det nye datasettet bør ha flere prøver, enten prøver i intervallet 0-5 Gy eller 0-10 Gy avhengig av hvor mange prøver det er praktisk å lage. De absorberte dosene burde i alle tilfeller være nærmere hverandre, da spranget fra 4 Gy og opp til 10 Gy vanskeligjør validering av modellen. Det bør være omtrent like mange prøver over 2 Gy som under, siden 2 Gy er klassifiseringsgrensen mellom høy og lav.

For å kunne bruke regresjon med alle variablene uten å utføre en variabelreduksjon bør det være mer enn 1024 prøver i kalibreringsdatasettet. Dette er ikke et en realistisk mengde prøver, men hvis PCR med tre komponenter skal brukes kan det holde med 14 prøver. Dette er nok et litt lavt tall da flere prøver kan tenke seg å trenge flere PCR komponenter for å gi en optimal estimering. Et nytt eksperiment bør ha så mange prøver som det er praktisk mulig å lage, men ikke mindre enn 14 for at residualen skal kunne ha 10 frihetsgrader ved bruk av tre PCR komponenter [70]. Med kun 14 prøver bør den høyeste dosen være 5 Gy.

Alternativt til å utføre et helt nytt forsøk kan være å gjøre de samme analysene som er vist i dette arbeidet, bare uten 10 Gy prøven. Dette kan gi bedre resultater, men vil medføre at det er enda færre prøver til å lage modellen. Det kan også være mulig å analysere de andre datasettene som Fattibene et al. [10] produserte for å se om de gir samme eller bedre resultater ved bruk av metodene i denne oppgaven.

6 Konklusjon

Denne oppgaven viser at det er sannsynlig at det finnes et R3-radikal i bestrålt alanin som overlapper med R1 og R2 bidragene i EPR-spekteret. Det er blitt vist at andelen til R3-radikalet mest sannsynlig oppfører seg annerledes enn andelene til R1- og R2-radikalene. R3-spekteret må innføres for å kunne forklare den manglende tilpasningen mellom målespektre og de teoretiske R1- og R2-spektrene.

Analysene av alanin datasettet viser at det er mulig å estimere R1-, R2- og R3-spektrene relativt likt som de kvantemekaniske simuleringene til Heydari et al. [9]. Metodene MCR og SMA klarer å estimere alle tre komponentene i samme analyse. MLCFA klarer å estimere R1- og R2-spektrene i en analyse og R3-spekteret i en analyse med utvalgte EPR-spektre. ICA og PCA klarer å estimere R1- og R2-spektrene, men ikke R3-spekteret. Alle metodene gir entydig estimering av R1-spekteret, og nesten entydig estimering av R2-spekteret. R3-spekteret blir estimert likt av MCR, MLCFA og SMA. ICA og PCA gir den samme tredje komponenten, men denne er en støykomponent. Det er også mulig å estimere R3-spekteret ved hjelp av residualanalyse av en minste kvadraters tilpasning med de teoretiske spektrene til R1 og R2. Alle de identifiserte R3-spektrene har en høyere andel støy enn de identifiserte komponentene til R1- og R2-spektrene.

Preprosessering av alanin datasettet ga ikke store forandringer i de identifiserte komponentene.

I alanin datasettet finnes det EPR-spektre som ligner på de teoretiske spektrene til R1 og R2. Dette gjør at korrelasjonen er høy mellom de identifiserte komponentene og enkelte av målespektrene.

De ulike metodene gir forskjellige estimater av andelene av R1, R2 og R3 i de målte EPR-spektrene. Alle metodene gir den forventede utviklingen av komponentene ved økt temperatur og tid, men andelene spriker med kjent teori [9].

Alt tatt i betraktning er MCR og SMA de metodene som estimerer spekterkomponentene og andelene best i forhold til tidligere studier. Av disse to metodene er MCR å foretrekke siden SMA tar utgangspunkt i at det finnes noen rene variabler i spektrene.

Denne oppgaven har også sett på om det er mulig å bruke Gorilla® Glass som dosimeter. Det kommer frem at det er mulig å lage regresjonsmodeller som klarer å estimere prøvene i kalibreringsdatasettet relativt nøyaktig, men det er ikke funnet noen metoder som klarer å

estimere prøvene ved kryssvalidering nøyaktig. Spesielt 10 Gy prøven har vist seg være vanskelig å tilpasse modellene.

Regresjonsmodellen som gir beste estimering av doser ved kryssvalidering er IPLS med syv variabler, men IPLS-variablene lar seg ikke reprodusere ved kryssvalidering.

For videre bruk er enten PCR eller PLS de anbefalte modellene. Disse to modellene gir omtrent samme RMSECV-verdi.

Det er forskjeller mellom de to datasettene på hvilken metode og preprosessering som er best. For SuperX kavitet er det PCR med sentrering som er best. Mens for rektangulær kaviteter er det PCR med MSC og sentrering som preprosessering, som er best.

Det er mulig å inndele prøvene inn i klasser ut fra høy/lavdose ved bruk av LDA med Mahalanobis avstand. K-gjennomsnittsalgoritme og PLSDA er ikke egnet til å finne klassene uavhengig av preprosessering og om Mahalanobis eller euklidsk avstand blir benyttet.

7 Referanser

[1] G. C. A. M. Vanhaelewyn, S.A. Amir, W. K. P. G. Mondelaers, and F. J. Callens, "Decomposition study of the electron paramagnetic resonance spectrum of irradiated alanine," *Spectrochimica Acta Part A*, pp. 387-397, 2000.

[2] J. Lilley, *Nuclear Physics - Principles and Applications*. West Sussex, England : John Wiley & Sons Ltd, 2001.

[3] H. Rosoff and D. von Winterfeldt, "A Risk and Economic Analysis of Dirty Bomb Attacks," *Risk Analysis, Vol. 27, No 3*, pp. 533-546, 2007.

[4] D. Regulla, "From dating to biophysics — 20 years of progress in applied ESR spectroscopy," *Appl Radiat Isot 52*, pp. 1023-1030, 2000.

[5] E. Sagstuen and E. O. Hole, "Radiation Produced Radicals," in *Electron Paramagnetic Resonance*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2009, pp. 325-382.

[6] E. Malinen, "EPR Dosimetry in Clinical Applications," in *Applications of EPR in Radiation Research*. Cham, Sveits: Springer International Publishing, 2014, pp. 509-538.

[7] W. W. Bradshaw, D. G. Cadena, G. W. Crawford, and H. A. W. Spetzler, "The Use of Alanine as a Solid Dosimeter," *Radiation Research, Vol. 17, No. 1*, pp. 11-21, 1962.

[8] E. Sagstuen, E. O. Hole, S. R. Haugedal, and W. H. Nelson, "Alanine Radicals: Structure Determination by EPR and ENDOR of Single Crystals X-Irradiated at 295 K," *J. Phys. Chem. A 101*, pp. 9763-9772, 1997.

[9] M. Z. Heydari, E. Malinen, E. O. Hole, and E. Sagstuen, "Alanine Radicals. 2. The Composite Polycrystalline Alanine EPR Spectrum Studiet by ENDOR, Thermal Annealing, and Spectrum Simulations," *J. Phys. Chem. A*, pp. 8971-8977, Aug. 2002.

[10] P. Fattibene, F. Trompier, and et. al., "EPR dosimetry intercomparison using smart phone touch screen glass," *Radiat Environ Biophys*, pp. 311-320, 2014.

[11] E. Malinen, M. Z. Heydari, E. Sagstuen, and E. O. Hole, "Alanine Radicals, Part 3: Properties of the Components Contributing to the EPR Spectrum of X-Irradiated Alanine Dosimeters," *Radiation Research Society*, pp. 23-32, 2003.

[12] J. A. Weil and J. R. Bolton, *Electron Paramagnetic Resonance, Elementary Theory and Practical Applications, Second Edition*.: John Wiley & Sons, Inc., 2007.

[13] C. Corvaja, "Introduction to Electron Paramagnetic Resonance," in *Electron Paramagnetic Resonance*.: John Wiley & Sons. Inc., 2009, pp. 3-35.

[14] E. Malinen and E. Sagstuen, *FYS390 - Elektron spinn resonans (ESR) spektroskopi*. Oslo, Norge: Fysisk institutt, Universitetet i Oslo, 2001.

[15] IAEA, *Use of electron paramagnetic resonance dosimetry with tooth enamel for retrospective dose assessment*. Wien, Østerriket: IAEA, 2002.

[16] Professor E. Malinen. (2016) Fysisk institutt, Universitetet i Oslo, Personlig kommunikasjon.

[17] Institutt for Energiteknikk. (20.04.16) Hva er radioaktivitet? [Online]. https://www.ife.no/no/ife/avdelinger/miljo_og_stralevern/faq-no/radioaktivitetogstraling

[18] Mirion Technologies. (20.04.16) Types of Ionizing Radiation. [Online]. https://www.mirion.com/introduction-to-radiation-safety/types-of-ionizing-radiation/

[19] G. C. A. M. Vanhaaelewyn, W. K. P. G. Mondelaers, and F. J. Callens, "Effect of Temperature on the Electron Paramagnetic Resonance Spectrum of Irradiated Alanine," *Radiation Research Society*, pp. 590-594, 1999.

[20] J. Tausjø and O.Klepp. (15.02.09 besøkt 11.04.16) Stråleskader, Store Norske Leksikon. [Online]. https://snl.no/str%C3%A5leskader

[21] K. Hofstad. (23.03.15 besøkt 05.04.16) Sievert, Store Norske Leksikon. [Online]. https://snl.no/sievert

[22] Universitetet i Oslo Fysisk institutt, *Laboratorieoppgaver i biofysikk - FYS290*. Oslo, Norge: Fysisk institutt, Universitetet i Oslo , Høstsemesteret 1996.

[23] P. Kierulf. (12.05.15 besøkt 23.04.16) Store medisinske leksikon. [Online]. https://sml.snl.no/aminosyrer

[24] F. Callens, K. van Laere, W. Mondelaers, P. Matthys, and E. Boesman, "A Study of the Composite Character of the ESR Spectrum of Alanine," *Appl. Radiat. Isot. Vol 47, No 11/12*, pp. 1241-1250, 1996.

[25] F. Trompier et al., "Overview of physical and biophysical techniques for accident dosimetry," *Radiation Protection Dosimetry*, no. 144, pp. 571-574, 2011.

[26] M. Marralea, A. Longoa, M.C. D'Ocad, A. Bartolottad, and M. Braia, "Watch glasses exposed to 6 MV photons and 10 MeV electrons analysed by means of ESR technique: A preliminary study," *Radiation Measurements*, no. 46, pp. 822-826, 2011.

[27] T. Ulseth and M. Abrahamsen. (22.06.15 besøkt 23.04.16) iPhone, Store norske leksikon. [Online]. https://snl.no/iPhone

[28] Corning Inc. (05.04.16) Corning - Gorilla® Glass. [Online]. http://www.corninggorillaglass.com/

[29] Corning Incorporated, "Corning Gorilla® Glass Product Information," New York, USA, 2015. [Online]. http://www.corninggorillaglass.com/en/glass-types/gorilla-glass-3-with-ndr

[30] K. Villeneuve, "EPR Spectral Analysis of X-Irradiated Alanine Dosimeter Subject to Thermal Annealing," Fysisk Institutt, Universitetet i Oslo, Oslo, Norge , 2015.

[31] IRSN. (25.04.16) IRSN. [Online]. http://www.irsn.fr/EN/Pages/home.aspx

[32] BRUKER. (30.03.16) Bruker Corporation, USA. [Online]. https://www.bruker.com/

[33] Professor E. O. Hole; Fysisk institutt, Universitetet i Oslo, Personlig kommunikasjon, 2016.

[34] MULTIBIODOSE, "Final publishable summary report of the project," 2013. [Online]. http://www.multibiodose.eu/News/MBD%20final%20publishable%20summary.pdf

[35] D. C. Lay, *Linear Algebra and Its Applications*. Boston, USA: Pearson Education, Inc, 2012.

[36] K. Hofstad. (14.11.14 besøkt 20.03.16) SI-systemet, Store Norske Leksikon. [Online]. https://snl.no/SI-systemet

[37] B. M. Wise et al., "Chemometrics Tutorial for PLS_Toolbox and Solo," Wenatchee, USA, 2006.

[38] R. Tauler and A. d. Juan:, *The Unscrambler Methods*.: CAMO Software AS, Juni 2006.

[39] A. Candolfia, R. De Maesschalcka, D. Jouan-Rimbauda, P.A. Haileyb, and L. Massarta D., "The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra," *Journal of Pharmaceutical and Biomedical Analysis*, no. 1, pp. 115-132, Oktober 1999.

[40] A. Kohler, C. Kirschner, A. Oust, and H. Martens, "Extended Multiplicative Signal Correction as a Tool for Separation and Characterization of Physical and Chemical Information in Fourier Transform Infrared Microscopy Images of Cryo-sections of Beef Loin," *Applied Spectoscopy, Volume 59, number 6*, pp. 707-716, 2005.

[41] R. A. Johanson and D. W.Wichern, *Applied Multivariate Statistical Analysis, 6. edition*. London, England: Pearson Education, Inc., 2007.

[42] J. Trygg, "Chemometrics made easy," *Homepage of Chemometrics*, 2003.

[43] H. Martens and M. Martens, *Multivariate Analysis of Quality: An Introduction*. West Sussex, England: John Wiley & Sons Ltd., 2001.

[44] A. Kohler et al., "Interpreting Several Types of Measurements in Bioscience," in *Modern Concepts in Biomedical Vibrational Spectroscopy*.: John Wiley & Sons, Inc, 2008, pp. 333-356.

[45] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, USA: Springer, 2013.

[46] R. Wehrens, *Chemometrics with R - Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Berlin, Tyskland: Springer - Verlag, 2011.

[47] A. Juan, J. Jaumot, and R. Tauler, "Multivariat Curve Resolution (MCR). Solving the mixture analysis problem," *The Royal Society of Chemistry*, May 2014.

[48] A. de Juana, S. Naveaa, J. Diewokb, and R. Taulera, "Local rank exploratory analysis of evolving rank-deficient systems," *Chemometrics and Intelligent Laboratory Systems*, pp. 11-21, September 2003.

[49] M. Maeder, "Evolving factor analysis for the resolution of overlapping chromatographic peaks," *Analytical Chemistry*, vol. 59, pp. 527-530, Februar 1987.

[50] Eigenvector Research. (15.03.16) PLS_Toolbox. [Online]. http://www.eigenvector.com/software/pls_toolbox.htm

[51] W. Windig, N. B. Gallagher, J. M. Shaver, and B. M. Wise, "A new approach for interactive self-modeling mixture analysis," *Chemometrics and Intelligent Laboratory Systems*, pp. 85–96, May 2005.

[52] W. Windig and J. Guilment, "Interactive self-modeling mixture analysis," *Anal. Chem.*, vol. 63, pp. 1425-1432, Mar. 1991.

[53] P. Persoone, R. De Gryse, and P. De Volder, "A new powerful transformation for maximum likelihood common factor analysis (MLCFA)," *Journal of Electron Spectroscopy and Related Phenomena*, pp. 225-232, 1995.

[54] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neurale Networks*, pp. 411-430, Mar. 2000.

[55] A. Hyvärinen, "Independent component analysis: recent advances," *Philosophical Transactions of the Royal Society*, p. 371, Desember 2012.

[56] J. Shlens, *A Tutorial on Independent Component Analysis*.: Google Research, 2014.

[57] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*. Stanford, USA: Springer, 2008.

[58] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B. Vol. 58*, pp. 267-288, 1996.

[59] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of Chemometrics*, pp. 166-173, 2003.

[60] G. G. Løvås, *Statistikk for univeristeter og høyskoler*, 26th ed. Oslo, Norge: Universitetsforlaget AS, 2004.

[61] W. Mendenhall and T. Sincich, *A Second course in Statistics, Regression Analysis, Seventh edition*. Essex, England: Pearson Education Limited, 2014.

[62] M. Clark, "A Comparison of Correlation Measures ," *Center for Social Research, University of Notre Dame*, May 2013.

[63] J. Jaumot, B. Igne, C. A. Anderson, J. K. Drennen, and A. de Juan, "Blending process modeling and control by multivariate curve resolution," *Talanta*, pp. 492-504, Sep. 2013.

[64] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*. Heidelberg, Tyskland: Springer-Verlag, 2008.

[65] MathWorks. (27.04.2016) [Online]. http://se.mathworks.com/index.html?s_tid=gn_logo

[66] MathWorks. (27.04.16) Statistics and Machine Learning Toolbox. [Online]. http://se.mathworks.com/products/statistics/

[67] H. Gävert, J. Hurri, J. Särelä, and A. Hyvärinen. (15.03.13 besøkt 18.02.16) Independent Component Analysis (ICA) and Blind Source Separation (BSS). [Online]. http://research.ics.aalto.fi/ica/fastica/

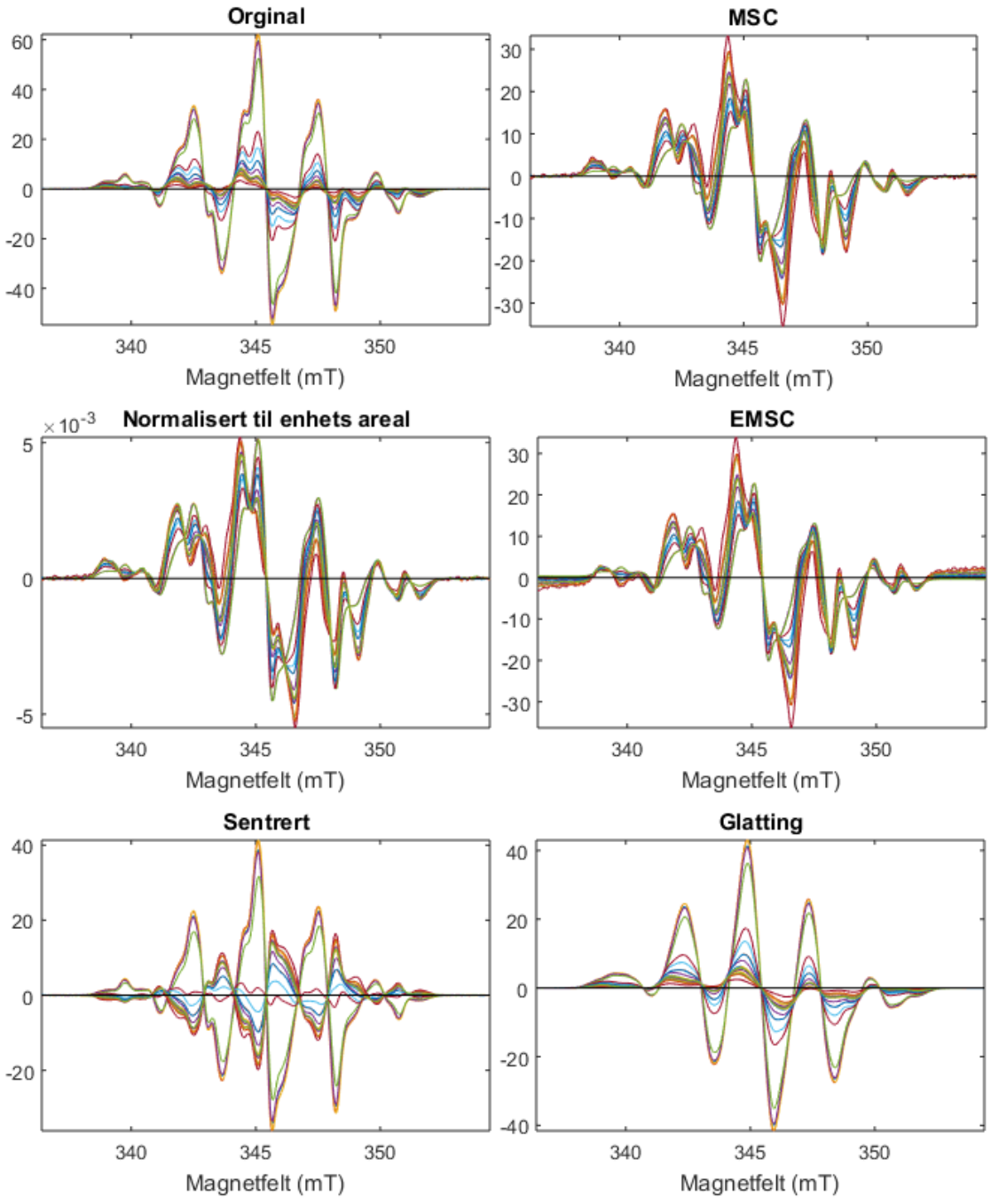
[68] H. Motulsky. (13.07.11 besøkt 29.04.16) stackexchange. [Online]. http://stats.stackexchange.com/questions/12900/when-is-r-squared-negative

[69] R. Wicklin. (15.02.12 besøkt 30.04.16) What is Mahalanobis distance? [Online]. http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance.html

[70] A. Grafen and R. Hails, *Modern Statistics for the Life Sicences*, 1st ed. New York, USA: Oxford University Press Inc., 2002.

8 Vedlegg

8.1 Vedlegg 1: Eksempler på preprosesseringer



Figur 8-1, eksempler hvordan preprosesseringsteknikkene påvirker EPR-spektrene, for teknikkene MSC, EMSC, Normalisering til enhetsareal og glatting med 50 målepunkters glattefilet. Det kommer fram at det er små forskjeller mellom normalisering, MSC og EMSC.

8.2 Vedlegg 2: Andeler av R1, R2 og R3* fra residualanalysene

Tabell 8-1, beste tilpasning ved bruk av minste kvadraters tilpasning til målespektrene ved bruk av tre komponenter R1, R2 og R3. R3* estimatet er basert på de uthevede rader (kontrollene).*

Temperatur [ºC]	Tid [min]	R1	R2	R3*	LOF	Korrelasjon	R^2
197	**1**	**0,610**	**0,373**	**0,017**	**0,020**	**1,000**	**1,000**
197	30	0,238	0,750	0,012	0,203	0,979	0,959
197	60	0,236	0,752	0,012	0,202	0,979	0,959
197	90	0,124	0,866	0,011	0,324	0,946	0,895
197	120	0,106	0,884	0,010	0,348	0,938	0,879
197	150	0,121	0,869	0,010	0,328	0,945	0,892
205	**1**	**0,598**	**0,387**	**0,016**	**0,018**	**1,000**	**1,000**
205	16	0,357	0,630	0,013	0,124	0,992	0,985
205	32	0,156	0,834	0,011	0,285	0,958	0,919
205	48	0,109	0,881	0,010	0,341	0,940	0,883
205	80	0,047	0,943	0,010	0,458	0,889	0,790
213	**1**	**0,605**	**0,378**	**0,017**	**0,044**	**0,999**	**0,998**
213	10	0,284	0,704	0,012	0,171	0,985	0,971
213	20	0,108	0,882	0,010	0,346	0,938	0,880
213	30	0,056	0,934	0,010	0,434	0,901	0,812
213	40	0,048	0,942	0,010	0,448	0,894	0,800
213	50	0,017	0,971	0,011	0,554	0,832	0,693
Kontroll 1	**0**	**0,595**	**0,390**	**0,015**	**0,016**	**1,000**	**1,000**
Kontroll 2	**0**	**0,592**	**0,393**	**0,015**	**0,027**	**1,000**	**0,999**

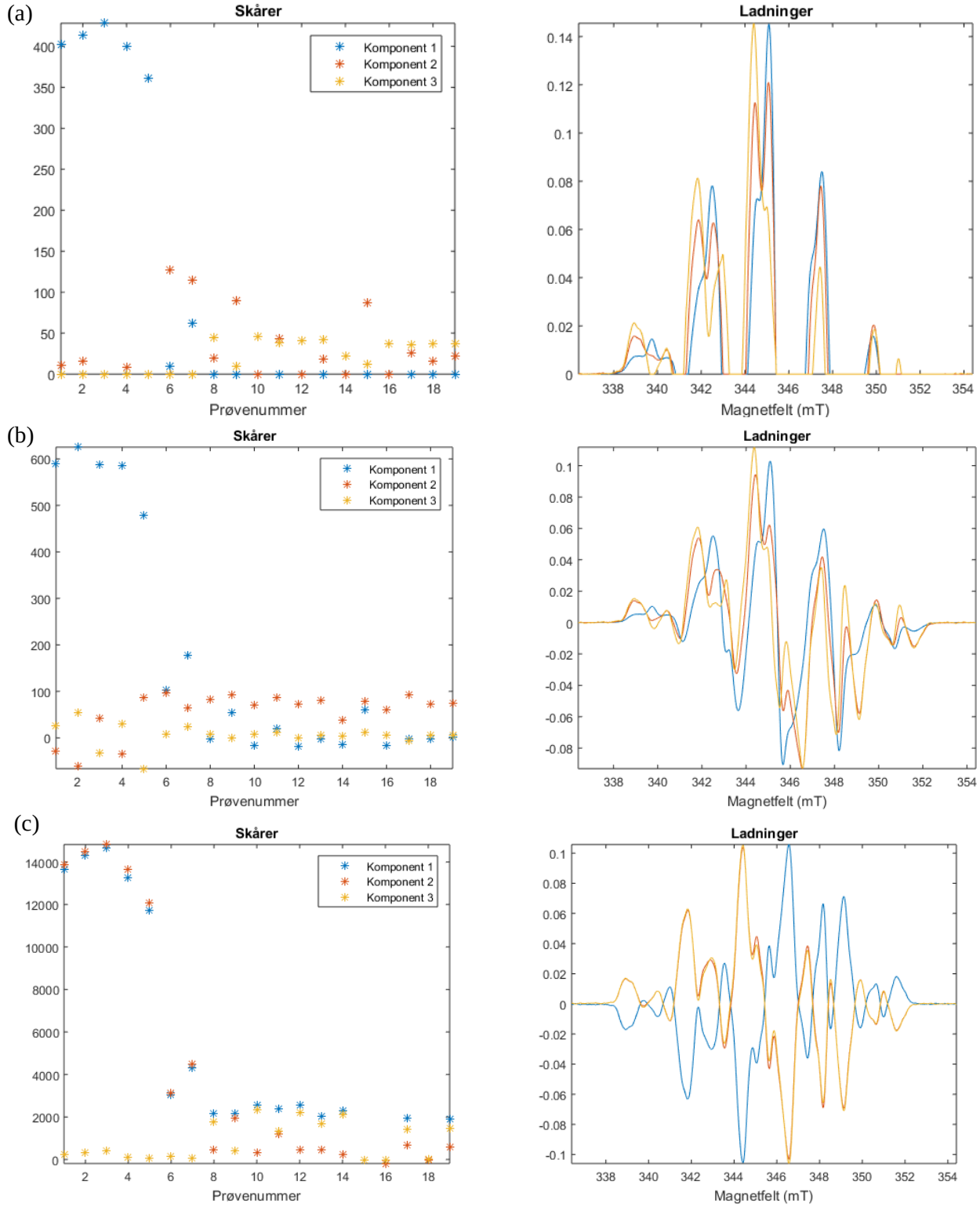
Tabell 8-2, beste tilpasning ved bruk av minste kvadraters tilpasning til målespektrene ved bruk av tre komponenter R1, R2 og R3. R3* estimatet er basert på EPR-spektrene som antas å inneholde mest R3* (uthevede rader).*

Temperatur [ºC]	Tid [min]	R1	R2	R3*	LOF	Korrelasjon	R^2
197	1	0,608	0,376	0,017	0,139	0,990	0,981
197	30	0,229	0,724	0,047	0,107	0,994	0,989
197	60	0,227	0,726	0,046	0,102	0,995	0,990
197	90	0,117	0,822	0,060	0,083	0,997	0,993
197	120	0,100	0,837	0,063	0,052	0,999	0,997
197	**150**	**0,115**	**0,824**	**0,061**	**0,059**	**0,998**	**0,997**
205	1	0,596	0,389	0,015	0,129	0,992	0,983
205	16	0,348	0,616	0,036	0,123	0,992	0,985
205	32	0,148	0,796	0,056	0,076	0,997	0,994
205	48	0,103	0,834	0,063	0,054	0,999	0,997
205	**80**	**0,043**	**0,877**	**0,079**	**0,033**	**1,000**	**0,999**
213	1	0,603	0,380	0,016	0,144	0,990	0,979
213	10	0,274	0,683	0,043	0,111	0,994	0,988
213	20	0,102	0,835	0,063	0,054	0,999	0,997
213	**30**	**0,052**	**0,872**	**0,076**	**0,030**	**1,000**	**0,999**
213	**40**	**0,045**	**0,878**	**0,077**	**0,048**	**0,999**	**0,998**
213	**50**	**0,016**	**0,888**	**0,097**	**0,083**	**0,997**	**0,993**
Kontroll 1	0	0,593	0,392	0,015	0,128	0,992	0,984
Kontroll 2	0	0,590	0,395	0,015	0,127	0,992	0,984

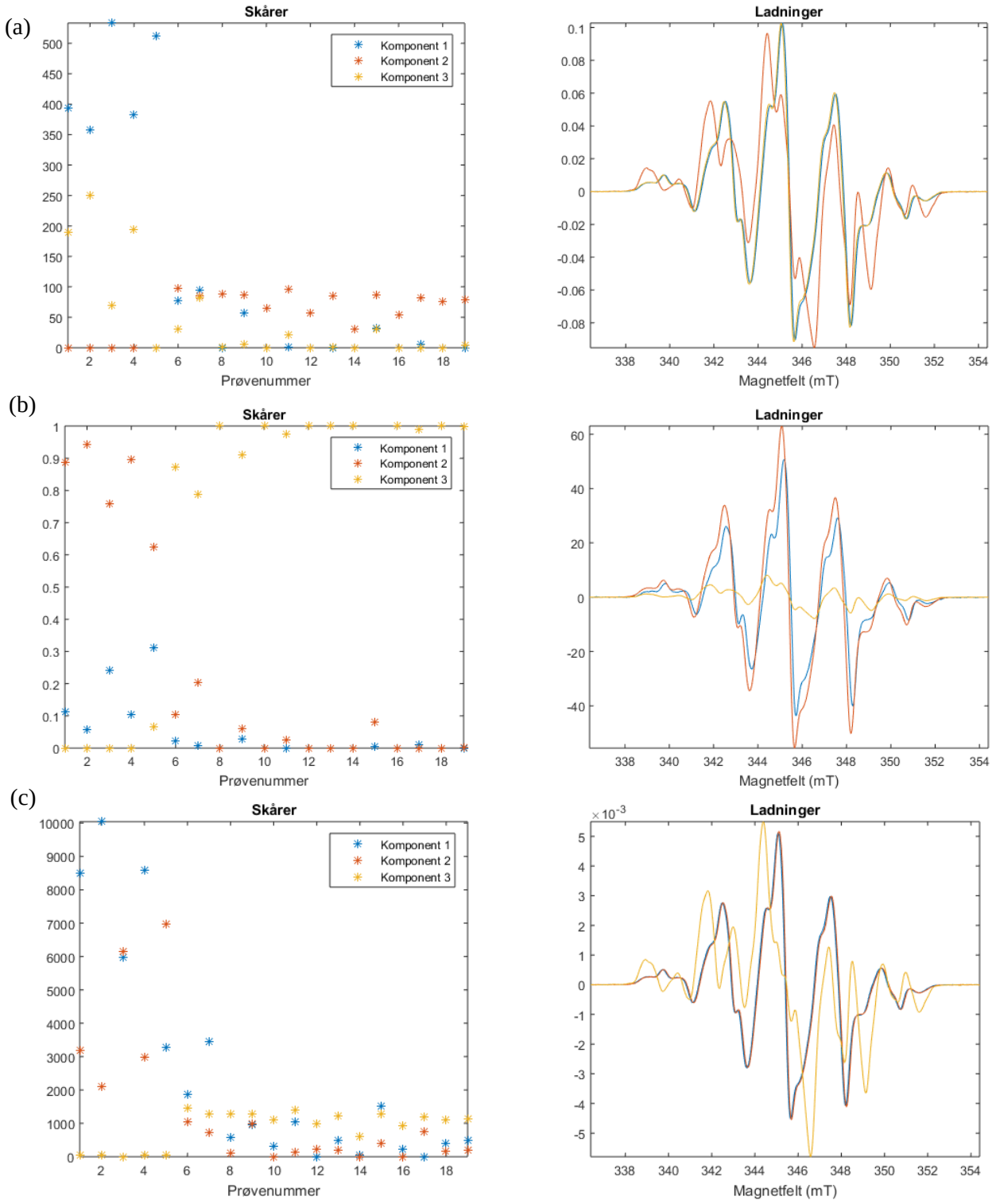
Tabell 8-3, beste tilpasning ved bruk av minste kvadraters tilpasning til målespektrene ved bruk av tre komponenter R1, R2 og R3. R3* estimatet er basert alle målingene.*

Temperatur [ºC]	Tid [min]	R1	R2	R3*	LOF	Korrelasjon	R^2
197	1	0,602	0,370	0,027	0,087	0,996	0,992
197	30	0,233	0,732	0,035	0,072	0,997	0,995
197	60	0,231	0,735	0,034	0,080	0,997	0,994
197	90	0,121	0,839	0,040	0,165	0,986	0,973
197	120	0,103	0,857	0,040	0,190	0,982	0,964
197	150	0,118	0,843	0,039	0,175	0,984	0,969
205	1	0,591	0,384	0,025	0,084	0,996	0,993
205	16	0,350	0,618	0,032	0,055	0,999	0,997
205	32	0,152	0,811	0,037	0,142	0,990	0,980
205	48	0,106	0,854	0,040	0,183	0,983	0,966
205	80	0,045	0,907	0,047	0,283	0,959	0,920
213	1	0,598	0,375	0,028	0,092	0,996	0,991
213	10	0,278	0,689	0,034	0,058	0,998	0,997
213	20	0,105	0,855	0,040	0,190	0,982	0,964
213	30	0,054	0,900	0,046	0,263	0,965	0,931
213	40	0,047	0,906	0,047	0,272	0,962	0,926
213	50	0,017	0,927	0,056	0,368	0,930	0,865
Kontroll 1	0	0,588	0,387	0,024	0,083	0,997	0,993
Kontroll 2	0	0,586	0,391	0,024	0,085	0,996	0,993

8.3 Vedlegg 3: MCR betingelser

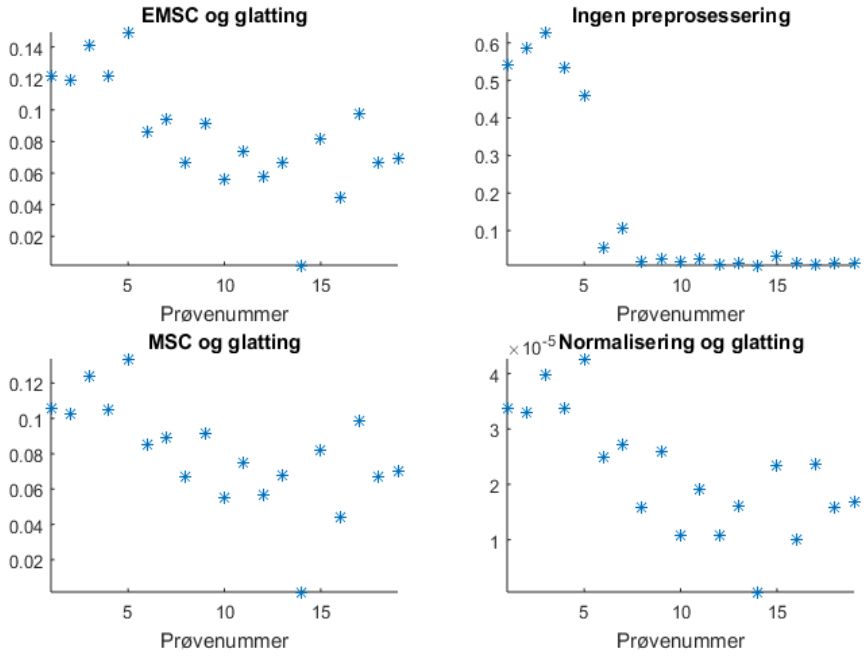


Figur 8-2, MCR analyse med betingelsene: (a) Standard betingelser, (b) negative skårer og ladninger og (c) negative ladninger og betingelsen maksimal kontrast til ladningene.

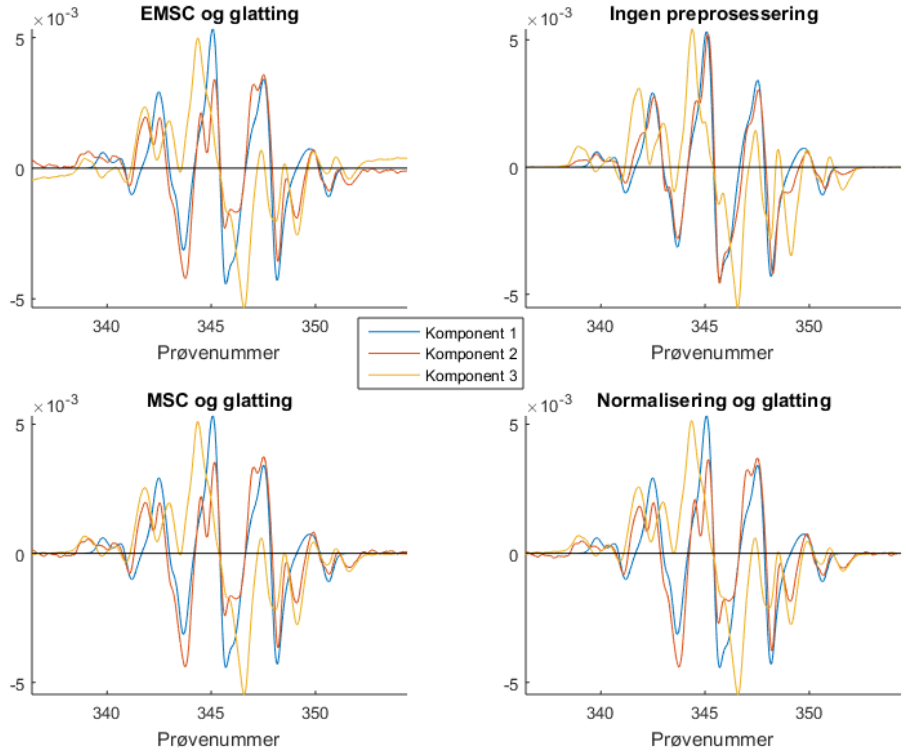


Figur 8-3, MCR analyse med betingelsene: (a) negative ladninger og betingelsen maksimal kontrast på skårene, (b) negative ladninger og betingelsen skårene skal summere seg opp til 1 og (c) negative ladninger som er normalisert til enhetsareal, med initialiseringsmetode distslct i Matlab®.

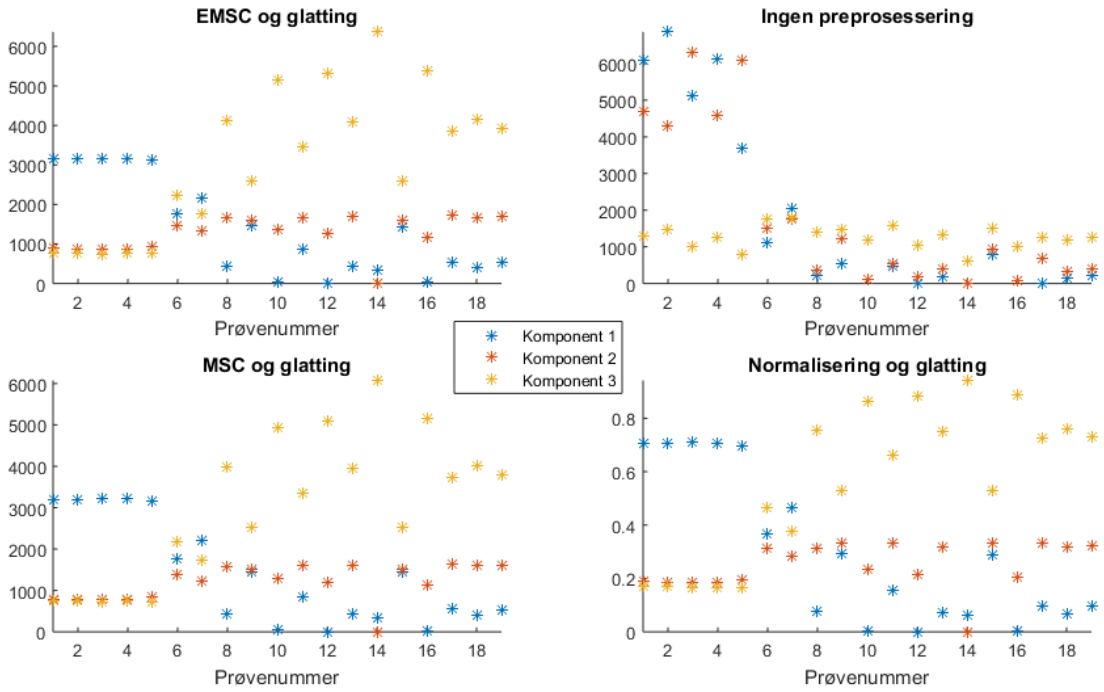
8.4 Vedlegg 4: Ladninger, skårer og residualer fra MCR



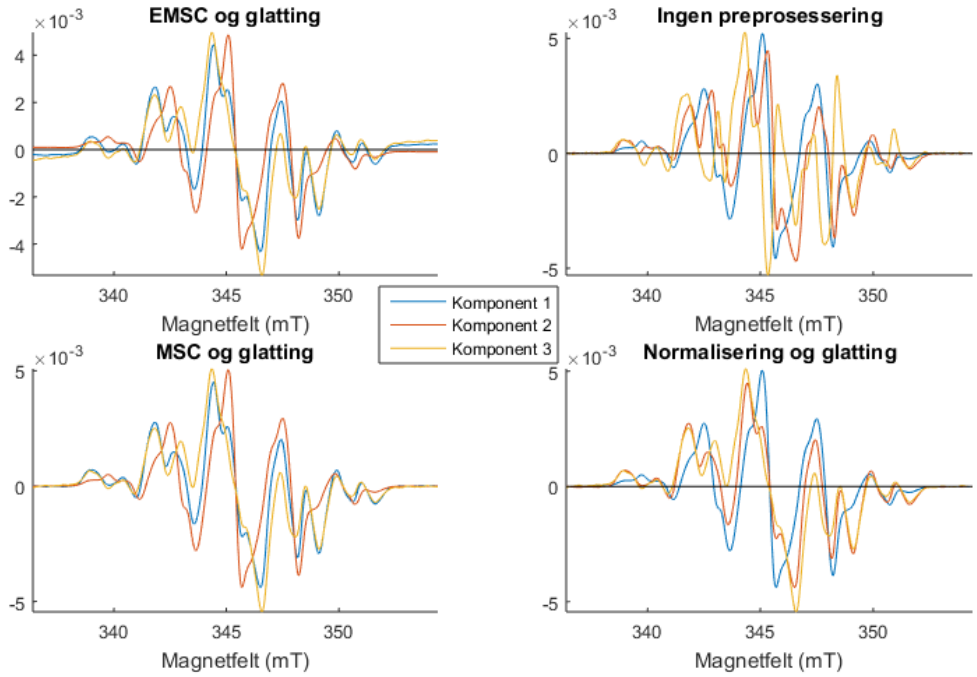
Figur 8-4, Q residualen fra MCR analyse av alanin datasettet med R1 og R2 antatt kjent. Q residualene er størst for ingen preprosessering. Q er størst for kontrollene og for 1 minuttsoppvarmede prøvene. De tre preprosesserte modellene har mer gjevt fordelt residual. Aksene er en relativ akse med den største skårverdien som 1.



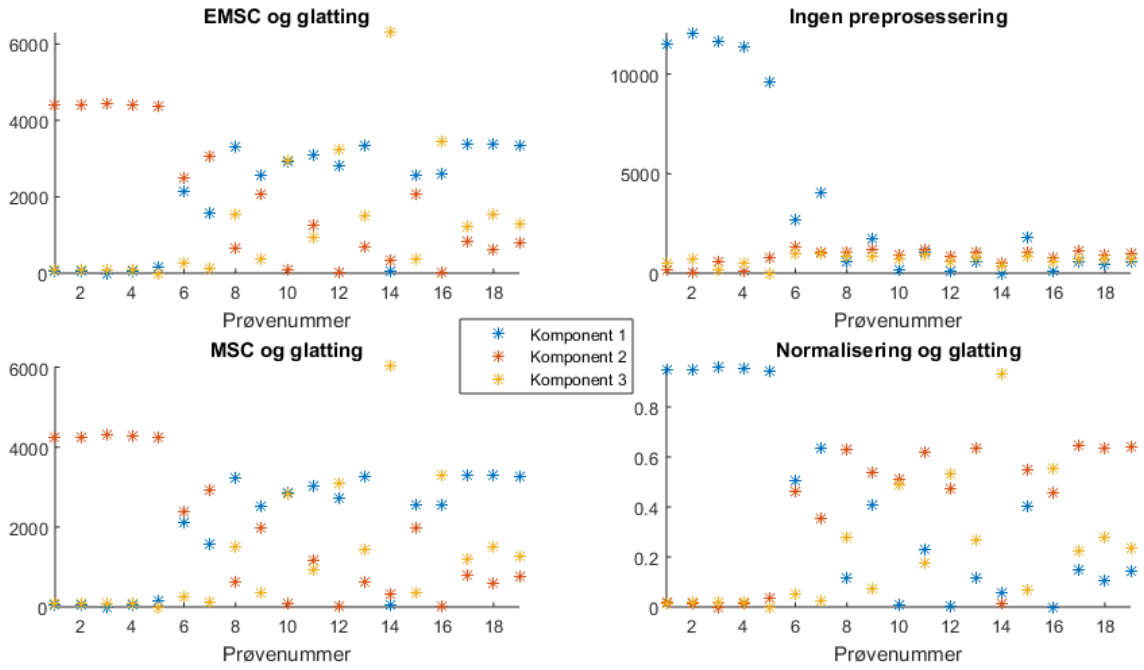
Figur 8-5, ladningene fra MCR analyser av alanin datasettet, med R1 antatt kjent, titlene angir preprosesseringsteknikk som ble brukt.



Figur 8-6, skårene fra MCR av alanin datasettet, med R1 antatt kjent. Titlene angir preprosesseringsteknikk som ble brukt.



Figur 8-7, ladningene fra MCR analyser med ingen antatt kjente spektre. Titlene angir preprosesseringsteknikk som ble brukt.



Figur 8-8, skårene fra MCR analyser med ingen antatt kjente spektre. Titlene angir preprosesseringsteknikk som ble brukt.

8.5 Vedlegg 5: Andeler av R1*, R2* og R3* i målespektrene

Tabell 8-4, viser andelene av PC1-3 i målespektrene for PCA på sentrert alanin datasett, med tilhørende korrelasjon og LOF-verdier. Andelen av PC1 går nedover med økende tid, PC2 går oppover mens PC3 holder seg ganske konstant. Korrelasjonen er høy (>0,91) for alle prøvene og LOF er ganske lav.

Temperatur [ºC]	Tid [min]	PC1	PC2	PC3	Korrelasjon	LOF
197	1	0,898	0,072	0,030	1,000	0,001
197	30	0,606	0,328	0,066	0,995	0,011
197	60	0,614	0,338	0,048	0,995	0,010
197	90	0,478	0,426	0,096	0,981	0,038
197	120	0,466	0,462	0,073	0,978	0,043
197	150	0,484	0,445	0,071	0,982	0,036
205	1	0,914	0,085	0,001	1,000	0,001
205	16	0,728	0,246	0,026	0,998	0,003
205	32	0,528	0,415	0,057	0,988	0,024
205	48	0,470	0,458	0,072	0,979	0,041
205	80	0,387	0,537	0,075	0,955	0,088
213	1	0,878	0,071	0,051	1,000	0,001
213	10	0,653	0,296	0,051	0,997	0,007
213	20	0,470	0,462	0,068	0,979	0,041
213	30	0,402	0,525	0,073	0,961	0,076
213	40	0,385	0,520	0,095	0,955	0,087
213	50	0,343	0,583	0,074	0,915	0,163
0	0	0,911	0,086	0,003	1,000	0,001
0	0	0,903	0,088	0,009	0,999	0,001

Tabell 8-5, andel av R1, R2*, R3*, R^2 og LOF verdier for tilpasning med minste kvadraters metode til måledataene ved bruk av estimerte, glattede spektre av alanin datasettet, fra MLCFA analysen.*

Temperatur [Cº]	Tid [min]	Andel R1*	Andel R2*	Andel R3*	R^2	LOF
197	1	0,71	0,26	0,02	0,99	0,09
197	30	0,42	0,46	0,12	0,92	0,28
197	60	0,42	0,46	0,12	0,91	0,30
197	90	0,30	0,56	0,13	0,90	0,32
197	120	0,28	0,58	0,14	0,90	0,31
197	150	0,30	0,57	0,14	0,90	0,32
205	1	0,70	0,27	0,03	0,99	0,10
205	16	0,53	0,39	0,09	0,95	0,22
205	32	0,34	0,53	0,13	0,90	0,31
205	48	0,29	0,58	0,13	0,91	0,30
205	80	0,21	0,69	0,10	0,95	0,22
213	1	0,71	0,27	0,02	0,99	0,10
213	10	0,47	0,43	0,10	0,94	0,25
213	20	0,29	0,59	0,13	0,91	0,29
213	30	0,23	0,67	0,10	0,95	0,22
213	40	0,22	0,69	0,09	0,95	0,22
213	50	0,17	0,82	0,01	0,98	0,13
0	0	0,70	0,27	0,03	0,99	0,10
0	0	0,70	0,27	0,03	0,99	0,10

Tabell 8-6, tilpassing av komponent 1-3 fra MCR med R1 og R2 antatt kjent og MSC som preprosessering. Andelene for hver komponent i de målte spektrene ble funnet ved minste kvadraters metode. Det viser seg at andelen R1 avtar med økende tid og ved økende temperatur, mens R2 og R3 øker. Korrelasjonen mellom de målespektrene og de tilpassede spektrene er også tatt med og det kommer fram at de estimerte spektrene er ganske god tilpasning til de målte dataene, LOF verdien er også lave for alle modellene <2,2 %.*

Temperatur [ºC]	Tid [min]	R1	R2	R3*	Korrelasjon	LOF
197	1	0,820	0,112	0,067	0,990	0,020
197	30	0,454	0,224	0,322	0,991	0,019
197	60	0,452	0,230	0,319	0,991	0,017
197	90	0,259	0,240	0,501	0,992	0,016
197	120	0,221	0,234	0,545	0,994	0,012
197	150	0,253	0,238	0,509	0,994	0,012
205	1	0,813	0,124	0,063	0,991	0,017
205	16	0,604	0,187	0,209	0,990	0,019
205	32	0,319	0,239	0,441	0,993	0,014
205	48	0,228	0,236	0,536	0,994	0,012
205	80	0,081	0,159	0,760	0,997	0,007
213	1	0,818	0,117	0,066	0,989	0,022
213	10	0,516	0,208	0,276	0,991	0,018
213	20	0,225	0,233	0,542	0,994	0,012
213	30	0,105	0,184	0,711	0,996	0,008
213	40	0,086	0,176	0,738	0,996	0,008
213	50	0,002	0,019	0,979	0,999	0,002
0	0	0,811	0,125	0,064	0,991	0,017
0	0	0,809	0,128	0,063	0,992	0,017

Tabell 8-7, tilpassing av komponent 1-3 fra MCR med R1 antatt kjent og preprosessering MSC. Andelene for hver komponent i de målespektrene ble funnet ved minste kvadraters metode. Her er komponent 1 et estimat for R1, komponent 2 et estimat for R3 og komponent 3 et estimat for R2. Andelen av komponent 1 går avtar økende tid, komponent 2 øker med økende tid, og komponent 3 øker kraftig og dominerer ved de lengste tidene.

Temperatur [ºC]	Tid [min]	Komponent 1	Komponent 2	Komponent 3	Korrelasjon	LOF
197	1	0,679	0,175	0,146	0,992	0,015
197	30	0,252	0,295	0,453	0,997	0,005
197	60	0,251	0,294	0,454	0,998	0,005
197	90	0,074	0,301	0,625	0,998	0,004
197	120	0,050	0,289	0,662	0,999	0,001
197	150	0,074	0,294	0,632	0,999	0,002
205	1	0,675	0,175	0,150	0,993	0,013
205	16	0,417	0,255	0,327	0,996	0,009
205	32	0,131	0,298	0,571	0,999	0,003
205	48	0,055	0,291	0,655	0,999	0,001
205	80	0,016	0,197	0,787	0,999	0,004
213	1	0,666	0,186	0,147	0,992	0,016
213	10	0,321	0,277	0,402	0,997	0,006
213	20	0,055	0,286	0,658	0,999	0,002
213	30	0,014	0,225	0,761	0,999	0,003
213	40	0,020	0,209	0,771	0,998	0,005
213	50	0,029	0,027	0,944	0,999	0,002
0	0	0,672	0,176	0,151	0,993	0,013
0	0	0,674	0,174	0,152	0,993	0,013

Tabell 8-8, tilpassing av komponent 1-3 for MCR med ingen kjente spektre og MSC som preprosessering. Andelene for hver komponent i de målespektrene ble funnet ved minste kvadraters metode. Andelen av komponent 2 minker med økende tid, mens andelen av komponent 1 og 3 øker med økende tid.

Temperatur [ºC]	Tid [min]	Komponent 1	Komponent 2	Komponent 3	Korrelasjon	LOF
197	1	0,021	0,977	0,002	0,999	0,002
197	30	0,581	0,391	0,028	0,998	0,003
197	60	0,585	0,388	0,027	0,998	0,003
197	90	0,706	0,123	0,171	0,998	0,004
197	120	0,694	0,082	0,224	0,999	0,001
197	150	0,697	0,118	0,185	0,999	0,002
205	1	0,031	0,965	0,004	0,999	0,002
205	16	0,378	0,611	0,010	0,997	0,005
205	32	0,668	0,207	0,126	0,999	0,002
205	48	0,692	0,092	0,216	0,999	0,001
205	80	0,493	0,027	0,480	0,998	0,005
213	1	0,055	0,921	0,024	0,997	0,006
213	10	0,500	0,488	0,012	0,998	0,004
213	20	0,685	0,090	0,225	0,999	0,002
213	30	0,562	0,019	0,419	0,999	0,003
213	40	0,522	0,030	0,448	0,997	0,007
213	50	0,085	0,020	0,895	0,999	0,002
0	0	0,036	0,960	0,005	0,999	0,001
0	0	0,032	0,967	0,001	0,999	0,002

Tabell 8-9, andeler av komponent 1-3 fra SMA analysen med kun positive verdier til alanin datasettet. Komponent 1 kan være et estimat på R3 siden andelen av komponent 1 går fra nesten 0, hos kontrollene, og opp til ca. 40% for de prøvene som har vært utsatt for høy temperatur/lang tid.

Temperatur [ºC]	Tid [min]	Komponent 1	Komponent 2	Komponent 3	Korrelasjon	LOF
197	1	0,001	0,939	0,061	1,000	0,000
197	30	0,393	0,473	0,133	0,997	0,000
197	60	0,395	0,473	0,132	0,998	0,000
197	90	0,435	0,422	0,143	0,999	0,000
197	120	0,443	0,413	0,144	0,999	0,000
197	150	0,438	0,419	0,143	1,000	0,000
205	1	0,027	0,914	0,059	1,000	0,000
205	16	0,312	0,574	0,113	0,997	0,000
205	32	0,419	0,444	0,137	0,996	0,000
205	48	0,438	0,420	0,142	1,000	0,000
205	80	0,457	0,395	0,148	0,989	0,000
213	1	0,081	0,838	0,081	1,000	0,000
213	10	0,356	0,520	0,124	0,994	0,000
213	20	0,437	0,422	0,141	0,999	0,000
213	30	0,452	0,402	0,146	0,997	0,000
213	40	0,455	0,397	0,147	0,993	0,000
213	50	0,466	0,383	0,151	0,946	0,000
0	0	0,021	0,922	0,058	1,000	0,000
0	0	0,000	0,950	0,050	1,000	0,000

Tabell 8-10, andeler av komponent 1-3 fra SMA med normerte verdier.

Temperatur [°C]	Tid [min]	Komponent 1	Komponent 2	Komponent 3	Korrelasjon	LOF
197	1	0,263	0,224	0,513	0,998	0,004
197	30	0,096	0,069	0,834	0,998	0,004
197	60	0,095	0,068	0,837	0,999	0,001
197	90	0,170	0,060	0,770	0,946	0,141
197	120	0,230	0,075	0,695	0,890	0,261
197	150	0,186	0,061	0,753	0,939	0,162
205	1	0,262	0,221	0,517	0,999	0,001
205	16	0,194	0,142	0,665	0,999	0,001
205	32	0,094	0,019	0,887	0,991	0,038
205	48	0,220	0,071	0,709	0,904	0,236
205	80	0,474	0,105	0,421	-0,143	1,244
213	1	0,266	0,216	0,518	0,997	0,006
213	10	0,143	0,104	0,753	0,999	0,002
213	20	0,227	0,071	0,701	0,897	0,252
213	30	0,419	0,104	0,477	0,233	0,971
213	40	0,449	0,109	0,443	0,023	1,127
213	50	0,808	0,063	0,129	-0,984	3,112
0	0	0,262	0,221	0,517	0,999	0,001
0	0	0,261	0,222	0,518	1,000	0,001

Tabell 8-11, andeler av IC 1-3 fra ICA med tre komponenter. Her er IC 1 et estimat på R3, IC 2 estimat på R2 og IC 3 er et estimat på R1. Før denne tabellen ble laget ble IC 3 skalert med -1.

Temperatur [°C]	Tid [min]	IC 1	IC 2	IC 3	Korrelasjon	LOF
197	1	0,016	0,232	0,752	1,000	0,000
197	30	0,016	0,462	0,522	1,000	0,001
197	60	0,035	0,454	0,511	1,000	0,001
197	90	0,005	0,590	0,405	0,999	0,002
197	120	0,024	0,604	0,372	1,000	0,001
197	150	0,025	0,583	0,392	1,000	0,001
205	1	0,042	0,229	0,729	1,000	0,000
205	16	0,043	0,359	0,598	0,999	0,001
205	32	0,035	0,536	0,429	1,000	0,000
205	48	0,025	0,598	0,377	1,000	0,001
205	80	0,027	0,695	0,278	0,996	0,009
213	1	0,002	0,237	0,760	1,000	0,000
213	10	0,027	0,419	0,554	1,000	0,001
213	20	0,029	0,598	0,373	1,000	0,001
213	30	0,029	0,676	0,295	0,997	0,006
213	40	0,003	0,706	0,291	0,996	0,009
213	50	0,027	0,756	0,218	0,980	0,040
0	0	0,040	0,231	0,729	1,000	0,000
0	0	0,050	0,229	0,721	1,000	0,000

8.6 Vedlegg 6: Glatting av R3*

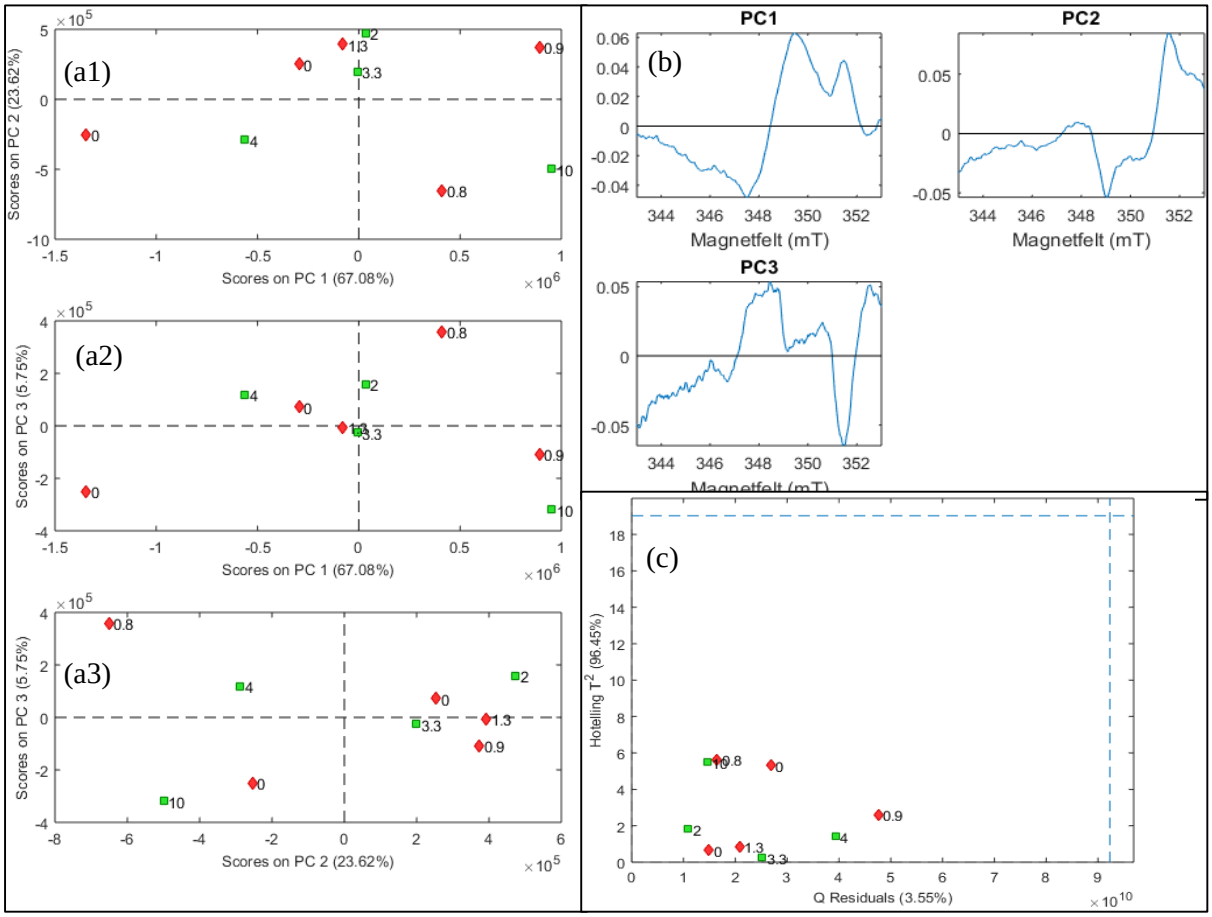
Tabell 8-12, korrelasjon, LOF og R^2 mellom det teoretiske R3-spektret (figur 2-4) og de estimerte R3-spektrene i figur 4-32, funnet fra (a) minste kvadraters metode, (b) MCR, (c) MLCFA og (d) SMA.*

	Preprosessering	**Spekter**	**Størrelsen på glattefilteret**	**Korrelasjon**	**LOF**	**R^2**
(a) Minste kvadraters metode						
		Alle	Ingen	0,252	1,13	0
			10	0,263	1,12	0
			30	0,362	1,02	0
			50	0,501	0,902	0,185
			100	0,681	0,737	0,456
		Kontroll	Ingen	0,689	0,751	0,435
			10	0,708	0,730	0,467
			30	0,832	0,560	0,686
			50	0,926	0,378	0,857
			100	0,959	0,297	0,912
		Utvalgte	Ingen	0,914	0,406	0,835
			10	0,918	0,396	0,843
			30	0,943	0,335	0,887
			50	0,960	0,282	0,920
			100	0,972	0,240	0,942
(b) MCR						
	MSC	R1 og R2 kjent	Ingen	0,774	0,644	0,585
			10	0,777	0,639	0,591
			30	0,803	0,601	0,639
			50	0,840	0,542	0,706
			100	0,896	0,462	0,786
	Ingen	R1 kjent	Ingen	0,648	0,799	0,360
			10	0,653	0,794	0,369
			30	0,694	0,746	0,443
			50	0,753	0,669	0,552
			100	0,834	0,558	0,688
	MSC	Ingen kjent	Ingen	0,787	0,627	0,607
			10	0,790	0,622	0,613
			30	0,817	0,581	0,662
			50	0,853	0,522	0,728
			100	0,903	0,45	0,797
(c) MLCFA						
		Tre komponenter	Ingen	-0,583	1,61	0
			10	-0,595	1,62	0
			30	-0,681	1,67	0
			50	-0,796	1,72	0
			100	-0,913	1,77	0
		To komponenter	Ingen	0,856	0,521	0,729
			10	0,861	0,513	0,737
			30	0,894	0,450	0,797
			50	0,933	0,360	0,870
			100	0,963	0,295	0,913
(d) SMA						
		Normerte EPR-spektre	Ingen	-0,784	1,79	0
			10	-0,787	1,79	0
			30	-0,814	1,80	0
			50	-0,851	1,791	0
			100	-0,906	1,741	0
		Forskjøvet EPR-spektre	Ingen	0,121	1,333	0
			10	0,124	1,331	0
			30	0,148	1,293	0
			50	0,185	1,220	0
			100	0,353	1,054	0

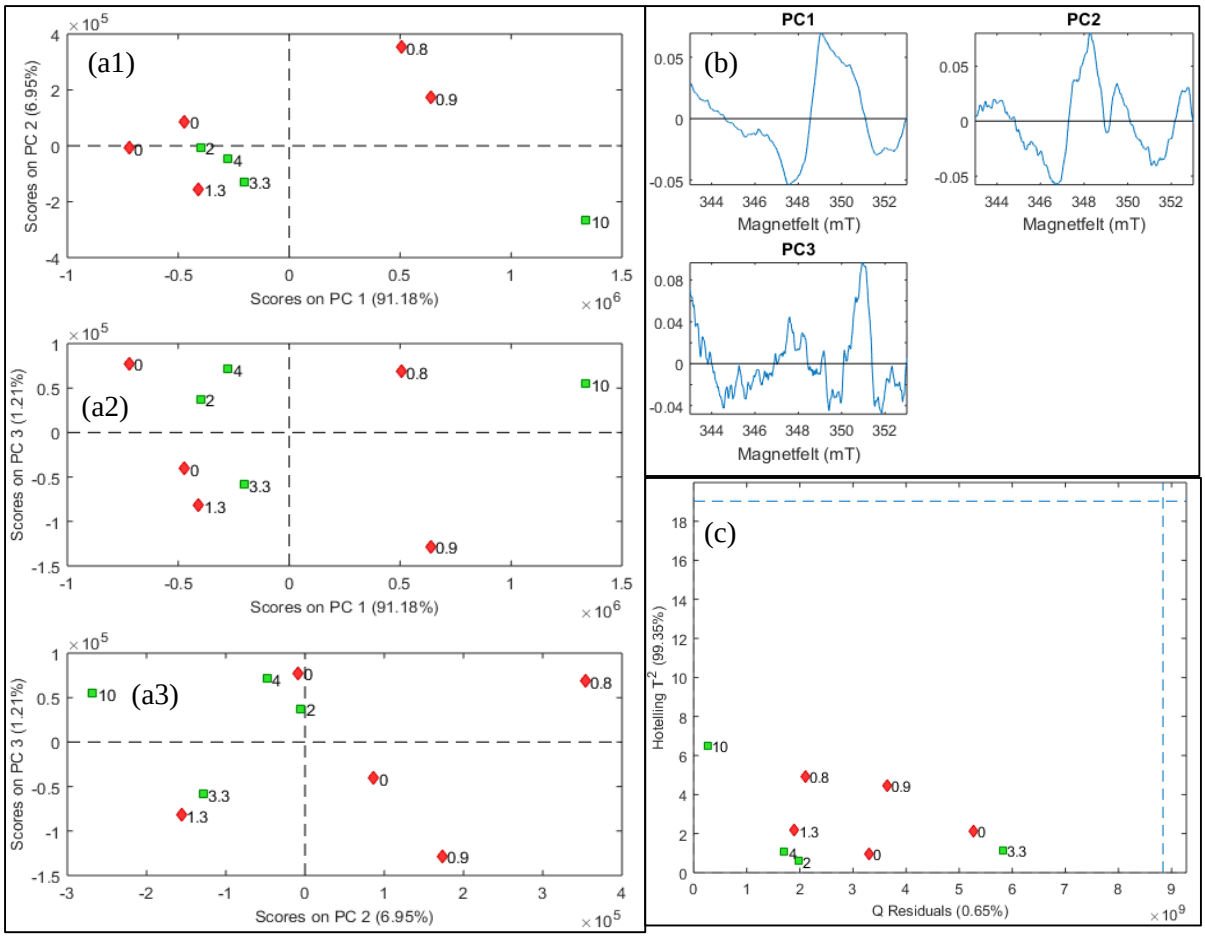
Tabell 8-13, korrelasjon, LOF og R^2 mellom det teoretiske R3-spekteret (figur 2-4) og de estimerte R3-spektrene i figur 4-32, funnet fra (a) PCA og (b) ICA.*

	Spekter	**Størrelsen på glattefilteret**	**Korrelasjon**	**LOF**	**R^2**
(a) PCA					
	Kun sentrering	Ingen	0,096	1,34	0
		10	0,099	1,337	0
		30	0,121	1,292	0
		50	0,149	1,241	0
		100	0,268	1,118	0
	MSC og sentrering	Ingen	0,047	1,274	0
		10	0,048	1,275	0
		30	0,057	1,275	0
		50	0,07	1,273	0
		100	0,114	1,205	0
(b) ICA					
		Ingen	-0,066	1,458	0
		10	-0,068	1,457	0
		30	-0,085	1,43	0
		50	-0,106	1,404	0
		100	-0,2	1,44	0

8.7 Vedlegg 7: PCA på Gorilla® Glass datasettene

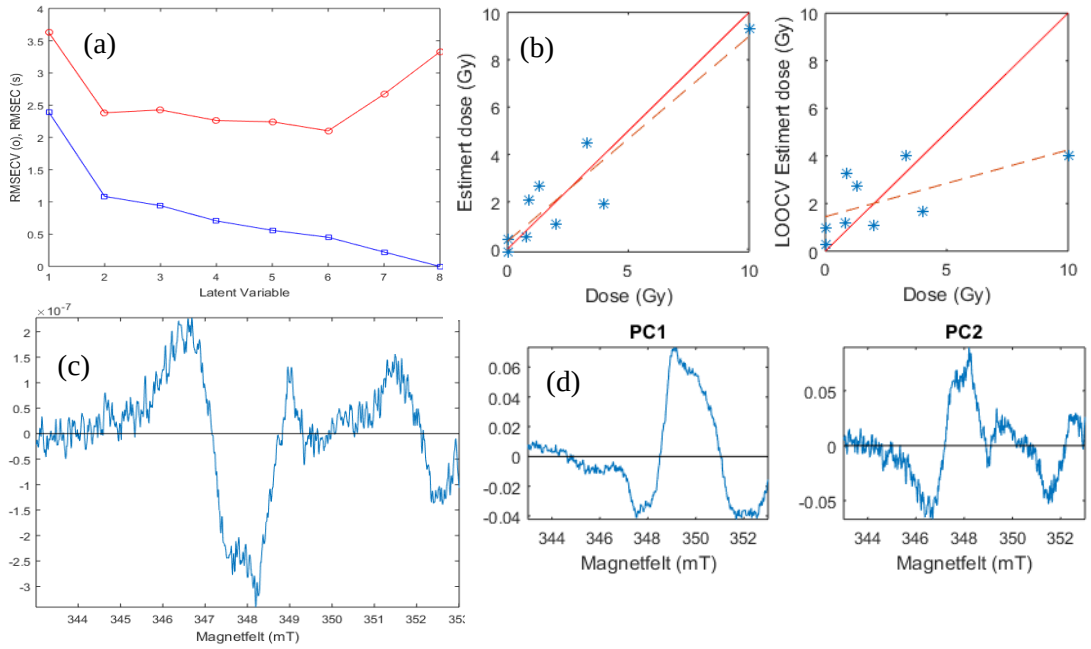


Figur 8-9, resultater fra PCA analyse med glatting og sentrering som preprosessering på rektangulær kavitet datasettet, (a) PC1 mot PC2 (a1), PC1 mot PC3 (a2) og PC2 mot PC3 (a3), (b) ladningene og (c) Q-residualer mot T^2 med 95% konfidensintervall grenser (stiplede linjer). Fargene står for hvor stor dose prøvene har mottatt: rød betyr lavdose (<2 Gy), grønn betyr høydose (≥2 Gy), tallen er eksakt dose prøvene har absorbert. PC1 står for 67,08, PC2 23,62 % og PC3 5,75 % av forklart varians.

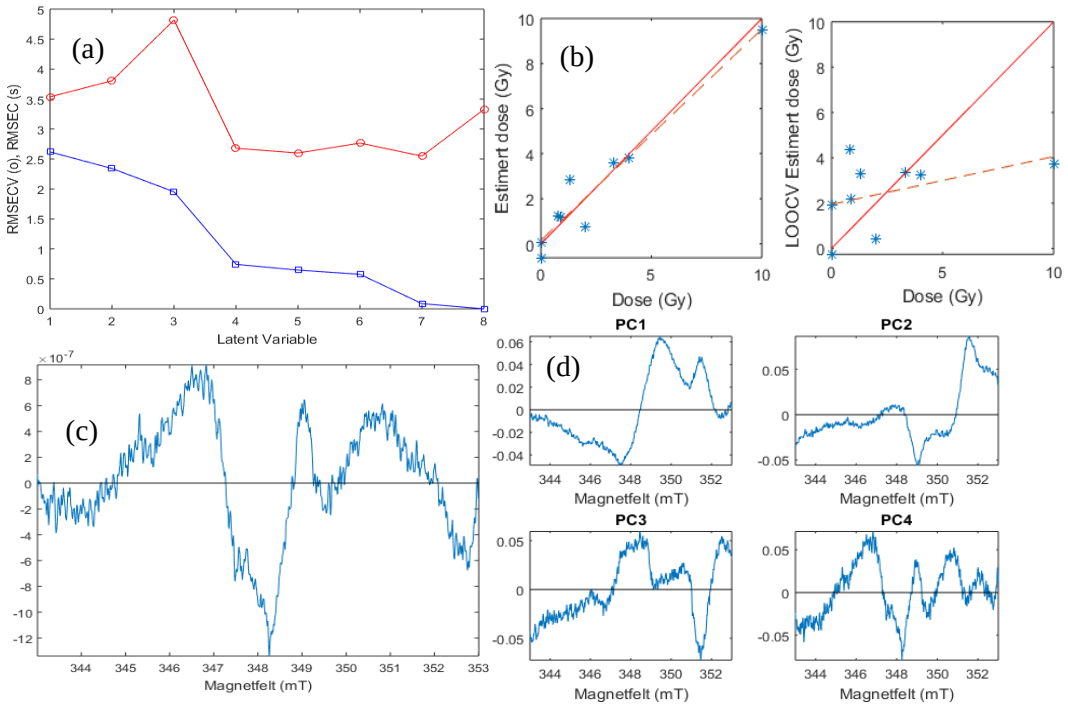


Figur 8-10, resultater fra PCA analysen med EMSC, glatting og sentrering som preprosessering på rektangulær kavitet datasettet, (a) PC1 mot PC2 (a1), PC1 mot PC3 (a2) og PC2 mot PC3 (a3), (b) ladningene og (c) Q-residualer mot T^2 med 95% konfidensintervall grenser (stiplede linjer). Fargene angir stor dose prøvene har mottatt: rød betyr lavdose (<2 Gy), grønn betyr høydose (≥2 Gy), tallene er eksakt dose prøvene har absorbert. PC1 står for 91,18 %, PC2 6,95 % og PC3 1,21 % av forklart varians.

8.8 Vedlegg 8: PCR analyser

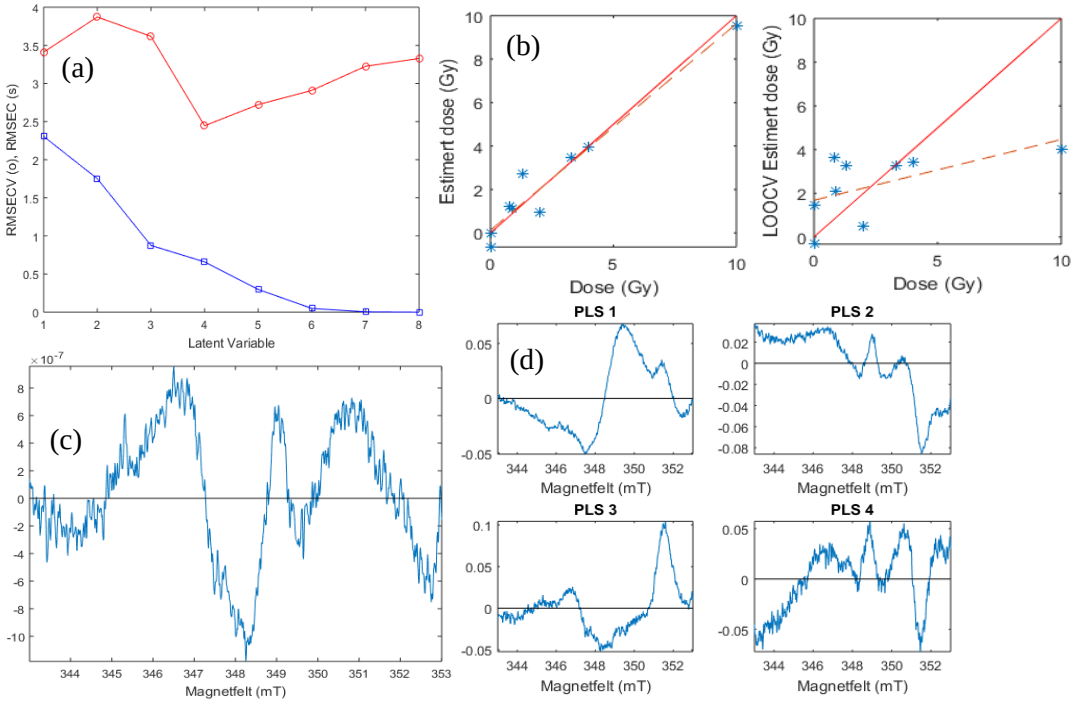


Figur 8-11, PCR analyse av SuperX kavitet datasettet med preprosessering sentrering og MSC, (a) RMSEC (blå) og RMSECV (rød), (b) estimerte doser uten kryssvalidering (venstre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir RMSEC 1,01 og RMSECV 1,02 (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for 85,1 % (PC1) og 8,2 % (PC2) av forklart varians hos variablene og 34,6 % (PC1) og 52,0 % (PC2) av forklart varians for responsene.

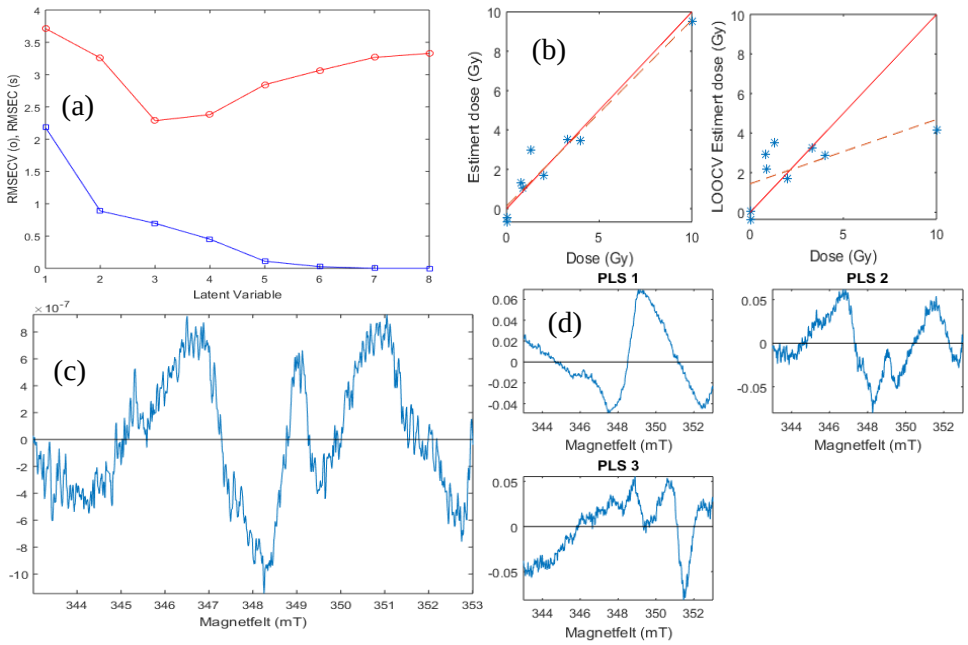


Figur 8-12, PCR analyse av sentrert rektangulær kavitet datasettet, (a) RMSEC (blå) og RMSECV (rød) (b) estimerte doser uten kryssvalidering (venstre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose. Beste tilpasning gir (stiplet linje) RMSEC 0,72 og RMSECV 1,31, og (d) ladningene. Ladningene står for 66,7 % (PC1), 23,6 % (PC2), 5,8 % (PC3) og 3,1 % (PC4) av forklart varians for variablene og står for 21,7 % (PC1), 15,7 % (PC2), 19,1% (PC3) og 37,3 % (PC4) av forklart varians hos responsen.

8.9 Vedlegg 9: PLS analyser

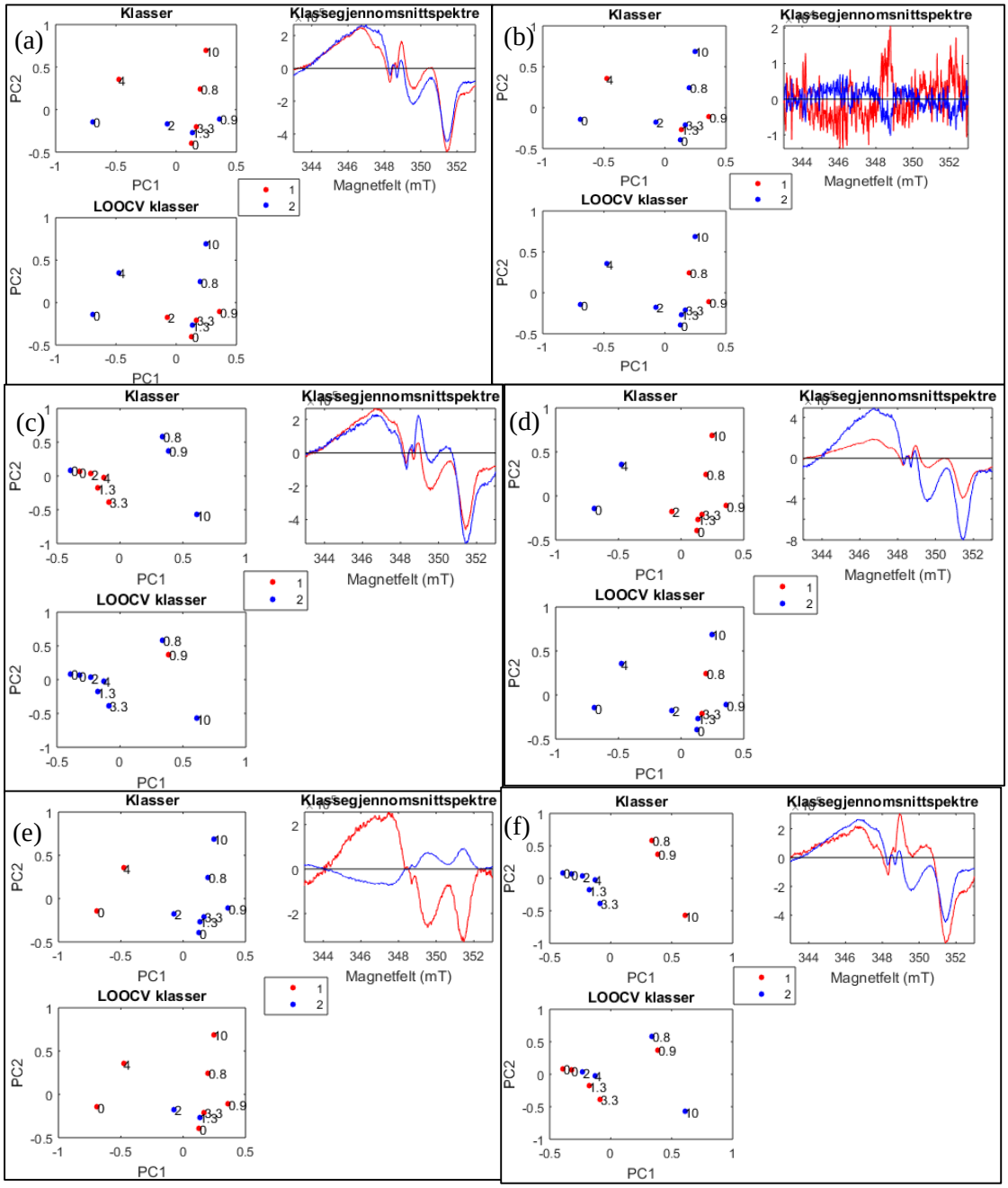


Figur 8-13, PLS regresjon av sentrert rektangulær kavitet datasett, (a) RMSEC (blå) og RMSECV (rød) (b) estimerte doser uten kryssvalidering (vestre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir 0,64 og RMSECV 1,18, (c) regresjonskoeffisientene og (d) ladningene. PLS 1 står for 62,6 %, PLS 2 for 24,0 % og PLS 3: 8,2 % av totalt forklart varians for variablene og for 39,3 % (PLS 1), 25,7 % (PLS 2) og for 3,8 % (PLS 3) av responsene.

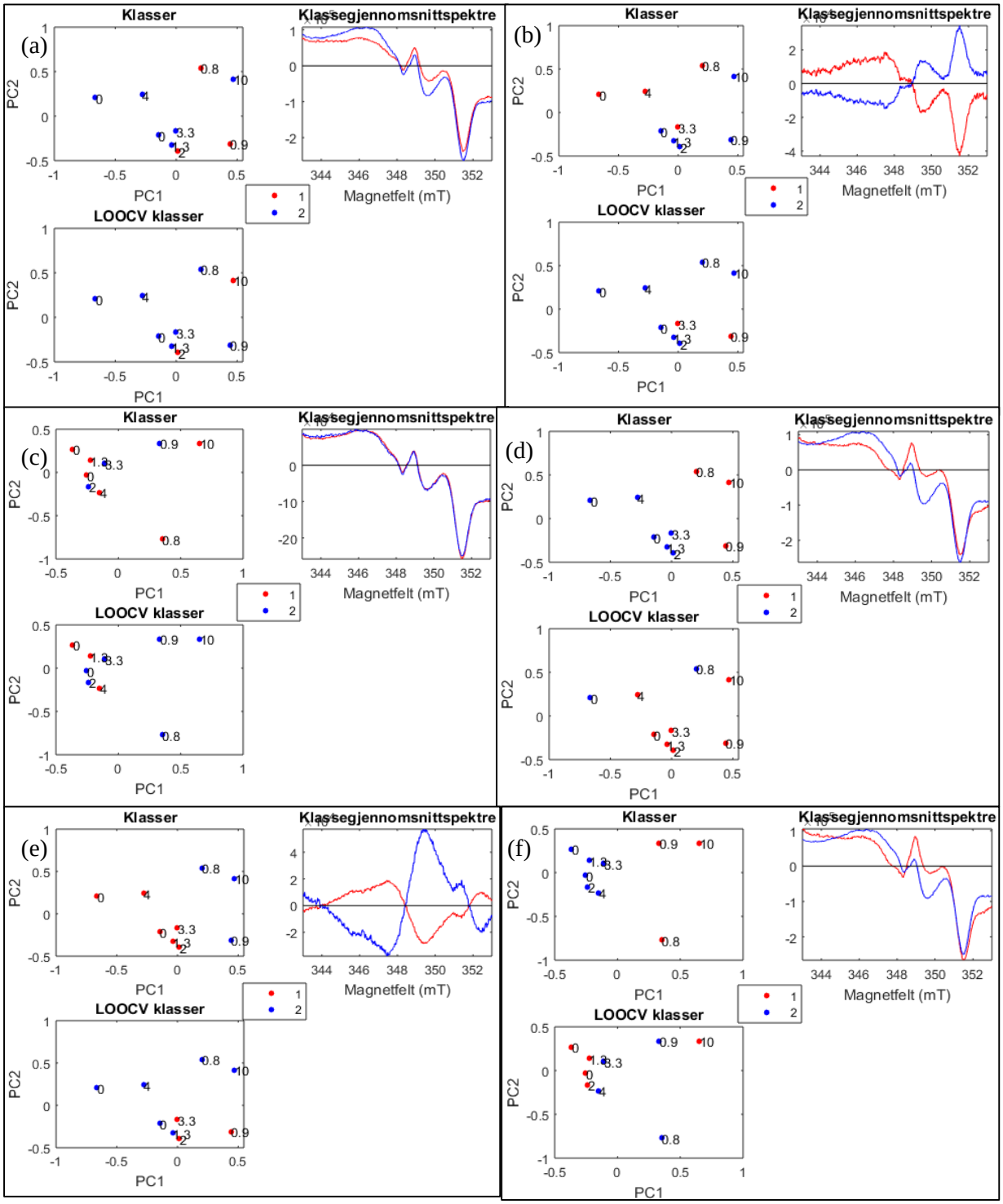


Figur 8-14, PLS regresjon av rektangulær kavitet datasett med preprosessering MSC og sentrering, (a) RMSEC (blå) og RMSECV (rød), (b) estimerte doser uten kryssvalidering (vestre) og med kryssvalidering (høyre). Den røde linjen angir riktig estimering av dose, tilpasning til beste linje (stiplet) gir RMSEC 0,68 og RMSECV 1,09, (c) regresjonskoeffisientene og (d) ladningene. Ladningene står for totalt forklart varians for variablene PLS 1: 81,0 %, PLS 2: 7,7 % og PLS 3: 8,7 % og for responsene PLS 1: 45,6 % PLS 2: 45,4 % og PLS 3: 3,5 %.

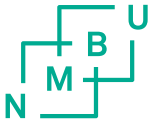
8.10 Vedlegg 10: K-gjennomsnittsanalyse



Figur 8-15, resultatene fra k-gjennomsnittsalgoritme av SuperX dataene, med (a,d) ingen, (b,e) gjennomsnittsentrering og (c,f) MSC som preprosessering. Mahalanobis (a,b,c) og euklidsk (d,e,f) avstand som avstandsmål. Til venstre i hvert panel vises klassetilhørigheten uten og med LOOCV, visst mot PC1 og PC2 verdiene. PCA blir brukt for å visualisering. Til høyre i hvert panel er klassegjennomsnittspektrene til klassifiseringsmodellene.



Figur 8-16, resultatene fra k-gjennomsnittsalgoritme av rektangulær kavitet dataene, med (a,d) ingen, (b,e) sentrering og (c,f) MSC som preprosessering. Med Mahalanobis (a,b,c) og euklidsk avstand (d,e,f) som avstandsmål. Til venstre i hvert panel vises klassetilhørigheten uten og med LOOCV, visst mot PC1 og PC2 verdiene. PCA blir brukt for å visualisering. Til høyre i hvert panel er klassegjennomsnittspektrene til klassifiseringsmodellene.



Norges miljø- og biovitenskapelig universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway