Norges miljø- og
biovitenskapelige
universitet

Masteroppgave 2018 60 stp
Fakultet for kjemi, bioteknologi og matvitenskap

Prediksjon av toksisitet i skjell basert på fettsyresammensetning

Prediction of toxicity in shellfish based on their fatty acid composition

Elise Lunde Gjelsvik
Master i kjemi

Forord

Denne masteroppgaven ble gjennomført ved Fakultet for kjemi bioteknologi og matvitenskap (KBM) ved Norges miljø og biovitenskapelige universitet (NMBU) i perioden august 2017 til mai 2018.

Jeg har alltid følt at statistikk er et viktig verktøy når det kommer til kjemiske analyser. Statistiske metoder kan bidra i stor grad til å gi verdifull innsikt i resultater. Jeg er veldig gald for at jeg fikk muligheten til å gjøre dette i oppgaven min.

Takk til Trygve Almøy for støtte og hjelp i prosessen med å lære multivariat statistikk. Takk til Kristian Hovde Liland for hjelp med programmering og CPLS. Takk til hovedveileder Dag Ekeberg for kjemisk veiledning. Tusen takk for at dere ville være med på den litt uvanlige ideen min om en kjemometrioppgave.

Tusen takk til mamma og pappa for støtte gjennom studietiden. Takk til gode venner for avkobling og morsomme stunder.

Tusen takk til jentene i Palasset, hadde aldri kommet gjennom dette uten dere!

Elise Lunde Gjelsvik

Ås, 11/05-2018

Sammendrag

Lipofile marine biotoksiner kan akkumuleres i skjell og være en helserisiko for mennesker hvis de konsumeres. Okadasyregruppen er toksingruppen som er den mest vanlige årsaken til diaréfremkallende skjellforgiftning (DSP) i Norge. Toksininnholdet kontrolleres i kommersielt omsatte skjell. Metoden for måling av toksininnhold har flere ulemper (Fux et al. 2008, Aanrud 2016). Et alternativ til analysen er å bruke multivariabel statistikk til å finne ut om skjellene er giftige.

Analyse av fettsyreprofiler i en skjellprøve kan være enklere og sikrere enn analyse av toksininnholdet. Statistiske metoder som PCR, PLS, CPLS og variabelseleksjon ble undersøkt for å finne en prediksjonsmodell for toksisitet basert på fettsyresammensetning. Metodene ble validert ved leave-one-out kryssvalidering og testsettvalidering. PCA ble kjørt for å se på grupperinger eller sammenhenger i variablene. Sammenligning av scoreplot og ladningsplot antydet at Blåskjell inneholder mer trans-fettsyrer og Stillehavsøsters inneholder med mettede fettsyrer.

ANOVA ble gjennomført for å vurdere forklaringsvariablene. Sted kom ut som signifikant med $\alpha = 0.05$. Det ble under denne analysen oppdaget en uteligger B-1443 Rundhaugen. Denne prøven representerte en ekstrem algeoppblomstring som kan føre til overestimering dersom den ble inkludert i modellen. Siden sted virket å ha effekt ble residualer hentet ut fra ANOVA og brukt som respons for noen regresjonsmetoder.

De tre beste metodene ble valgt ut til å være PCR med logtransformering, PLS med logtransformering og CPLS med logtransformering. CVANOVA og Tukey *post hoc*-test viste at CPLS med 5 komponenter var den beste metoden.

Abstract

Lipophilic marine biotoxins can be accumulated in different shellfish and can be a health risk to humans if consumed. Okadaic acid is the toxin group which most commonly causes Diarrhetic Shellfish Poisoning (DSP) in Norway. Commercially distributed shellfish are controlled for toxins. The analysis for detection of toxins have some disadvantages (Fux et al. 2008, Aanrud 2016). One option instead of this analysis is the use of multivariate statistics to discover toxic shellfish.

Analysis of fatty acid profiles in shellfish can be easier and more accurate then measuring toxin content. Statistical methods such as PCR, PLS, CPLS and forward selection was explored to obtain a prediction model for the toxicity based on the fatty acid composition. The methods were validated using leave-one-out crossvalidation and testset validation. PCA was examined to explore groupings or realtions in the data. Comparisons between scoreplots and loadingplots indicated that Blue mussle contains more trans fatty acids and Pacific oyster contains more saturated fatty acids.

ANOVA was performed to evaluate the explanatory variables. Samplearea was determined to be significant ($\alpha = 0.05$). During this analysis an outlier, B-1443 Rundhaugen, was detected. This sample represents an extreme algae bloom which could lead to overestimation if included. Samplearea seemed to have an effect on toxicity and residuals from the ANOVA were assesed as a respons during the regression methods.

The three best methods were selected as PCR with logtransformation, PLS with logtransformation and CPLS with logtransformation. CVANOVA and Tukey *post hoc*-test suggested that CPLS containing 5 components was the best method.

Forkortelser

ANOVA	Variansanalyse
CPLS	Kanonisk Partial Least Squares Regresjon
CVANOVA	Variansanalyse av kryssvaliderte prediksjoner
DSP	Diarefremkallende skjellforgiftning
DTX-1/2/3	Dinophysistoksin-1/2/3
EI	Elektronioniseringskilde
ESI	Elektronspray ioniseringskilde
FAME	Fettsyremetylester
GC	Gasskromatografi
HPLC	Høy-ytelses væskekromatografi
LOOCV	Leave-One-Out Kryssvalidering
LS	Minste kvadrater
MBA	Musebioassay
MS	Massespektrometri
MSEP	Mean Square Error of Prediction
OA	Okadasyre
PCA	Prinsipalkomponentanalyse
PCR	Prinsipal komponent regresjon
PE	Prediksjonsfeil
PLS	Partielle Minste Kvadraters Regresjon
RMSEP	Root Mean Square Error of Prediction
RMSECV	Root Mean Square Error of Cross Validation

Innhold

1 Innledning

1.1 Marine biotoksiner

Lipofile marine biotoksiner kan akkumuleres i forskjellige skjelldyr og være en helserisiko for mennesker dersom de konsumeres. For å beskytte folkehelsen er det opprettet overvåkningsprogrammer for marine biotoksiner for å detektere disse i skjelldyrvev. I Norge har Mattilsynet opprettet et tilsynsprogram som inkluderer 18 helårlige prøveuttakssteder og 36 prøveuttakssteder i sommerhalvåret (Mattilsynet 2018). Blåskjell dyrket i Norge og andre typer skjell solgt i butikk og omsatt kommersielt kontrolleres for toksininnhold. Tilsynsprogrammet startet opp i 2006 og er en videreføring av overvåkning av blåskjell startet av Statens næringsmiddeltilsyn i 1988. Det tas både vannprøver som sendes til marinebiologer for analyse av algeinnhold og prøver av skjell som sendes til Algelaboratoriet på Institutt for mattrygghet og infeksjonsbiologi (MatInf) ved Norges miljø- og biovitenskapelige universitet (NMBU). Giftene prøvene analyseres for er diaréfremkallende skjellforgiftning (DSP), paralytisk skjellforgiftning (PSP) og nevrotoksisk skjellforgiftning (NSP). Av disse er den mest vanlige i Norge DSP som refererer til gastrointestinal ubalanse som et resultat av inntak av skjell infisert med dinoflagellattoksiner (Lee et al. 1989).

Det er tidligere vist at antallet giftige skjell øker når mengden toksiske plankton i et havområde øker (Yasumoto et al. 1978). Skjellene blir giftige når de filterer vannet for alger som inneholder toksiner. Toksinene er delt inn i åtte grupper etter kjemisk struktur; azaspiracid (AZA), brevetoksin, domoisyre (DA), okadasyre (OA), pectenotoksin (PTX),

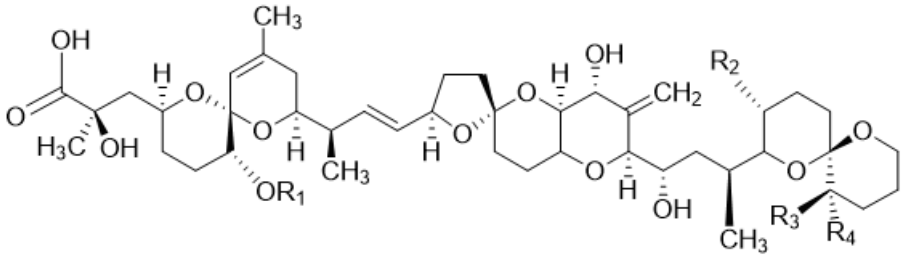
saxitoksin (STX) sykliske imin og yessotoksin (YTX). Av disse er OA-gruppen den vanligste årsaken til DSP-forgiftning (Steidinger 1993). I Europa er de mest vanlige toksingruppene OA, DTX, PTX og YTX og gruppene detekteres noen ganger som blandinger (Alarcan et al. 2018).

Foreløpig er analyser av skjellvev den beste metoden for kontrollering av toksinkonsentrasjonene og matsikkerheten, men metoden har flere ulemper (Fux et al. 2008). Noen av disse er problemer ved prøvetaking, forskjell i opptak av toksiner mellom arter og interferens grunnet matriks. Et alternativ til denne metoden kan være å istedet for analysen bruke multivariabel statistikk til å forutse ut om skjellene er gifitge.

1.2 Okadasyre-gruppen

Okadasyre (OA) er et lipofilt fykotoksin som akkumuleres i fordøyningsorganer hos muslinger (Svensson & Förlin 2003). I denne gruppen regnes også syrens strukturelle analoger dinophysistoksin-1 (DTX-1), dinophysistoksin-2 (DTX-2) og dinophysistoksin-3 (DTX-3). DTX-3 brukes som en samlebetegnelse på OA-toksiner som har en fettsyreester i R_1 hvor de tre andre har en OH-gruppe. Dannelsen av DTX-3 er studert i kamskjell hvor det ble vist at skjellene produserer DTX-3 selv og at det ikke finnes DTX-3 i dinoflagellatene som produserer OA-toksinene (Suzuki et al. 1999). Altså er DTX-3 et biotransformasjonsprodukt fra OA, DTX-1 og DTX-2.

Det er tidligere vist at OA-gruppen hindrer proteinfosfatase (Holmes et al. 1990). Dette kan hindre biologisk aktivitet og være grunnen til forgiftningssymptomene ved inntak. I Norge produseres toksiner fra OA-gruppen av algene *Dinophysis acuta* og *Dinophysis Norvegica* som er to dinoflagellater (Lee et al. 1989). OA-gruppen er illustrert i figur 1.1.



Figur 1.1: Okadasyre-gruppen

Tabell 1.1 viser hvordan de strukturelle analogene til OA-gruppen, DTX-1, DTX-2 og biotransformasjonsproduktet DTX-3, vil se ut.

Tabell 1.1: Oversikt over hvordan OA-gruppen vil substituere og hvordan de strukturelle analogene vil se ut.

Toksin	R_1	R_2	R_3	R_4
OA	H	CH_3	H	H
DTX-1	H	CH_3	CH_3	H
DTX-2	H	H	H	CH_3
DTX-3 fra OA	Fettsyre	CH_3	H	H
DTX-3 fra DTX-1	Fettsyre	CH_3	CH_3	H
DTX-3 fra DTX-2	Fettsyre	CH_3	H	CH_3

Oppblomstringen av alger er en prosess som ikke har en kjent syklus. At syklusen ikke er kjent hindrer helårlig industriell produksjon av skjell. De fleste industrier som dyrker blåskjell forbys å høste skjellene deler av året (Séchet et al. 1990). Konstant kontroll av toksisiteten i skjellene kan bidra til mer kjennskap om denne syklusen og da bedre planlegging av industriell produksjon.

1.3 Skjellarter

Blåskjell (*Mytilus edulis*) er en muslingart som vokser langs store deler av norskekysten. Blåskjell er i Norge mye brukt som mat og plukkes og selges året rundt. Kamskjell (*Pecten maximus*) er en musling i bløtdyr familien som skiller seg fra andre muslingarter på grunn av større lukkemuskler. Lukkemuskelen fra kamskjell selges som en delikatesse. O-skjell (*Modiolus modiolus*) ligner på blåskjell og er i likhet med dem spiselige, men er større og lever dypere nede i havet. Stillehavsøsters (*Crassostrea gigas*) er en type østers som i nyere tid har spredd seg utover kysten av Norge. Alle disse skjellartene kan akkumulere toksiner fra giftige alger og derfor kontrolleres de av Mattilsynet.

De fire forskjellige skjellartene kan påvirkes ulikt av samme algetoksin (Yasumoto et al. 1978). Det er tidligere gjennomført studier som sammenligner opptaket og toksisiteten av de forskjellige skjellartene. Studiene har vist at blåskjell akkumulerer mer DSP-toksiner enn østers (Lindegarth et al. 2009), men østers produserer i større grad DTX-3 enn blåskjell (Torgersen et al. 2008). Kamskjell har en mer effektiv mekanisme for å håndtere DSP-toksiner og dermed akkumuleres de i mindre grad i forhold til blåskjell (Bauder et al. 2001). Informasjon om hvordan O-skjell påvirkes av DSP-toksiner er manglende, men Aanrud 2016 viste at O-skjell inneholder noen andre fettsyrer enn de tre andre artene.

1.4 Toksiske grenseverdier

Grenseverdiene for toksiske skjell bestemmes i Norge av EU og er satt til å være 160 µg OA-ekvivalent/kg skjellmat (EU-RL-MB 2015). Det blir benyttet en sikkerhetsmargin på ti ganger konsentrasjoner som gir symptomer på skjellforgiftning. En OA-ekvivalent tar hensyn til at DTX-2 har toksisitet på 0,6 ganger toksisiteten til OA. DTX-3 derivert fra DTX-2 har dermed også tilsvarende lavere toksisitet.

1.5 Kjemisk analyse

Tidligere ble det brukt bioassay til deteksjon av marine biotoksiner. Bioassay er metoder som benytter levende organismer til å bestemme biologisk aktivitet. Lenge har musebioassay (MBA) vært den mest vanlige måten å måle toksiner i skjell (Lawrence et al. 2011). Denne metoden er veldig lite spesifikk fordi frie fettsyrer kan føre til musedød og dermed gi falskt positivt resultat (Suzuki et al. 1996). Av etiske hensyn burde unødvendig bruk av forsøksdyr unngås. MBA har også indikert at OA-gruppen og DTX-1 har kreftfremkallende effekter på huden til mus (Suganuma et al. 1988). På grunn av dette er metoden nå erstattet med en kjemisk analyse basert på LC-MS/MS og MBA brukes ikke for analyse av skjell til kommersiell omsetning i Norge lengre.

Bestemmelse av toksisiteten i skjell og analyse av fettsyreprofiler består av to helt forskjellige metoder. Toksisitet bestemmes med analyse på LC-MS/MS mens fettsyreprofiler bestemmes av analyse med GC-MS.

1.5.1 Analyse av toksiner med LC-MS/MS

Bestemmelse av toksisitet i skjell følger en metode basert på van den Top et al. 2011 og EU-RL-MB 2015. Metoden analyserer toksiner fra gruppene AZA, OA, DTX fra OA, PTX og YTX. Analyse av uhydrolysert skjellmateriale finner konsentrasjonen til OA, DTX-1 og DTX-2, mens et ekstra hydrolysesteg trengs for å finne DTX-3 fra OA. Toksinene kontrolleres i henhold til EU-direktiver.

Analysen gir noen ganger negative resultater både for prøvene og for kontrollprøver, som vitner om en svakhet i metoden (Aanrud 2016). På grunn av dette var det ønskelig med en optimalisering av metoden. Et alternativ kan være å erstatte metoden med multivariat statistikk for prediksjon av toksisiteten.

1.5.2 Fettsyreprofiler

Fett kan deles inn i tre fraksjoner; polare lipider, nøytrale lipider og frie fettsyrer. En vanlig måte å analysere fett i en prøve er ved bestemmelse av fettsyreprofil. Det er vanlig å derivatisere fettmolekylene ved omdannelse til fettsyreestere (Fatty Acid Methyl Ester, FAME) før analyse. Polare og nøytrale lipider bundet med ester-bindinger omestres mens frie fettsyrer forestres. Dersom en total fettsyreprofil skal bestemmes må disse to metodene kombineres for å gi en komplett analyse.

Fettsyreprofiler bestemmes som oftest ved bruk av gasskromatografi kombinert med et massespektrometer som detektor (GC-MS). Sertifisert referansemateriale (CRM) brukes ofte som ekstern standard til kvantifisering av fettsyrene.

1.6 Separasjon

Kromatografi er en separasjonsteknikk med en stillestående stasjonærfase og en mobilfase som beveger seg langs den stasjonære fasen. Substansene som analyseres danner konstant nye likevekter med den mobile og stasjonære fasen mens de beveger seg gjennom mobilfasen. Dermed separeres de fra hverandre.

1.6.1 Høy-ytelses væskekromatografi

Kromatografi deles inn etter hvilken type stasjonærfase eller mobilfase som brukes. Når mobilfasen er en eller flere væsker, kalles systemet høy-ytelse væskekromatografi (HPLC). HPLC bruker høyt trykk til å presse løsning gjennom lukkede kolonner som inneholder fine partikler som gir høy-oppløselige separasjoner.

1.6.2 Gasskromatografi

Gasskromatografi (GC) er en type kromatografi hvor mobilfasen er en gass. I GC analyseres prøver som kan fordampes uten å brytes ned. Prøven forflyttes gjennom kolonnen uten at den løses i mobilfasen. Ulikt kokepunkt for analyttene i prøvene og retensjon i kolonnen fører til separasjon i prøven.

1.7 Massespektrometri

I massespektrometri analyseres en analytt i forhold til masse-til-ladnings forholdet, m/z. I en ionekilde ioniseres prøven av høy-energetiske elektroner og ladede ioner dannes. Ionene filtreres av magnetiske eller elektriske felt i masseanalysatoren. En detektor detekterer så de ønskede ionene.

1.7.1 Ionisering

Før en prøve kan analyseres i massefilteret må den ioniseres. Prøven må være i gassform for å kunne ioniseres. For GC er prøven allerede en gass og den mest vanlige ioniseringsmåten er en elektronioniseringskilde (EI). Denne består av et filament som bombarderer prøven med elektroner inni et kammer. Det fører til fragmentering av noen ioner. Denne fragmenteringen gir informasjon om strukturen til analyttene i prøven.

Dersom MS skal brukes sammen med HPLC må prøven overføres fra væskefase til gassfase før ionisering. Den vanligste måten dette gjøres på er en elektronspray ioniseringskilde (ESI). I ESI brukes et elektrisk felt til å samle ladningen ved væskens overflate slik at det dannes en en dråpespray med høy ladning (de Hoffmann & Stroobant 2007). Sprayen føres gjennom enten en inert gass eller et oppvarmet kapillærør for å fjerne mobilfasen. Prøven er da ionisert og fortsetter inn i analysatoren.

1.7.2 Kvadrupoler

Kvadrupol massefilter er en analysator som bruker stabiliteten til banen i oscillerende elektriske felt for å separere ioner med hensyn til deres m/z forhold (de Hoffmann & Stroobant 2007). Analysatoren består av fire hyperbolske staver plassert parallelt i forhold til hverandre. Stavene ovenfor hverandre har samme ladning og nærliggende staver har motsatt ladning. Et ladet ion i området mellom stavene vil trekkes mot stavene med motsatt ladning. Dersom stavene endrer ladning før ionet treffer, endrer ionet retning. Ioner som er enten for store eller for små vil treffe stavene og ødelegges slik at kun de ønskede ionene går videre inn i detektoren.

1.7.3 Trippel kvadrupoler

En trippel-kvadrupol er et MS-system med tre kvadrupoler koblet sammen på rad. Den første kvadrupolen brukes som et massefilter som filtrerer eller analyserer ionene fra ionekilden. Den andre kvadrupolen fungerer som en reaksjonscelle hvor fragmentering induseres. Den siste kvadrupolen analyserer fragmentene som dannes i reaksjonscellen.

1.7.4 Fotomultiplikator

Fotomultiplikator er den mest vanlige detektoren brukt med MS. Ionene fra analysatoren akselereres med høy fart mot en konversjonsdynode hvor sekundære partikler emitteres som positive ioner, negative ioner, elektroner og nøytroner. Ionet av interesse endrer ladning når de treffer konversjonsdynoden. De sekundære partiklene konverteres til elektroner ved den første dynoden og akseleres så mot neste elektrode hvor flere sekundære elektroner produseres. Denne prosessen fortsetter over et gitt antall elektroder og danner dermed en forsterket strøm før måling.

1.8 Kjemometri

Kjemometri er læren om å hente ut data fra kjemiske systemer ved bruk av datastyrte metoder. Multivariat statistisk analyse, andvent matematikk og datavitenskap blir brukt til å løse kjemiske problemer. I beskrivende applikasjoner, blir egenskapene til kjemiske systemer modellert med hensikten å lære om de underliggende forholdene og strukturene til systemet som modellforståelse og identifisering.

I takt med den økende bruken av svært avanserte analytiske instrumenter i analyseforsøk har den resulterende datamengden økt betraktelig. En enkel analyse kan i dag gi flere datapunkter enn det er mulig å analysere manuelt. Dette krever mer avanserte metoder for å håndtere datamengdene. Multivariat statistikk kan enkelt analysere store datamengder og gi en mer grundig forståelse av datasettet enn kun visuell inspeksjon.

Målet med statistiske regresjonsmetoder (PCR/PLS) er å finne noen få lineære kombinasjoner av de originale variablene i **X** og bruke kun disse kombinasjonene i regresjonslikningen. På denne måten blir irrelevant informasjon forkastet og kun den relevante delen av variablene i **X** blir brukt videre.

1.9 Formål

Konstruksjon av regresjonsmetoder for tolkning og prediksjon er et viktig område innenfor anvendt statistikk og kjemometri.

Prinsipalkomponentanalyse (PCA) testes for å se på fordelingen i egenverdier, egenvektorer og eventuelle grupperinger mellom artene, stedene eller fettsyrene. Multivariate prediksjonsmetoder som Prinsipalkomponentregresjon (PCR), Partial Least Squares Regresjon (PLS) og Kanonisk Partial Least Squares Regresjon (CPLS) testes. I tillegg er variabelseleksjonsmetoden forlengs utvelgelse testet. Validering av resultatene og evnen til prediksjon beregnes med den kvadratiske gjennomsnittsfeilen for prediksjon (RMSEP) ved bruk av leave-one-out kryssvalidering og testsettvalidering. Test- og ka-

libreringsett defineres fra datasettet for å kontrollere prediksjonsevnen metoden oppnår. De beste metodene sammenlignes ved bruk av to-veis variansanalyse av kryssvaliderte prediksjoner (CVANOVA).

Formålet med denne oppgaven er å bruke multivariat statistikk for å finne en prediksjonsmodell for toksisitet i skjell basert på fettsyresammensetningen. Dersom giftige skjell har ulik fettsyresammensetning enn andre skjell kan dette gi nyttig informasjon som kan brukes til detektering av de giftige skjellene.

LC-MS/MS analysen som brukes til å kvantifisere toksinene kan være ustabil (Aanrud 2016). Dersom multivariat statistikk kan brukes til å finne en sammenheng mellom fettsyresammensetningen og toksisiteten kan denne analysen erstattes. Dette avhenger av en sterk sammenheng for at fettsyrene kan brukes til prediksjon. I tillegg til at LC-MS/MS metoden er noe ustabil krever den en del prøvepreparering. For å detektere DTX-3 i prøvene kreves i tillegg et ekstra hydrolysesteg. Å bruke GC til å finne fettsyreprofilen er en noe lettere metode. Erstatting av LC-MS/MS metoden med GC kan spare tid og arbeid.

1.10 Dataprogrammer

I denne oppgaven ble alle kalkuleringer utført ved bruk av R 3.2.3. Oppgaven er skrevet ved bruk av LaTeX. Strukturformler er tegnet ved bruk av ChemDraw 17.0.

2 Metodikk

Kompliserte målinger og store datasett er blitt mer vanlig ettersom maskiner og måleinstrumenter er blitt mer avanserte. Multivariat statistikk kan brukes til å forenkle store datasett til mindre, mer oversiktlige matriser og vektorer. Dette kapittelet presenterer grunnleggende statistikk og metoder for å finne prediksjonsmodeller.

2.1 Statistisk modell

For å beskrive resultatet fra en prøve brukes en statistisk modell. Responsen lagres i vektoren $\mathbf{y}$, forklaringsvariablene i matrisen $\mathbf{X}$ mens β betegner regresjonskoeffisientene til modellen.

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \times \underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon} \tag{2.1}$$

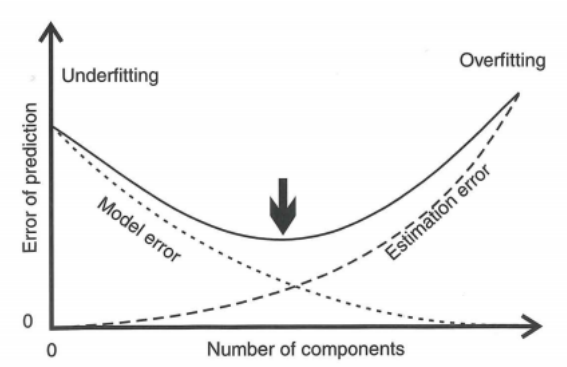
Alle reelle data inneholder støy som modelleres gjennom sannsynlighetsmodeller. I denne modellen er støyen modellert i feilleddet epsilon (ϵ) som antas å være uavhengig, normalfordelt med forventning null og ukjent varians, $\epsilon \sim N(0, \sigma^2 I)$. I en lineær modell er feilen ϵ definert som avstanden mellom $\mathbf{X}_i\beta$ og $\mathbf{y}_i$ altså $\epsilon = \mathbf{y}_i - \mathbf{X}_i\beta$.

Responsvektoren $\mathbf{y}$, matrisen med forklaringsvariabler $\mathbf{X}$, regresjonsfoeffisientene β og feilleddet ϵ ser ut som vist under. Innholdet i $\mathbf{y}$ vil endre seg etterhvert som forskjellige metoder analyseres, men stukturen vil forbli den samme.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix}, \qquad \mathbf{X} \times \beta = \begin{bmatrix} x_{11} & x_{12} & . & x_{1p} \\ x_{21} & x_{22} & . & x_{2p} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ . \\ . \\ \beta_p \end{bmatrix} \qquad og \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

Når et stort antall variabler introduseres i modellen øker sjansen for at estimeringsprosessen inkluderer støy og andre falske effekter fra kalibrasjonsdataene i den resulterende kalibrasjonsmodellen (Martens & Næs 1989). Da blir modellen overtilpasset. Når for få variabler introduseres i modellen kan det bety at modellen ikke er stor nok til å fange den viktige variasjonen i datasettet og mye av støyen kan forbli umodellert. Forholdet mellom prediksjonen og den sanne verdien vil bli dårlig og modellen vil da være undertilpasset.

Ettersom modellen blir mer og mer kompleks, kan den adaptere mer kompliserte underliggende strukturer og forventningsskjevheten blir redusert, men det fører også til økning av estimeringsfeilen og variasjonen (Hastie et al. 2001). Et sted mellom ligger den optimale modell kompleksiteten som gir lavest prediksjonsfeil, markert i figur 2.1.



Figur 2.1: Prediksjonfeil mot kompleksitet av modellen (Martens & Næs 1989)

2.1.1 Notasjoner

Fete små bokstaver (**y**) er vektorer mens fete store bokstaver (**X**) er matriser. $\mathbf{X}^T$ indikerer en transponert matrise definert ved ombytting av rader og kolonner i den opprinnelige matrisen **X**. Under regresjon indikerer **X** en $n \times p$ matrise med forklaringsvariabler mens **y** er en $n \times 1$ respons vektor. *Tr* angir trasen til en matrise som er summen av diagonalelementene og I er en identitetsmatrise. Alle parametre i modellen er angitt ved greske bokstaver og når parametrene estimeres brukes notasjonen hatt over den greske bokstaven (for eksempel $\hat{\mu}$ som estimat for forventningen) for å angi at dette ikke er den sanne verdien, men en tilnærmet gjetning.

2.1.2 Forventning

Forventningen til en variabel y er definert som den gjennomsnittlige verdien i utvalget og angitt med den greske bokstaven μ.

$$E(Y \mid x) = \mu_{Y|x} = \beta^T x \tag{2.2}$$

I praksis kan ikke forventningen finnes, men må estimeres som: $\hat{\mu}_{Y|x} = \hat{\beta}^T x$. Hvor $\hat{\beta}$ er et estimat for β og $\hat{\mu}$ er et estimat for μ. Når antallet observasjoner øker blir estimatet for β sikrere og gir dermed et bedre estimat for forventningen.

2.1.3 Varians og standardavvik

For ethvert datasett vil det oppstå forskjeller i dataene. Dette kan komme av forskjeller i utvalget som for eksempel biologiske, genetiske osv., eller eventuelt endringer av parametre. Et mål på denne spredningen er varians. Varians er angitt som σ^2 og er nærmere bestemt det gjennomsnittlige kvadratavviket. Den betingede variansen for y er gitt ved lign. 2.3.

$$Var(Y \mid x) = Var(\epsilon) = \sigma^2 I \tag{2.3}$$

Når variansen estimeres brukes $\hat{\sigma}^2$ for å angi at dette ikke er den sanne variansen, men en estimert verdi.

Standardavvik er et mål på spredningen i et datasett og finnes som kvadratroten av variansen. Dersom standardavviket er lite tyder det på at datapunktene ligger nært gjennomsnittet, mens stort standardavvik tyder på at datapunktene er spredt utover et større område. Standardavviket er definert som σ og når standardavviket estimeres brukes $\hat{\sigma}$.

2.1.4 Variansanalyse

Variansanalyse (ANOVA) er en betegnelse på metoder for å teste ulikheten mellom to eller flere grupper i en populasjon. Den observerte variansen i responsen deles inn i komponenter som hører til forskjellige kilder av variasjon. Den enkleste formen for ANOVA gir en statistisk test for likhet mellom gjennomsnittet av grupper og bruker F-tester til å sammenligne forskjellene (Montgomery 2013). En signifikant forskjell mellom grupper finnes hvis observasjonsstatistikken fra F-testen overstiger testobservatoren for et valgt signifikansnivå α.

2.1.5 Kovarians

Det er ofte en sammeheng mellom X og Y som påvirker måten disse varierer i forhold til hverandre. I multivariat analyse måles flere variabler, $x_1, \ldots, x_K$ for et antall objekter N. Hver av disse variablene har et gjennomsnitt og en varians, og derfor et standardavvik. I tillegg kan en kovarians mellom hvert par av variabler defineres. Kovarians er mål på den lineære avhengigheten mellom to variabler og er definert i lign. 2.4.

$$\sigma_{xy} = E(\mathbf{X} - \mu_x)(\mathbf{Y} - \mu_y) \tag{2.4}$$

Når kovariansen estimeres brukes $\hat{\sigma}_{xy}$ og ligningen gitt i 2.5.

$$\hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2.5}$$

På matriseform er kovariansen mellom variablene på sentrert form gitt i lign. 2.6.

$$\hat{\Sigma}_{xx} = \frac{(\mathbf{X}^T\mathbf{X})}{n-1} \tag{2.6}$$

Kovariansmatrisen som dannes fra lign. 2.6 får formen:

$$\Sigma_{xx} = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & \cdot & cov(x_1, x_k) \\ cov(x_2, x_1) & var(x_2) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ cov(x_k, x_1) & \cdot & \cdot & var(x_k) \end{bmatrix}$$

Dersom kovariansmatrisen til $\mathbf{X}$ har rang $n < p$, kan den totale variasjonen til $\mathbf{X}$ forklares av de første n prinsipalkomponentene (Mardia et al. 1979). Matrisen har ikke full rang siden $n < p$ og dermed vil de resterende $p - n$ egenverdiene være tilnærmet lik null.

2.1.6 Korrelasjon

Korrelasjon er et mål på den lineære sammenhengen mellom variabler og defineres ved å dividere kovariansen med produktet til standardavviket mellom variablene. Korrelasjon har fordelen over kovarians ved at den er uavhengig av variabelenhetene og tar verdier i intervallet [-1,1]. Korrelasjonen i et utvalg er gitt i lign. 2.7.

$$\widehat{Corr(X,Y)} = \hat{\rho}(X,Y) = \frac{\widehat{Cov(X,Y)}}{\hat{\sigma}_X * \hat{\sigma}_Y} \tag{2.7}$$

På matriseform er korrelasjonen mellom variablene på sentrert form gitt i lign. 2.8.

$$\hat{\rho} = (diag\hat{\Sigma})^{-1/2})\hat{\Sigma} \times diag\hat{\Sigma}^{-1/2} \tag{2.8}$$

2.1.7 Egenverdier og egenvektorer

Når man har en $p \times p$ matrise $\mathbf{X}$ og en $p \times p$ identitetsmatrise $\mathbf{I}$ defineres løsningene $\lambda_1, \lambda_2, \ldots, \lambda_p$ til polynomlikningen $\mid \mathbf{X}^T\mathbf{X} - \lambda\mathbf{I} \mid = 0$ som egenverdiene til $\mathbf{X}$ (Johnson & Wichern 2002). Matrisen er $(\mathbf{X}^T\mathbf{X})$ hvor $\mathbf{X}$ er sentrert.

$$(\mathbf{X}^T\mathbf{X})e_i = \lambda_i e_i \tag{2.9}$$

En egenvektor for en $p \times p$ matrise $\mathbf{X}$ er en vektor $\mathbf{e}$ med tall slik at lign. 2.9 for en skalar λ oppfylles. Skalaren λ er egenverdien til $\mathbf{X}^T\mathbf{X}$ dersom det er en ikke-triviell løsning for $\mathbf{e}$ slik at $\mathbf{e}$ blir egenvektoren korresponderende til λ (Lay et al. 2016).

Relevante komponenter er et viktig begrep i forhold til komponentene som inkluderes i metoden som velges. Dersom alle egenvektorene e_i som gir lign. 2.10 ligger i området gitt av prediktoren for $\mathbf{y}$ gitt $\mathbf{X}$ (vist i lign. 2.2) er disse de relevante egenvektorene med korresponderende relevante egenverdier (Næs & Helland 1993). Altså må egenvektoren være korrelert til $\mathbf{y}$ for å kunne være relevant. Dersom de relevante egenverdiene er små fører dette til dårlig prediksjonsevne.

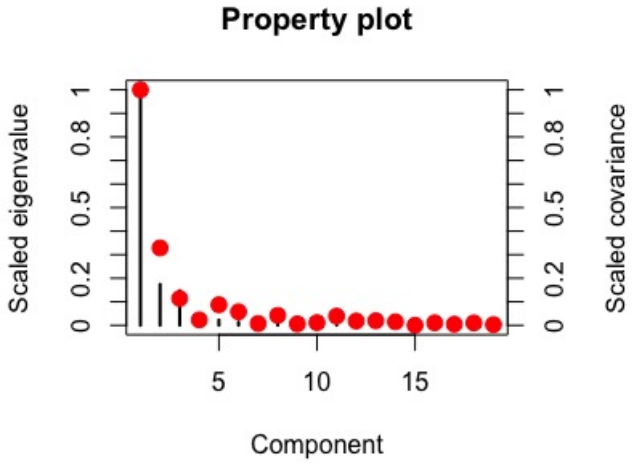
$$e_i^T \sigma_{xy} \neq 0 \tag{2.10}$$

Dersom egenvektorene ikke er korrelert til $\mathbf{y}$ er de irrelevante og oppfyller ling. 2.11.

Dersom de irrelevante egenverdiene er store fører dette til dårlig prediksjonsevne.

$$e_i^T \sigma_{xy} = 0 \tag{2.11}$$

En måte å illustrere egenverdiene i forhold til korrelasjonen eller kovariansen med responsen er et plot kalt egenskapsplot (Sæbø et al. 2015). Denne typen plot viser de skalerte egenverdiene mot korrelasjonen eller kovariansen mellom egenvektoren til komponenten og responsen. Dette gir en god indikator for hvilke egenverdier som er mest relatert til responsen. Figur 2.2 viser et eksempel på et egenskapsplot med egenverdier og de røde prikkene viser skalert kovarians mellom prinsipalkomponentene og responsen. Kovariansen er kun ulik null for de relevante komponentene.



Figur 2.2: Egenskapsplot med skalerte egenverdier mot skalert kovarians

Ideelt sett skal et slikt plot ha egenverdier som synker som i figur 2.2 og kovarians eller korrelasjon som synker i takt med egenverdiene. Det betyr at de første komponentene som har de høyeste egenverdiene også skal ha høyest kovarians med responsen.

2.1.8 Kollinearitet

Variablene i **X** er kollineære dersom kollonnene i **X** er tilnærmet eller helt lineært avhengige. Altså er **X** kollineær dersom minst en av X-variablene kan skrives som en tilnærmet eller eksakt lineær kombinasjon av de andre.

Kollinearitet kan skyldes avhengighet mellom variablene eller målinger hvor antallet variabler er for lite. Dersom n er mindre enn p vil det alltid være kollinearitet i dataene selv om det ikke er kollinearitet i populasjonen. Et annet tegn på kollinearitet er stor forskjell i størrelsene på egenverdiene.

2.1.9 Residualer

Residualer er et anslag på det ukjente feileddet som kan påvirke effekten av modellen. Dette kan for eksempel være støyen eller den irrelevante variabiliteten i **X** og **y**.

Residualene fanger opp all variasjonen i responsen som modellen ikke klarer. Det er ønskelig med en modell som forklarer mest mulig av variasjonen ved hjelp av forklaringsvariablene. Da blir residualene lik null og kan brukes til å si noe om modellens forklaringsevne, hvor godt modellen forklarer forklaringsvariablene. Residualene finnes som avstanden mellom observasjonene og utvalgsgjennomsnittet vist i lign. 2.12.

$$\hat{\epsilon}_i = y_i - \hat{y}_i \tag{2.12}$$

Hvor $\hat{\epsilon}_i$ er residualene mens y_i er responsen og $\hat{y}_i = \hat{\beta}^T \mathbf{x}_i$.

2.1.10 Uteliggere

En uteligger er en observasjon som skiller merkbart fra de andre observasjonene i utvalget og vekker mistanke om at den kan ha kommet fra en annerledes mekanisme

(Khanmohammadi 2014). Uteligger deteksjon er en veldig viktig del av eksplorativ multivariat dataanalyse. En variabel kan få veldig stor varians når noen verdier viker veldig fra gjennomsnittet. Dersom en eller flere målinger er langt unna gjennomsnittet eller de andre målingene kan det føre til overestimering av variansen og dermed standardavviket. Da er det viktig å sjekke om målingen er riktig eller om det kan være en såkalt uteligger.

Uteliggere kan være tegn på at det har skjedd en feil. Det kan være feil i måling, registrering, instrumenter osv. En uteligger kan også være en måling eller observasjon som ikke er representativ for populasjonen. Dersom en modell blir tilpasset med en uteligger kan dette føre til over- eller underestimering av parametre (som regresjonskoeffisienten β i lign. 2.1) og kan gi en dårlig prediksjon.

2.2 Minste kvadraters metode

Minste kvadraters (Least Squares, LS) metode består av å finne verdier for regresjonskoeffisienten β som minimerer kvadratsummen av avstanden mellom målingene og de tilpassede verdiene, altså minimere residualene.

Skalarproduktet som skal minimeres kan skrives med residualleddet ϵ som gitt i lign. 2.13.

$$\epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{2.13}$$

Ved å minimere lign. 2.13 kan regresjonskoeffisienten $\hat{\beta}$ estimeres ved lign. 2.14.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.14}$$

Matrisen $\mathbf{X}^T\mathbf{X}$ er invertert, som krever at kolonnene til $\mathbf{X}$ er lineært uavhengige. Dersom $n < p$ har ikke matrisen full rang og da vil ikke $(\mathbf{X}^T\mathbf{X})^{-1}$ eksistere. Et annet stort problem

med minste kvadraters regresjon er forekomsten av kollinearitet i datasettet (Mandel 1982). Dersom det er multikolinearitet mellom variablene gir LS et veldig ustabilt estimat av β som kan føre til dårlig prediksjon.

$$trVar(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{2.15}$$

Når $n < p$ viser lign. 2.15 at estimatet for β blir veldig ustabilt. Mange egenverdier vil være svært lave og dette fører til stor varians for β. Algoritmen modifiseres ofte for å hanskes med dette problemet, og den vanligste modifikasjonen er prinsipal komponent regresjon.

2.3 Variabelseleksjon

Variabelseleksjon er en metode hvor antallet forklaringsvariabler blir redusert for å lage en submodell som bare inneholder den relevante informasjonen fra forklaringsvariablene. Variabelseleksjon brukes for å forbedre prestasjonen til modellen og gi bedre prediksjoner. Dersom det er flere variabler enn prøver ($p > n$), er det mulig å finne det antallet variabler som er korrelert til egenskapen som skal predikeres ved en tilfeldighet eller ved overtilpasning (Andersen & Bro 2010). En slik modell kan føre til veldig dårlig prediksjon når den brukes på nye prøver.

Variablene med lave korrelasjoner til responsen fjernes slik at kun de variablene med høy korrelasjon blir igjen i modellen. Variabelseleksjon kan dermed brukes for å øke korrelasjoner mellom variablene for å forbedre prestasjonen og prediksjonskapabiliteten til modellen (Seisonen et al. 2016). Variablene som inkluderes i modellen skal i tillegg være relativt ukorrelerte til hverandre.

2.3.1 Forlengs utvelgelse

Forlengs utvelgelse (forward selection) er en metode for stegvis utvelgelse av signifikante variabler. Denne utvelgelsesmetoden begynner med en modell som ikke inneholder noe. Første steg for å finne metode er at alle korrelasjons koeffisientene blir beregnet mellom $\mathbf{y}$ og hver x_i. Forklaringsvariabelen med lavest korrelasjonskoeffisient inkluderes i metoden dersom p-verdien er lavere enn et valgt signifikansnivå (α). Neste steg skjer kun dersom den første variabelen inkluderes i metoden. Da blir t-verdien for de resterende variablene beregnet og den variabelen med høyest absolutt t-verdi inkluderes i metoden dersom p-verdi < α. Metoden fortsetter med beregning av t-verdier og inkludering av variabler til p-verdi > α.

Hovedproblemet med denne typen utvelgelse er at det blir gjort et stort antall t-tester. Under utvelgelsen dannes det en $k \times p$ matrise med ladninger som definerer variabelen. Et problem med dette er at ladningene består av 0 og 1 ettersom om variabelen tas med i metoden eller ikke. Dette gir en veldig rigid og lite fleksibel metode som ikke kan måle seg med metoder fra PCR og PLS som har ladninger basert på egenvektorene til matrisen.

2.4 Prinsipalkomponentanalyse

Prinsipalkomponentanalyse (PCA) er en metode for datareduksjon. Dataene som brukes antas å være sentrerte og nye variabler blir laget som lineære kombinasjoner av de originale; med scoringer som definerer lengden og størrelsen. Matematisk blir konstruksjonen av de nye variablene oppnådd ved å finne egenvektorene for variansmatrisen ($\hat{\Sigma}_{xx}$) til de originale variablene. Egenvektorene blir da ladningene for konstruksjon av nye variabler og korresponderende egenverdier forteller hvor mye av den originale variansen som fanges i hver nye variabel (Næs et al. 2002). Variablene som blir laget er ukorrelerte i forhold til hverandre og under konstruksjonen av regresjonsmodellen lages en vektor med scoringer som blir vektet i regresjonen. Egenverdiene inneholder så

mye som mulig av variasjonen og har blitt konstruert til å ha maksimal varians mellom alle lineære kombinasjoner av variablene. Metoden PCA benytter er gitt i lign. 2.16.

$$\underset{n\times k}{\mathbf{Z}} = \underset{n\times p}{\mathbf{X}} \times \underset{p\times k}{\mathbf{E}_k} \tag{2.16}$$

Egenverdiene lages ifølge 2.16 hvor $Var(z_i) = \lambda_i$ og det antas at k < p. Egenverdiene for komponentene dannes derfor i matrisen $\mathbf{E}_k$ i synkende rekkefølge $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$ til $k = n$.

$$\mathbf{E}_k = \begin{bmatrix} e_1 & e_2 & \ldots & e_k \end{bmatrix}$$

Summen av de første k egenverdiene dividert med summen av alle egenverdiene representerer andelen av den totale variasjonen forklart av de første k prinsipalkomponentene (Mardia et al. 1979). Dette gir $(\lambda_1 + \ldots + \lambda_k)/(\lambda_1 + \ldots + \lambda_p)$.

En metode som inneholder få komponenter og de største egenverdiene gir ofte en bedre submodell enn når mange komponenter er med. Ved bruk av veldig mange komponenter kan støy fra feilleddet tas med, forstyrre metoden og føre til overtilpassning. Derfor er det lurt å gjennomføre komponentseleksjon basert på validerings kriterier som kryssvalidering eller visuell inspeksjon av ladningene for forståelse av komponentene. Et nyttig verktøy for å bestemme hvor mange prinsipalkomponenter som skal beholdes er et screeplot. Dette er et plot av egenverdiene i rekkefølgen til prinsipalkomponentene. Antallet komponenter velges på det punktet i plottet hvor de gjenværende egenverdiene er relativt små og har omtrentlig samme størrelse (Johnson & Wichern 2002).

2.5 Prinsipal komponent regresjon

Prinsipal komponent regresjon (PCR) er regresjon av et valgt sett prinsipalkomponenter $\mathbf{Z}$ som maksimerer variansen i $\mathbf{X}$ mot responsen $\mathbf{y}$. Minste kvadrater brukes som regresjonsmetode på de utvalgte komponentene. Hensikten med PCR er å uttrykke hovedinformasjonen i variablene $\mathbf{X}$ med et mindre antall variabler, altså prinsipalkomponentene til $\mathbf{X}$. PCR bruker lign. 2.16 for å estimere prinsipalkomponentene.

Regresjonskoeffisenten $\hat{\beta}_{\mathbf{Z}}$ for komponentene estimeres så ved bruk av minste kvadrater vist i lign. 2.17.

$$\begin{aligned} \hat{\beta}_Z &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T y \\ &= (\mathbf{E}_k^T\mathbf{X}^T\mathbf{X}\mathbf{E}_k)^{-1}\mathbf{E}_k^T\mathbf{X}\mathbf{y} \end{aligned} \tag{2.17}$$

For matrisen $\mathbf{X}$ med forklaringsvariabler estimeres regresjonskoeffisienten $\hat{\beta}_{\mathbf{X}}$ ved lign. 2.18.

$$\begin{aligned} \hat{\beta}_{\mathbf{X}} &= \mathbf{E}_k \times \hat{\beta}_Z \\ &= \mathbf{E}_k(\mathbf{E}_k^T\mathbf{X}^T\mathbf{X}\mathbf{E}_k)^{-1}\mathbf{E}_k^T\mathbf{X}\mathbf{y} \\ &= \mathbf{E}_k(\lambda_k)^{-1}\mathbf{E}_k^T\mathbf{X}^T \end{aligned} \tag{2.18}$$

Med PCR løses mye av kolinearitets problemene og mer stabile regresjonsligninger og prediksjoner oppnåes (Næs et al. 2002). Stabiliteten kommer av at variablene med minst varians fjernes fra regresjonen.

$$tr(Var(\hat{\beta})) = \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i} \tag{2.19}$$

Lign. 2.19 viser at den totale variansen til $\hat{\beta}$ minkes ved at de minste egenverdiene fjernes. Med veldig små egenverdier blir den totale variansen veldig stor. Dette gir en metode med bedre prediksjonsevne og et mer stabilt estimat for β enn LS oppnår. Et

problem som oppstår er at dette er en forventningsskjev estimator. Forventningsskjevheten er gitt i lign. 2.20.

$$(E\hat{\beta} - \beta)^T (E\hat{\beta} - \beta) = \sum_{i>k}^{p} (e_i^T \beta)^2 \tag{2.20}$$

2.6 Partial least square regresjon

Partial Least Squares regresjon (PLS) er en metode for å se på sammenhengen av en matrise og en vektor, **X** og **y**. Istedet for å velge ut et visst antall prinsipalkomponenter som for PCR, bruker PLS faktorer bestemt ved å maksimere kovariansen mellom **y** og alle mulige lineære kombinasjoner av **X** (Næs et al. 2002). Dette leder til komponenter som er mer direkte relatert til **y** enn prinsipalkomponentene. I PLS antas det at systemet som undersøkes er påvirket av noen få underliggende latente variabler. Et av målene med PLS er å estimere antallet latente variabler før prediksjon. Scoringene for PLS er gitt i lign. 2.21.

$$\underset{n\times k}{\mathbf{T}} = \underset{n\times p}{\mathbf{X}} \times \underset{p\times k}{\mathbf{W}} \tag{2.21}$$

Hvor **W** er en matrise med vekt-ladningene for regresjonen gitt som lineære kombinasjoner av **X**. Scoringene i **T** både predikerer **y** og modellerer **X** (Wold et al. 2001). Den første vekt-vektoren ($\mathbf{w}_1$) er den første egenvektoren til den kombinerte varians-kovarians matrisen ($\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$), og de følgende vektorene er egenvektorene til de deflaterte versjonene av samme matrise (Wold et al. 2001). Kovariansen mellom **X** og **y** for komponentene maksimeres ved vekt-vektoren $\mathbf{w_k}$. Ligningen er gitt i 2.22 og er basert på deflatering av **X** og **y**.

$$\mathbf{w}_k = \mathbf{w}_k^T \mathbf{X}_{k-1} \mathbf{y}_{k-1} \tag{2.22}$$

Ladningene til komponentene bestemmes med minste kvadraters tilpasning og ladningsvektorene ($\mathbf{t}_k$) er kolonnene i scorings matrisen $\mathbf{T}$. Disse brukes til redusering av $\mathbf{X}$ vist i lign. 2.23.

$$\mathbf{p}_k = \frac{\mathbf{X}^T \mathbf{t}_k}{\mathbf{t}_k^T \mathbf{t}_k} \tag{2.23}$$

Regresjonskoeffisienten $\hat{\beta}_T$ brukt i PLS prediktoren beregnes med lign. 2.24.

$$\hat{\beta}_T = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{y} \tag{2.24}$$

Regresjonskoeffisienten $\hat{\beta}_\mathbf{X}$ brukt i PLS prediktoren beregnes med lign. 2.25.

$$\hat{\beta}_x = \mathbf{W}(\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{y} \tag{2.25}$$

PLS finner de latente variablene i $\mathbf{X}$ som har størst kovarians til $\mathbf{y}$. Denne automatiske effekten finnes ikke i PCR. PCR gjør det best når de irrelevante egenverdiene er relativt små eller relativt store. De relevante egenverdiene til PCR og PLS er egenverdiene til komponentene som inkluderes i metoden.

Både PCR og PLS lager et antall lineære kombinasjoner av $\mathbf{X}$ som er mindre enn p. Siden disse lineære kombinasjonene må estimeres fra datasettet fører dette til et bidrag av tilfeldighet og en økning i variansen innad i metoden som lages (Helland 2001).

2.6.1 Kanonisk PLS

Kanonisk PLS (CPLS) er en utvidelse av PLS som bruker tilleggsinformasjon som vekter og ekstra målinger til å redusere antallet komponenter (Indahl et al. 2009). CPLS bruker kanonisk korrelasjon til å se på korrelasjonen mellom en lineær kombinasjon av variablene i et datasett og en lineær kombinasjon av variablene i et annet datasett.

Det beregnes midlertidige ladningsvekter gitt som $W_0 = \mathbf{X}^T[\mathbf{Y} \quad \mathbf{Y}_{add}]$ hvor $\mathbf{X}$ og $\mathbf{Y}$ er deflatert. Beregningen av de midlertidige scoringsvektorene $\mathbf{Z}_0$ er vist i lign. 2.26.

$$\underset{n\times(q+1)}{\mathbf{Z}_0} = \underset{n\times p}{\mathbf{X}} \times \underset{p\times(q+1)}{\mathbf{W}_0} \tag{2.26}$$

Kanonisk korrelasjon brukes så til å justere ladningsvektene ved at det blir funnet to vektorer $\mathbf{a}$ og $\mathbf{b}$ som maksimerer korrelasjonen mellom matrisene $\mathbf{X}$ og $\mathbf{Y}$ (Liland 2009). Vektoren $\mathbf{a}$ og $(q+1)\times 1$ vektoren $\mathbf{b}$, hvor q er antallet tilleggsresponser, er definert slik at de maksimerer korrelasjonen mellom $\mathbf{Yb}$ og $\mathbf{Z}_0\mathbf{a}$. Disse brukes til å produsere ladningsvektene ved $\mathbf{W}_0\mathbf{a} = \mathbf{w}$ som gjenntas for hver komponent.

Ekstra responsvariabler kan brukes til å spenne et større rom for de midlertidige ladningsvektorene (Liland 2009). Denne nye responsen vektes ikke når metoden brukes til prediksjon og brukes kun når metoden lages. Kategoriske variabler gjøres om til dummyvariabler med koding 1 eller 0. Responsen blir en matrise $\mathbf{Z}_0$ med dimensjon $n\times(q+1)$ med den originale $n\times 1$ responsen $\mathbf{Y}$ og q flere $n\times 1$ vektorer med nye responser. Dette fører til ladningsvekter som er bedre tilpasset fordi de utnytter denne ekstra informasjonen som ikke gis i PLS eller PCR.

2.7 Prediksjon

Å predikere er å anslå responsen ($\hat{y}$) ved en gitt verdi av $\mathbf{x}$. Prediksjonsmodellen er gitt i lign. 2.27.

$$\hat{y} = \bar{y} + \hat{\beta}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{2.27}$$

Hvor $\hat{\mathbf{y}}$ er responsen som predikeres, $\bar{\mathbf{y}}$ er gjennomsnittet av responsvektoren $\mathbf{y}$ tatt over kalibrasjonsettet, $\bar{\mathbf{x}}$ er gjennomsnittet til forklaringsvariablene tatt over kalibrasjonsettet og $\hat{\beta}$ er regresjonskoeffisienten for kalibreringsettet. $\mathbf{x}$ er de nye kjente forklarings-

variablene.

Lign. 2.27 er prediksjonsmodellen for ikke standardiserte forklaringsvariabler. Når forklaringsvariablene i kalibreringsettet standardiseres må også de nye forklaringsvariablene **x** standardiseres. Den standardiserte prediksjonsmodellen er gitt i lign. 2.28.

$$\hat{y} = \bar{y} + \hat{\beta}^T \mathbf{x}^* \tag{2.28}$$

Før **x** kan settes inn i lign. 2.27 må den standardiseres som vist i lign. 2.29. De samme verdiene som er brukt til standardiseringen av kalibreringsettet må brukes til standardisering av de nye forklaringsvariablene. Standardisering av hver variabel ($\mathbf{x}_i^*$) i $\mathbf{x}^*$ er vist i lign. 2.29.

$$x_i^* = \frac{x_i - \bar{x}_{i,cal}}{SD(x_{i,cal})} \tag{2.29}$$

Dersom modellen som er gitt stemmer og prediksjonsfeilen estimeres som $E(y - \hat{y})^2$ blir den beste prediktoren for responsen lign. 2.30. Med denne nås nedre grense for prediksjon som er σ^2.

$$\hat{y} = E(y \mid \mathbf{x}) = \beta^T \mathbf{x} \tag{2.30}$$

Nullmetoden til en statistisk metode er det punktet hvor ingen forklaringsvariabler inkluderes i metoden. Da blir leddet $\hat{\beta} = 0$ og prediksjonen blir $\bar{\mathbf{y}}$, gjennomsnittet til responsen. Dersom prediksjonsfeilen til en metode er større enn prediksjonsfeilen til nullmetoden, er det bedre å bruke ($\bar{y}$) som prediksjonsmodell.

2.7.1 Prediksjonsfeil

Den største forskjellen mellom prediksjon og estimering er at estimering utnytter informasjonen i både responsvektoren og forklaringsvariablene til å estimere parametrene i modellen. Prediksjon bruker kun informasjonen tilgjengelig i forklaringsvariablene og estimerte parametre til å predikere en variabel mens den sanne responsvektoren kun brukes til å evaluere kvaliteten på prediksjonen. Estimering kan ikke evalueres på denne måten fordi all kjent informasjon brukes til å estimere ukjente parametre.

Prediksjonsfeil (PE) er et mål på hvor godt en metode vil forutse nye observasjoner. Når $n < p$, blir $E\hat{\beta} \neq \beta$ og alle estimatene vil være forventningsskjeve. Prediksjonsfeilen kan hovedsaklig forklares av tre deler: modellfeilen, estimeringsfeilen og feilleddet ϵ (Helland & Almøy 1994). Når antallet variabler eller komponenter øker synker modellfeilen siden mer av variansen i x modelleres. Samtidig øker estimeringsfeilen ettersom at antallet parametre som må estimeres øker. Når antallet forklaringsvariabler nærmer seg antallet prøver i utvalget øker prediksjonsfeilen. Når $\hat{\beta} = \beta$ viser lign. 2.31 at $\sqrt{\sigma^2}$ er nedre grense for prediksjon.

$$\sqrt{(E\hat{\mathbf{Y}} - \mathbf{Y})^2} = \theta = \sqrt{\sigma^2 + [(E\hat{\beta} - \beta)^T \Sigma_{xx} (E\hat{\beta} - \beta)] + trVar\hat{\beta}\Sigma_{xx}} \tag{2.31}$$

Prediksjonsfeilen i lign. 2.31 tar forventningen over alle mulige fremtidige forklaringsvariabler og responser. Denne er frigjort fra alle kalibreringsdata siden $E(\hat{\beta})$ og $var(\hat{\beta})$ er ukjente. Dette gjør at det ikke er mulig å finne den sanne prediksjonsfeilen.

Når antallet komponenter inkludert i metoden er lavt, blir den sterkt påvirket av at estimatene for β er forventningsskjeve. Med mange komponenter vil $E\hat{\beta}$ nærme seg den sanne β som betyr at skjevheten reduseres. Med et høyere antall komponenter inkluderes mer støy og variansen øker. Etterhvert som veldig mange komponenter inkluderes, blir variansen veldig høy. Derfor aksepteres skjevhet i metoden mot at variansen blir lavere. Der finner man det optimale antallet komponenter for hver metode.

Prediksjonsfeilen faller raskere i forhold til antall komponenter ved bruk av PLS i for-

hold til PCR ettersom at PLS tar hensyn til kovariansen mellom **y** og **X**. Dette kan gi PLS-komponentene en bedre prediksjonsevne.

2.8 Validering av prediksjonskvalitet

For å bestemme hvilken av de konstruerte prediksjonsmodellene som fungerer best kan det brukes flere forskjellige metoder. For regresjon brukes ofte valideringskriterier som tar modellens evne til prediksjon med i betraktning.

Kriterier som $R^2, R^2_{Adj}, MallowsC_p$ og AIC leder ofte til overtilpasning fordi de kun måler graden av tilpasning og ikke betydningen av prediksjonskvaliteten til modellen (Höskuldsson 2000). Disse målene er ikke basert på selve prediksjonen, de er kun basert på hvor godt datasettet passer til modellen. Når mange forklaringsvariabler inkluderes i modellen kan alle disse kriteriene tvinges til å ha høye verdier og da er modellen overtilpasset.

Derfor brukes ofte RMSEP, R^2_{pred}, kryssvalidering og testsettvalidering til validering av prediksjonsevnen til modellen.

2.8.1 Root Mean Square Error of Prediction

Root Mean Square Error of Prediction (RMSEP) er en valideringsmetode som baseres på den kvadrerte forskjellen mellom den sanne responsen og den predikerte responsen. For svært avanserte regresjonsmetoder er det ikke mulig å beregne forventningen til $\hat{\beta}$ eller variansen til $\hat{\beta}$. Derfor brukes $\widehat{RMSEP} = \hat{\theta}$ gitt i lign. 2.32 som et estimat for den forventede prediksjonsfeilen.

$$\widehat{RMSEP} = \hat{\theta} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_{i,j})^2} \tag{2.32}$$

Hvor m er antall observasjoner i testsettet eller kalibreringsettet under kryssvalidering.

Den sanne responsen for observasjon i er angitt av y_i og $\hat{y}_{i,j}$ angir de predikerte responsene funnet ved enten kryssvalidering eller testsett validering hvor j angir forskjeller i metode og komponent. Kvadratroten tas av den kvadrerte estimatforskjellen mellom $\mathbf{y}$ og $\hat{\mathbf{y}}$. Dette betyr at utslaget en eventuell uteligger i testsettet vil gi på kvaliteten av prediksjonen reduseres.

Antallet komponenter som holdes igjen bestemmes fra et plot av RMSEP-verdiene for metoden. Der plottet når et punkt hvor RMSEP er lav sammenlignet med resten og RMSEP ikke endrer seg veldig ved å holde igjen en komponent til, angir hvor mange komponenter som skal være med i prediksjonsmodellen.

2.8.2 R^2_{pred}

Et annet mål på kvaliteten på en prediksjonsmodell er R^2_{pred}. Dette er et mål på korrelasjonen i modellen og er gitt i lign. 2.33.

$$R^2_{pred(j)} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{(i),j})^2}{\sum_{i=1}^{2}(y_i - \bar{y})^2} \tag{2.33}$$

Når $R^2_{pred(j)}$ er høy, blir en stor andel av variansen i datasettet forklart av modellen. Da ansees modellen for å være god. Dersom $R^2_{pred(j)}$ er negativ vil nullmetoden gi en bedre prediksjonsmodell. Da vil $\bar{\mathbf{y}}$ gi en bedre prediksjon av responsen enn modellen.

2.8.3 Kryssvalidering

Dersom det ikke er nok data tilgjengelig for å lage et kalibrering- og testsett kan kryssvalidering brukes til å sjekke prestasjonen av prediksjonsmodellen.

Leave-One-Out Kryssvalidering (LOOCV) er en validerings metode hvor en observasjon holdes utenfor mens en modell tilpasses på de resterende obervasjonene. Den første observasjonen fjernes fra datasettet og modellen blir tilpasset basert på de $n-1$

gjenværende observasjonene. Den nye modellen blir testet ved å sammenligne kvadratforskjellen mellom den predikerte verdien $\hat{\mathbf{y}}$ og den sanne $\mathbf{y}$ for den fjernede observasjonen. Den første observasjonen settes tilbake inn i datasettet og prosedyren gjentas ved å fjerne observasjon to. Dette fortsetter til alle observasjonene har blitt fjernet en gang. Kvaliteten på modellen kontrolleres gjennom den kvadratiske gjennomsnittsfeilen til kryssvalidering (RMSECV) gitt i lign. 2.34.

$$\widehat{RMSECV} = \hat{\theta}_{CV} = \sqrt{\sum_{i=1}^{n} (\hat{y}_{CV,i} - y_i)^2 / n} \tag{2.34}$$

Kvadratsummen for forskjellen mellom den predikerte verdien $\hat{\mathbf{y}}$ og den sanne $\mathbf{y}$, beregnes for alle de parallelle modellene for å finne RMSECV. Denne estimerer prediksjonsevnen til den kryssvaliderte metoden.

2.8.4 Kalibreringsett og testsett

Den beste valideringen av en modell er om den konsistent predikerer Y-verdiene presist for observasjoner med helt nye X-verdier (Wold et al. 2001). Dette gjøres ofte ved å dele datasettet i to, et kalibreringsett og et testsett. Kalibreringsettet brukes til å tilpasse en prediksjonsmodell. Basert på denne modellen estimeres $\hat{\beta}$. Testsettet settes så inn i modellen med den estimerte $\hat{\beta}$ og responsen $\hat{\mathbf{y}}$ predikeres så for testsettet. Den gjennomsnittlige kvadratavstanden mellom den predikerte $\hat{\mathbf{y}}$ og den sanne responsen $\mathbf{y}$ beregnes. Dersom forskjellen er stor vil ikke prediksjonsmodellen være robust og den vil gi dårlige prediksjoner i fremtiden.

Når prediksjonsmodellen blir konstruert brukes kun dataene fra kalibrasjonsettet. Testsettet holdes utenfor modellen og brukes kun til å vurdere kvaliteten på prediksjonene modellen gir.

2.9 Metodevalidering

For å sjekke om en metode kan ansees for å være signifikant bedre enn en annen, sjekkes det om forskjellen i prediksjonsevne er så stor at den ikke kan komme av kun tilfeldig støy (Cederkvist et al. 2005). En måte å vurdere modellene er to-faktor ANOVA for å teste for signifikante forskjeller mellom metodene. For hver metode og hver prøve beregnes forskjellen mellom responsen og det kryssvaliderte estimatet basert på gjenværende prøver ved bruk av den gitte metoden (Indahl & Næs 1998). Metoden kalles CVANOVA ettersom at det er variansanalyse av kryssvaliderte prediksjoner. Modellen er gitt i lign. 2.35.

$$z_{ij} = \mu + A_i + \tau_j + \epsilon_{ij} \tag{2.35}$$

Hvor i angir observasjon i = 1,..., M, og j angir metode j = 1, ..., n. τ_j antas å være normalfordelt med forventning null og varians σ_τ^2. ϵ_{ij} antas å være normalfordelt med forventing null og varians σ_ϵ^2. Responsen z_{ij} angir den kvadratiske forskjellen mellom den predikerte responsen og den sanne responsen (MESP). Hver prøve regnes som representativ for en større populasjon. Dermed kan effekten av prøvene regnes som en tilfeldig effekt og dette blir en blandet modell med både tilfeldige effekter og faste effekter.

2.9.1 Tukey par-vis kontrast

Tukey tester sammenligner alle mulige par av forventninger og baseres på studentfordelingen q (Tukey 1949). Etter variansanalyse (ANOVA) hvor nullhypotesen om lik forventning mellom metodene forkastes, kan det være ønskelig med parvis test av forventningene. Denne type test er en *post hoc* analyse fordi den ikke er planlagt eller gjennomført før etter at hovedanalysen er gjort. Tukey bruker observatoren gitt i lign. 2.36.

$$\hat{\Gamma}_{ij} = \hat{\tau}_i - \hat{\tau}_j \tag{2.36}$$

Hypotesene som testes sjekker forskjellen mellom gruppegjennomsnittene og er gitt i lign. 2.37.

$$\begin{aligned} H_0 &: \tau_i = \tau_j \\ H_1 &: \tau_i \neq \tau_j \end{aligned} \tag{2.37}$$

Hvor $i \neq j$. Denne testen kontrollerer den eksperimentvise feilraten ved det valgte signifikansnivået α (Montgomery 2013). Teststatistikken er gitt ved lign. 2.38.

$$T_\alpha = q_\alpha(a, f)\frac{MS_E}{n} \tag{2.38}$$

Hvor MS_E kommer fra lign. 2.35. Når $\hat{\Gamma}_{ij}$ er større enn T_α er forskjellen mellom forventningene signifikant med nivå α. $q_\alpha(a, f)$ angir den kritiske verdien for studentfordelingen med a grupper til sammenligning og f frihetsgrader.

3 Resultater

Beregninger på datasettene ble gjennomført ved bruk av R Studio og R Commander. R-koden som er benyttet er gitt i vedlegg B. Datasettet ble analysert med PCA og regresjonsmetoder som forlengs utvelgelse, PCR, PLS og CPLS. Alle resultatene er validert med enten kryssvalidering eller testsettvalidering. Metodevalidering ble gjennomført ved bruk av CVANOVA.

3.1 Datasett

I 2016 ble det gjennomført en masteroppgave i samarbeid med Fakultetet for kjemi, bioteknologi og matvitenskap (KBM) og Institutt for mattrygghet og infeksjonsbiologi (MatInf) (Aanrud 2016). Oppgaven analyserte 33 skjellprøver fra fire forskjellige arter og 14 forskjellige lokasjoner. Resultatet av masteroppgaven ble blant annet en 33×60 matrise med prøver av skjell mot fettsyreprofilene, art, lokasjon og toksisitet som trengte analyse. I denne matrisen består radene, beskrevet som n, av skjellprøvene mens kollonnene, beskrevet som p, består av informasjonen om prøvene. Datamatrisen er gitt i vedlegg C og ble bestemt ved bruk av GC-MS. FAMEne ble regnet om til en fettsyreprofil hvor arealet for hver topp viser hvor stor prosentandel av den totale mengden fettsyrer som hver fettsyre utgjør i hver av prøvene. Denne oppgaven består av behandling og analyse av disse dataene med vekt på prediksjon.

Dataene for FAME ble standardisert før PCR og PLS ved at hver variabel ble dividert med standardavviket sitt (tatt over kalibrasjons datasettet). Dette resulterer i at alle va-

riablene får forventning lik 0 og standardavvik lik 1. Dette svarer til å gi hver variabel den samme vekten, den samme viktigheten i analysen (Wold et al. 2001).

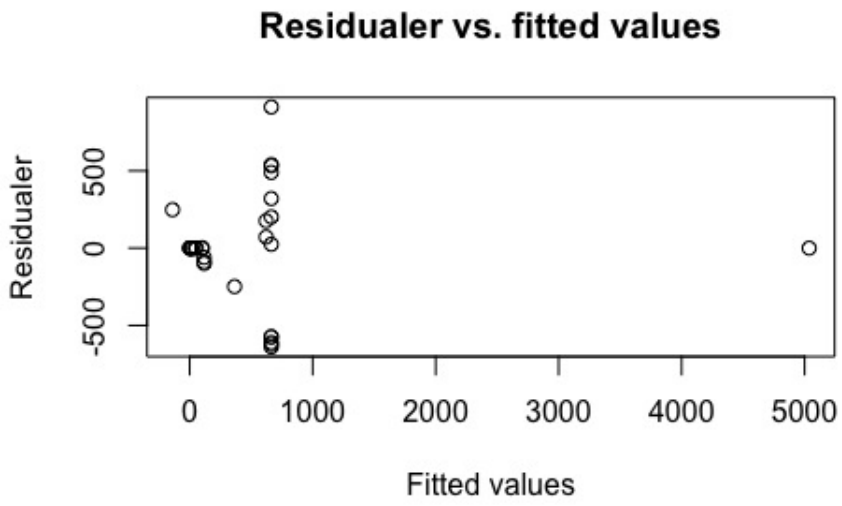
En av fettsyrene gav en FAME-profil som ikke var spesifikk for den fettsyren. Toppen bestod av C16:2 t9,12 og en annen fettsyre som ikke var mulig å skille fra transfettsyren. Målingene ble derfor vudert til å være null eller tilnærmet lik null, og C16:2 t9,12 er fjernet fra datasettet.

3.2 Variansanalyse

For å se om noen av forklaringsvariablene var signifikante ble en variansanalyse (ANOVA) gjennomført. Modellen er vist i lign. 3.1 med mengden toksiner som respons og art og sted som forklaringsvariabel. Dette er en to-faktor modell uten samspill.

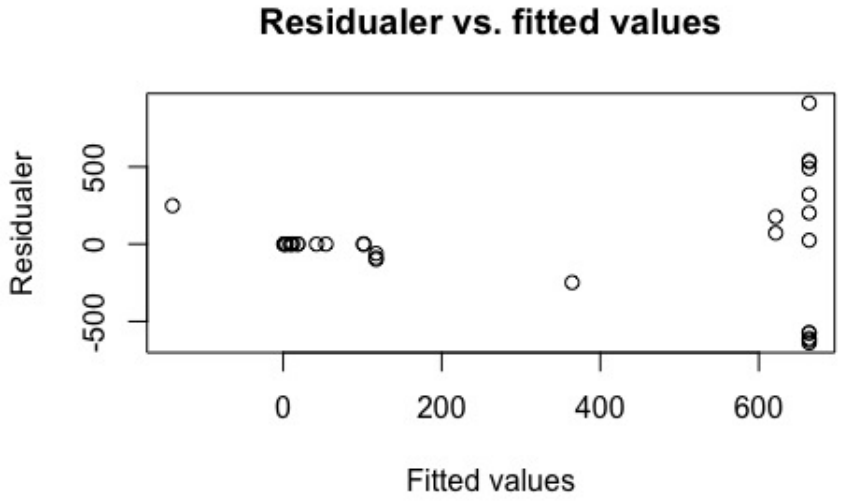
$$Total.toksin = Art + Sted \tag{3.1}$$

Den lineære modellen viste at et av stedene var signifikante med $p-verdi < 0.05$. Stedet som viste en signifikant p-verdi var Rundhaugen. For å se på spredningen i datapunktene ble residualene og og de tilpassede verdiene lagret i datasettet og plottet mot hverandre. Dette plottet er vist i figur 3.1.



Figur 3.1: Residualer plottet mot tilpassede verdier for modellen gitt i lign. 3.1

Prøven med toksisitet over 5000 ble utfra datasettet identifisert som B-1443 Rundhaugen. Stedet er altså det samme som var signifikant i ANOVA-analysen. Denne prøven ble vurdert som en uteligger og fjernet fra datasettet. ANOVA ble gjennomført igjen på det nye reduserte datasettet med modell 3.1. Residualene og de tilpassede verdiene for det nye datasettet er plottet mot hverandre i figur 3.2.



Figur 3.2: Residualer plottet mot tilpassede verdier for modellen gitt i lign. 3.1 med uteliggeren B-1443 Rundhaugen fjernet

Figur 3.2 viser at det ikke er konstant varians i datasettet, men en gruppe med lav varians og en gruppe med høy varians.

En tabell med resultatene fra ANOVA-analysen er vist i tabell 3.1.

Tabell 3.1: ANOVA-tabell for modell 3.1 med uteliggeren B-1443 Rundhaugen fjernet

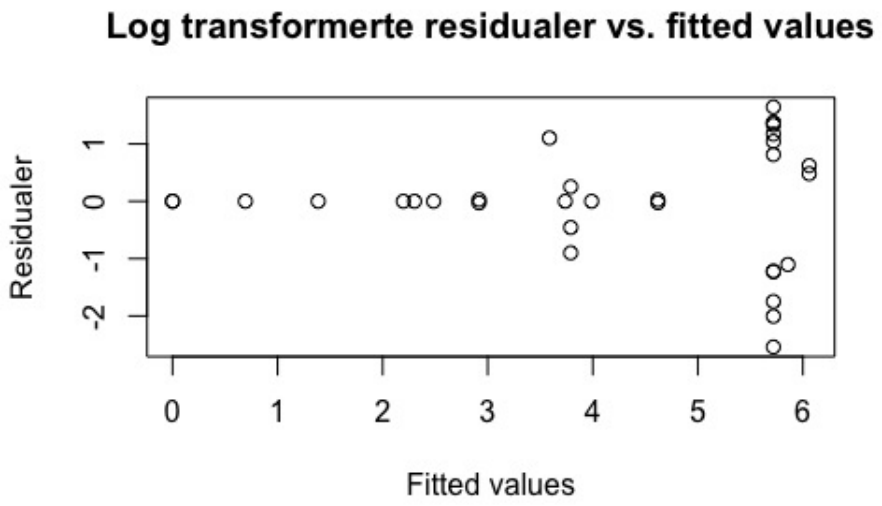
Faktor	Df	Sum Sq	Mean Sq	F value	P value
Sted	12	2617818	218152	0.9759	0.5056
Art	2	433432	216716	0.9694	0.3993
Residualer	17	3800287	223546	-	-

Tabell 3.1 viser at etter at uteliggeren ble fjernet er ikke sted signifikant lengre. Alle resultatene som er presentert etter denne analysen er basert på et datasett hvor uteliggeren B-1443 Rundhaugen er fjernet.

Det ble også gjennomført en ANOVA-analyse for en modell med logaritmisk transformasjon av responsen. Dette ble gjort for å se det kunne påvirke variansen. Modellen er vist i lign. 3.2.

$$log(Total.toksin) = Art + Sted \tag{3.2}$$

Residualene for modellen ble lagret og plottet mot de tilpassede verdiene vist i figur 3.3.

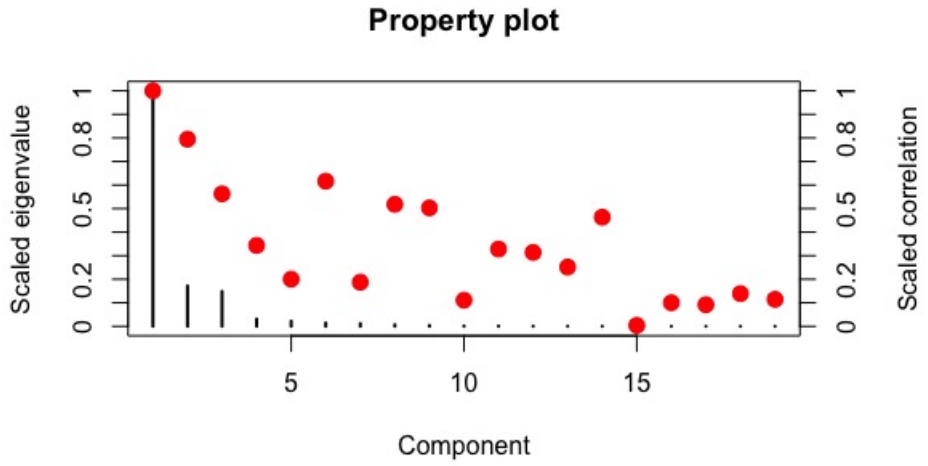


Figur 3.3: Residualer plottet mot tilpassede verdier for den logtransformerte modellen gitt i lign. 3.2

Figur 3.3 viser at variansen ikke er konstant i datasettet etter logtransformering av responsen.

3.3 Analyse av relevante komponenter

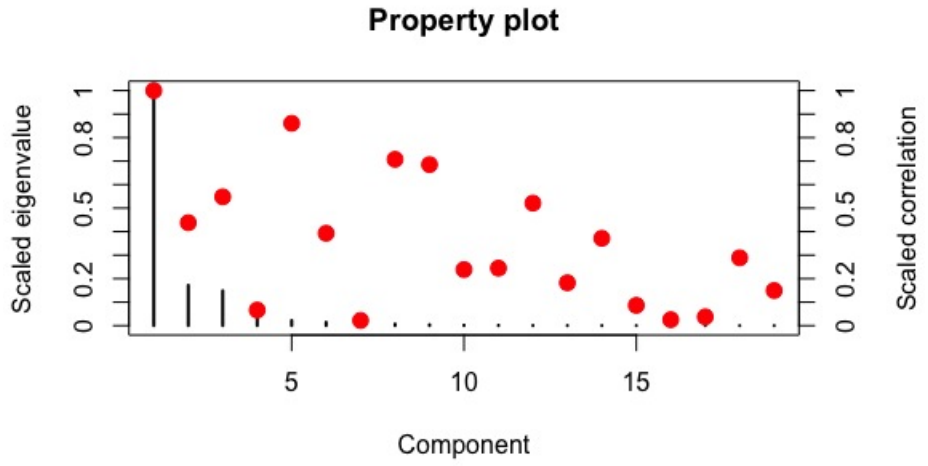
For å se hvilke prinsipalkomponenter som er høyest korrelert til responsen ble et egenskapsplot laget. Figur 3.4 viser egenskapsplottet for fettsyrene analysert med korrelasjonsmatrise.



Figur 3.4: Egenskapsplot hvor de svarte strekene er komponenter med tilsvarende egenverdier og de røde prikkene er korrelasjonen.

Figur 3.4 viser skalerte egenverdier til komponentene og korrelasjonen til responsen. Figuren viser at de første 5 komponentene har korrelasjon til responsen som synker gradvis før komponent 6 gjør et stort hopp opp igjen og får en høy korrelasjon til responsen.

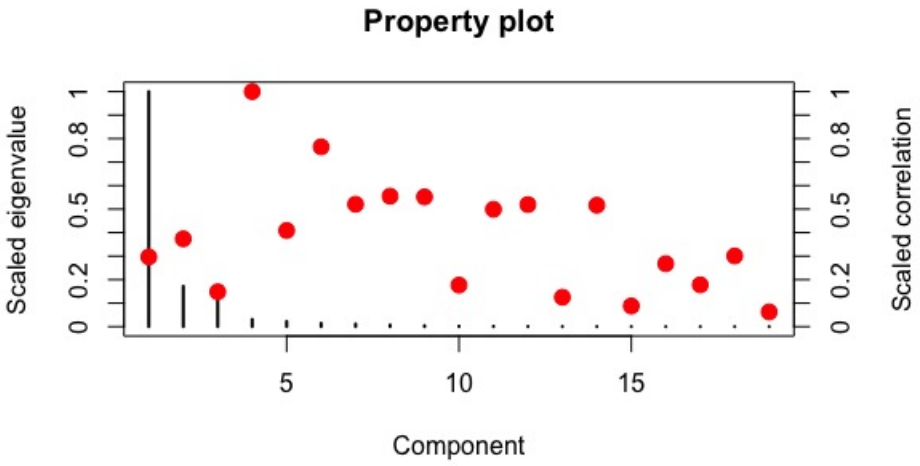
Et egenskapsplot ble også laget for logtransformeringen av responsen og er vist i figur 3.5.



Figur 3.5: Egenskapsplot for logaritmisk transformasjon av originaldatene hvor de svarte strekene er komponentene med tilsvarende egenverdier og de røde prikkene er korrelasjonen

Figur 3.5 viser større spredning i størrelsen på korrelasjonen til responsen etter log-transformering. I denne figuren har komponenter med små egenverdier enda høyere korrelasjon til responsen enn det figur 3.4 med originaldatene viser.

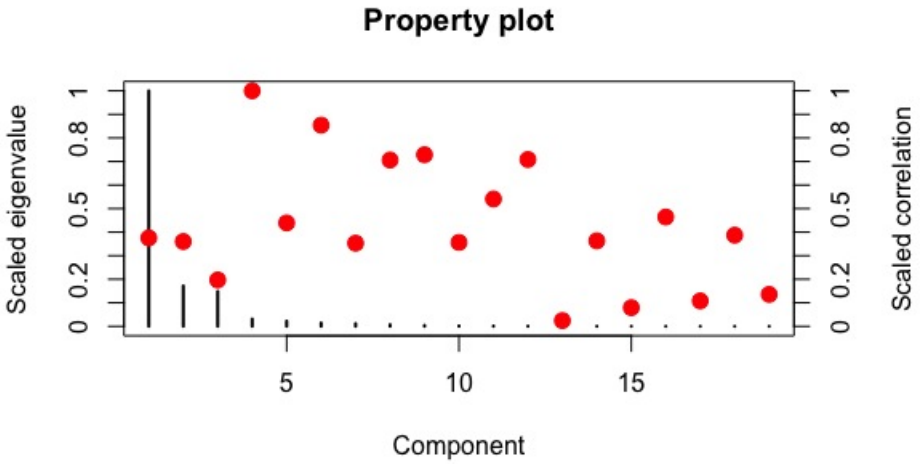
Et egenskapsplot for residualene fra modell 3.1 ble laget og er vist i figur 3.6.



Figur 3.6: Egenskapsplot med korrelasjonsmatrise og residualer hvor de svarte strekene er komponentene med tilsvarende egenverdier og de røde prikkene er korrelasjonen.

Figur 3.6 viser en lavere korrelasjon til responsen for komponent 1, 2 og 3. Dette kan tyde på at dette er komponenter som inneholder informasjon om art og sted. Siden responsen i denne modellen er residualene korrigeres det for de to faktorene og det kan føre til denne endringen.

Et egenskapsplot for modellen med residualene fra logtransformeringen av toksinkonsentrasjonen ble laget og er vist i figur 3.7.



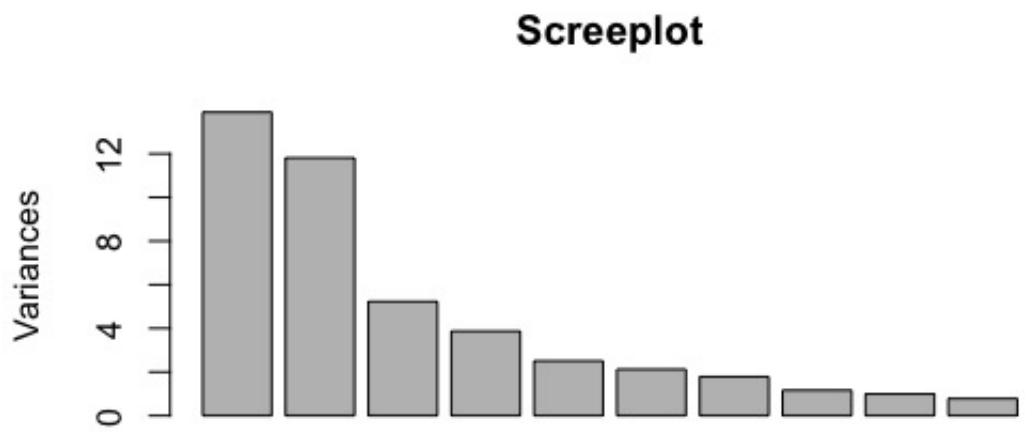
Figur 3.7: Egenskapsplot for residualer etter logaritmisk transformasjon av responsen hvor de svarte strekene er komponentene med tilsvarende egenverdier og de røde prikkene er korrelasjonen

Figur 3.7 viser lav korrelasjon for komponent 1, 2 og 3 som figur 3.6. Dette plottet viser mange komponenter som har små egenverdier og høy korrelasjon til responsen.

Alle fire plottene av relevante komponenter viser flere komponenter med små egenverdier som har høy korrelasjon til responsen, et problem som gjør at prediksjonsmodellen kan bli dårlig.

3.4 Prinsipalkomponentanalyse

Det ble utført PCA på fettsyrene for å se på sammenhenger i datasettet. For å illustrere hvor mye variasjon som forklares av de i første komponentene lages et screeplot. Screeplottet er gitt i figur 3.8 og viser hvordan egenverdiene til datasettet vil se ut.

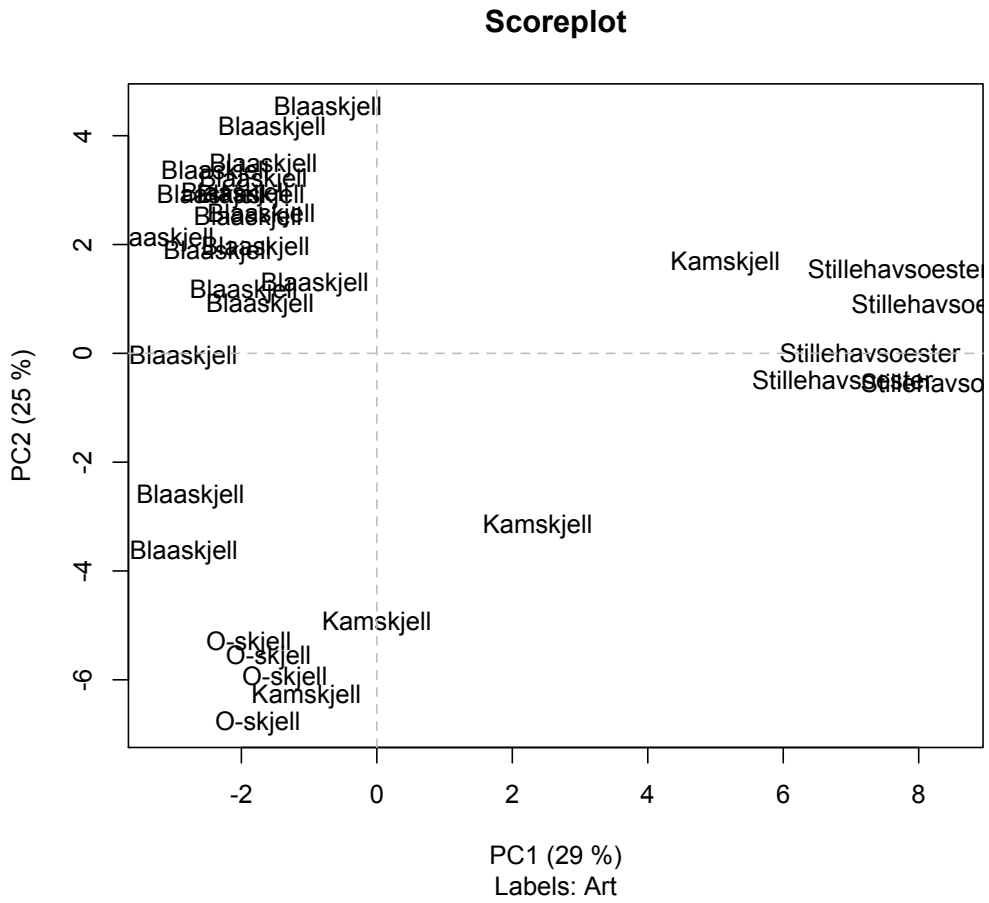


Figur 3.8: Screeplot av egenverdiene fra PCA

Figur 3.8 viser at de to første egenverdiene inneholder mye av variansen. De to første egenverdiene står for 53,56% av den totale variansen.

3.4.1 Scoreplot og ladningsplot

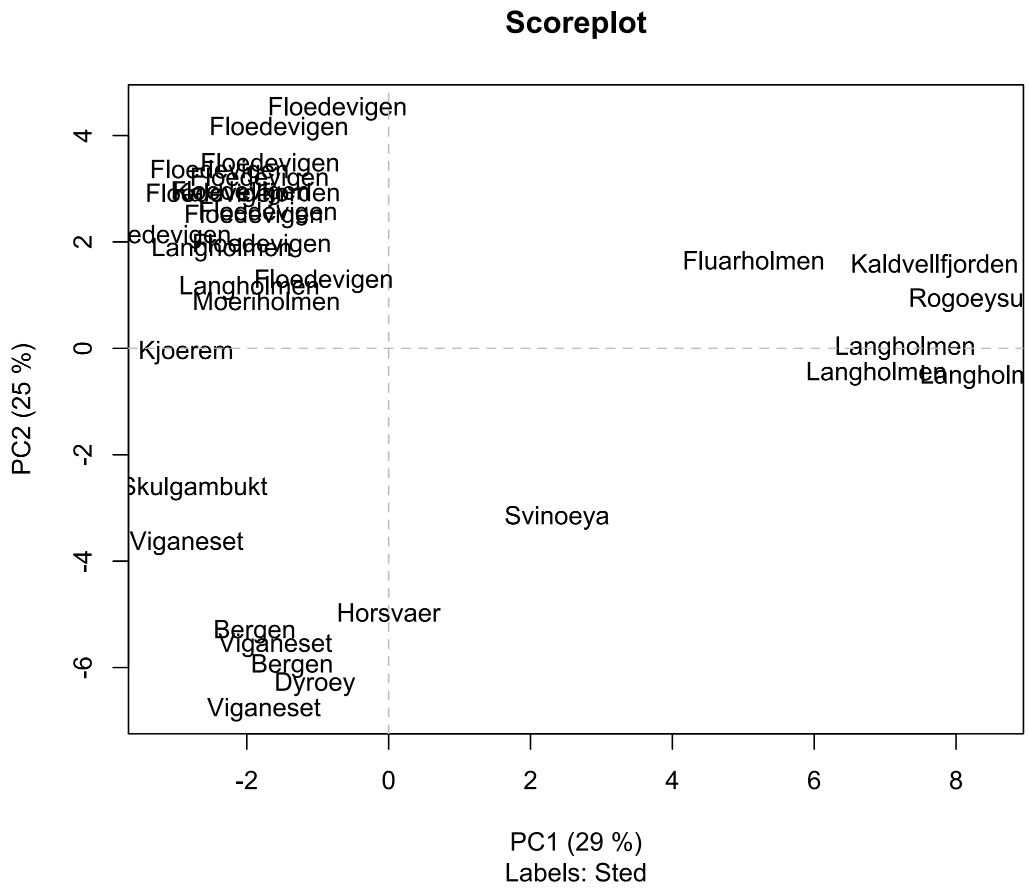
For å se på spredningen i dataene ble det laget et scoreplot for prøvene etter PCA. Plottet er vist i figur 3.9 og viser prinsipalkomponent 1 (PC1) mot prinsipalkomponent 2 (PC2) basert på korrelasjonsmatrisen til datasettet med art som grupperingsnavn.



Figur 3.9: Scoreplot som viser grupperinger i datasettet etter PCA merket etter art

Plottet i figur 3.9 viser tydelig gruppering mellom de forskjellige artene.

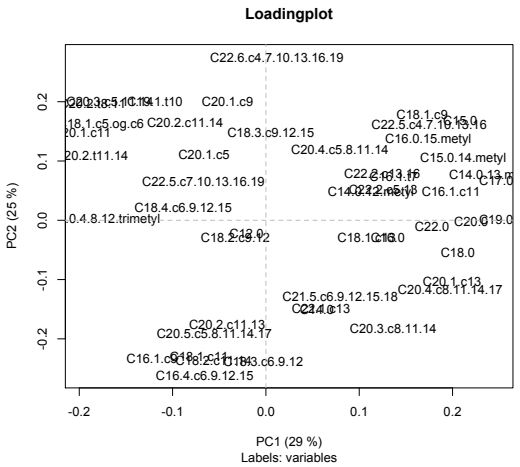
Det ble også laget et scoreplot for prøvene med sted som grupperingsnavn. Dette plottet er vist i figur 3.10.



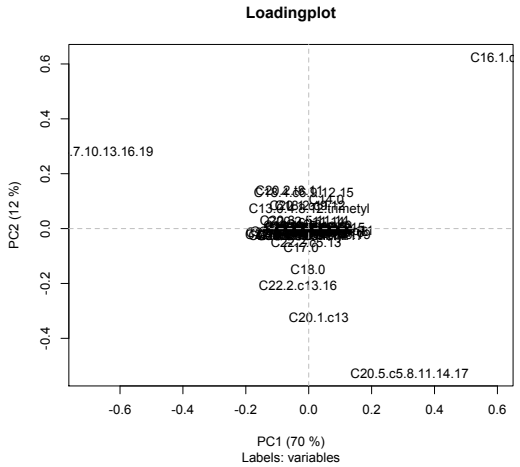
Figur 3.10: Scoreplot som viser grupperinger i datasettet etter PCA merket med prøvetakings sted

Plottet i figur 3.10 viser gruppering mellom noen av stedene. Få av prøvene kommer fra samme sted som gjør at det ikke er så mye informasjon å hente fra dette plottet.

Det ble også laget ladningsplot for prøvene. PCA kan utføres både på korrelasjonsmatrisen og kovariansmatrisen til dataene. Ladningsplottet for korrelasjonsmatrisen en vist i figur 3.11 mens ladningsplottet for kovariansmatrisen er vist i figur 3.12.



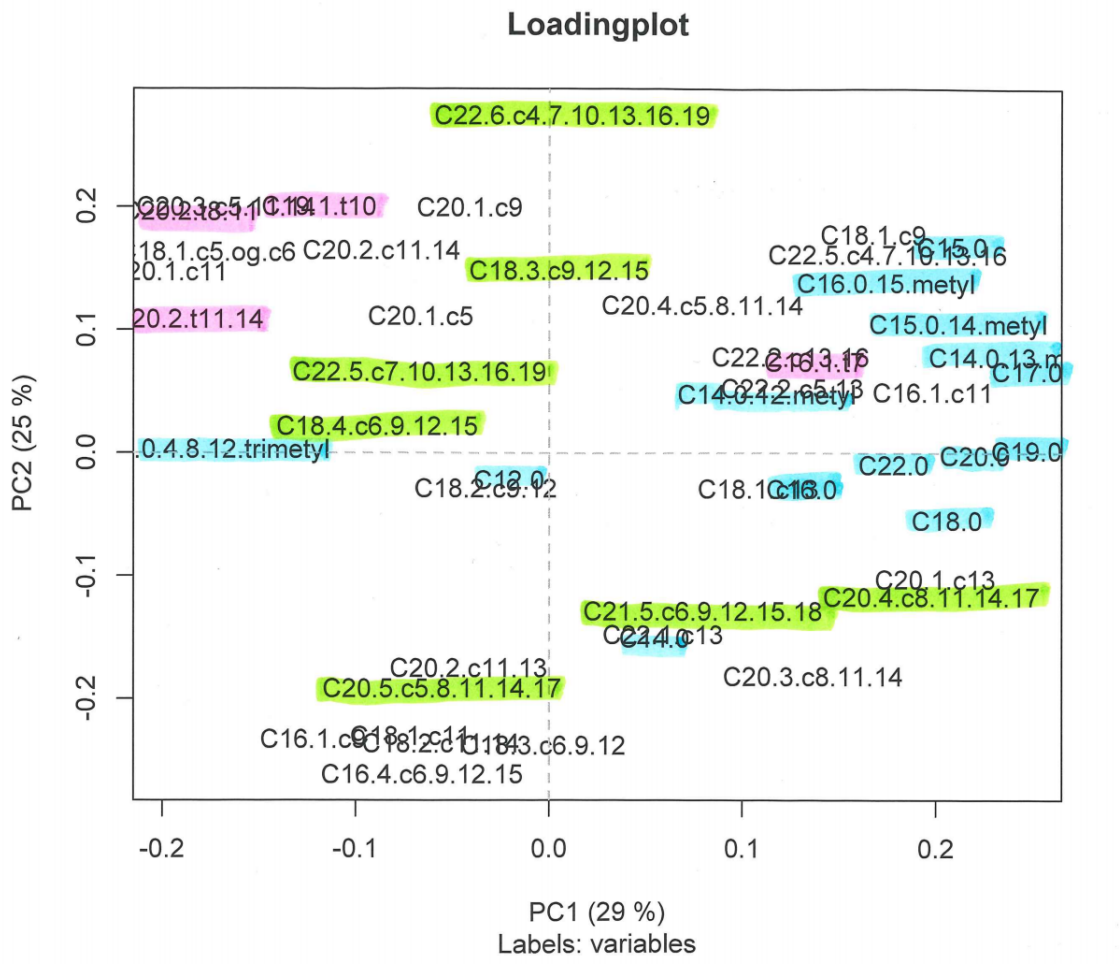
Figur 3.11: Ladningsplot av fettsyrer basert på korrelasjonsmatrisen



Figur 3.12: Ladningsplot for fettsyrer basert på kovariansmatrisen

Med korrelasjonsmatrisen blir 29% forklart av PC1 mens 25% blir forklart av PC2. Med kovariansmatrisen blir 70% forklart av PC1 mens 12% blir forklart av PC2. For kovariansmatrisen ender alle fettsyrene opp i midten av plottet som gjør det vanskelig å se forskjeller og grupperinger. Med bakgrunn fra disse ladningsplottene er det valgt å bruke korrelasjonsmatrisen til analyse av datasettet i denne oppgaven.

Figur 3.13 viser ladningsplottet for korrelasjonsmatrisen til fettsyrene med fargeinndeling av omega-3 fettsyrer i grønn, mettede fettsyrer i blå og trans-fettsyrer i rosa. Fettsyrene som ikke er merket er cis-fettsyrer.



Figur 3.13: Ladningsplot for kovariansmatrisen med fargeinndeling av omega-3 fettsyrer i grønn, mettedet fettsyrer i blå og trans-fettsyrer i rosa. Resten er cis-fettsyrer

Figur 3.13 viser to grupperinger av omega-3 fettsyrer, en gruppering av mettede fettsyrer og tre samlede trans-fettsyrer med en som ligger lengre unna. De resterende cis-fettsyrene ligger jevnt spredt utover plottet.

3.5 *Estimering av nedre grense for prediksjon*

Nedre grense for prediksjon angir hvor langt ned RMSEP-verdiene kan komme som vist i kapittel 2.7.1. Å kunne estimere denne grensen kan gi en indikasjon på hvor god metoden er. Dersom laveste RMSEP for metoden ligger nært den nedre grensen for prediksjon anses metoden for å gi god prediksjon. Bayes-PLS (Helland et al. 2012) er

brukt til å estimere σ^2, altså nedre grense for prediksjon. Resultatene er oppgitt som $\sqrt{\sigma^2}$ ettersom at denne verdien sammenlignes med RMSEP-verdiene som er tatt kvadratrot av.

Når det originale datasettet ble kjørt ble σ^2 estimert til å være 25402.42. Da blir σ estimert til å være 159.389. Bayes-PLS ble også kjørt for modellen basert på residualene fra modell 3.1 og σ^2 ble estimert til å være 19711.53. Da estimeres σ til 140.398.

3.6 Regresjonsanalyse

For å finne den beste mulige metoden for prediksjon ble det utført regresjonsanalyser med flere forskjellige metoder. Alle metodene ble kjørt med standardisering av variablene og er validert ved bruk av testsettvalidering eller leave-one-out kryssvalidering.

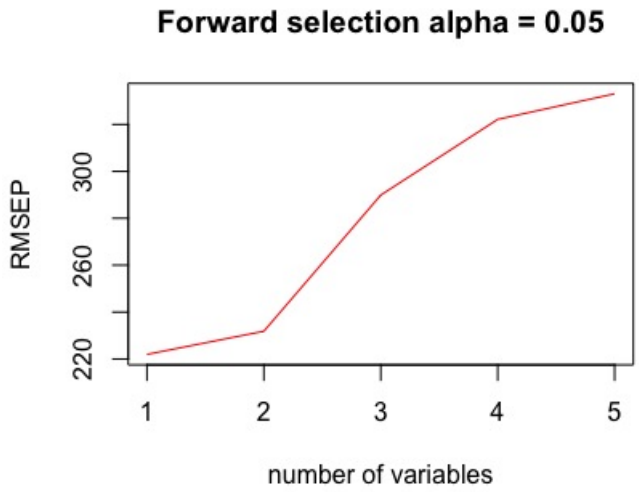
3.6.1 Nullmetoden

Gjennomsnittet ($\bar{y}$) for toksisiteten i skjell er beregnet til å være 317.38. RMSEP-verdien ble beregnet som 477.6 og RMSEP-verdien for analyse av residualer ble 355.7. Disse gir nullmetodene for de respektive metodene og blir de beste prediksjonsmodellene dersom ingen metoder gir en lavere RMSEP-verdi.

3.6.2 Forlengs utvelgelse

Variabelseleksjon i form av forlengs utvelgelse ble gjennomført for å finne en metoden med lavest mulig antall forklaringsvariabler som gir god prediksjon.

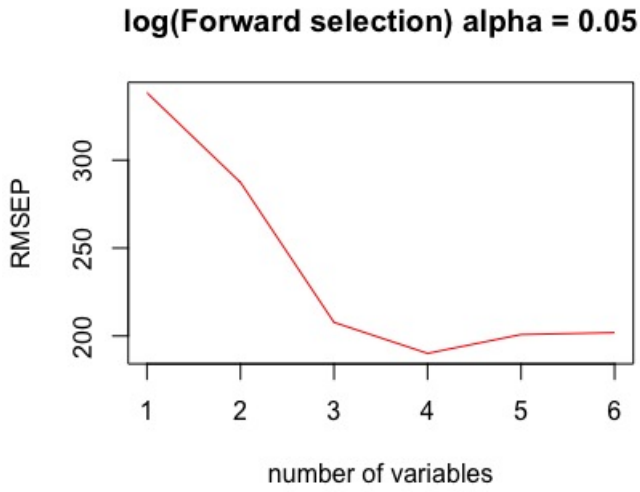
Metoden ble validert med testsettvalidering hvor datasettet ble delt opp i et kalibreringsett og et testsett. Kalibreringsettet brukes til konstruksjon av metoden, mens testsettet ble holdt utenfor. Metoden tilpasses med testsettet og RMSEP-verdien beregnes. Forlengs utvelgelse med alfa 0.05 er vist i figur 3.14.



Figur 3.14: Forlengs utvelgelse med alfa 0.05

Figur 3.14 viser at laveste RMSEP-verdi ble bestemt til å være 221.95 med 1 variabel.

Forlengs utvelgelse for logtransformert respons med alfa 0.05 er vist i figur 3.15.

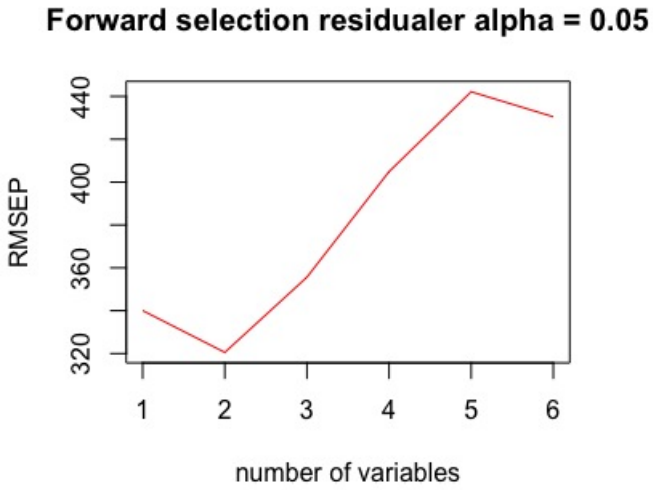


Figur 3.15: Forlengs utvelgelse med logtransformert respons og alfa 0.05

Figur 3.15 viser at laveste RMSEP-verdi ble bestemt til å være 190.11 med 4 variabler.

3.6.3 *Forlengs utvelgelse med analyse av residualer*

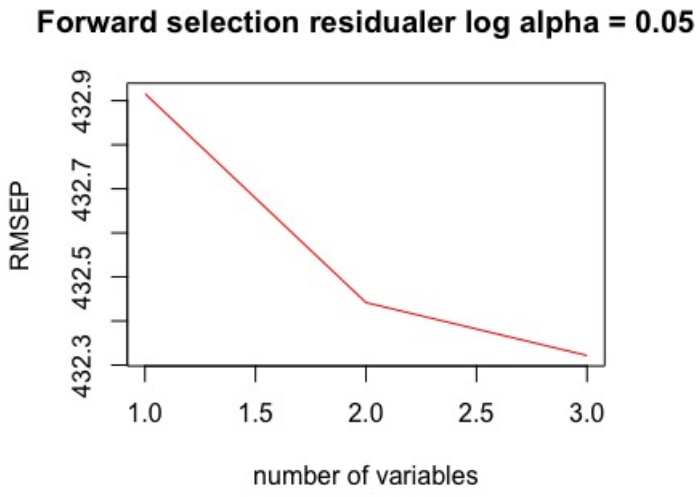
Forlengs utvelgelse på residualene ble også kjørt med $\alpha = 0.05$ og resultatene er vist i figur 3.16.



Figur 3.16: Forlengs utvelgelse av residualer med alfa 0.05

Figur 3.16 viser at laveste RMSEP-verdi ble bestemt til å være 320.55 med 2 komponenter.

Utvelgelsen ble også kjørt for residualer fra den logtrasformerte responsen fra ANOVA med $\alpha = 0.05$ og resultatet er vist i figur 3.17.



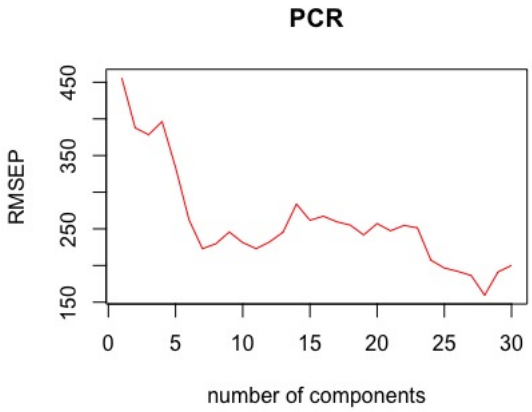
Figur 3.17: Forlengs utvelgelse på residualer med logtransformert respons og alfa 0.05

Figur 3.17 viser at laveste RMSEP-verdi ble bestemt til å være 432.32 med 3 komponenter.

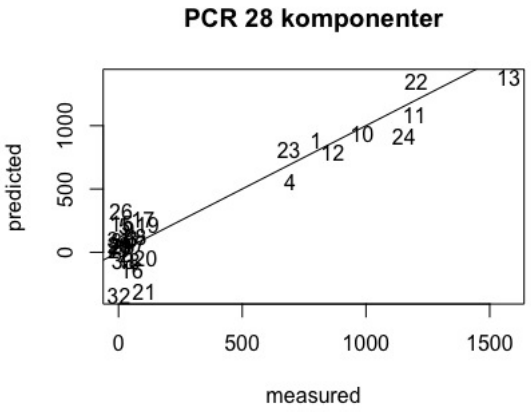
Forlengs utvelgelse ble i tillegg gjennomført med $\alpha = 0.20$. For den metoden ble 17-18 variabler holdt igjen. Dette α-nivået ble vurdert for høyt til å presenteres i oppgaven.

3.6.4 Prinsipal komponent regresjon

PCR ble utført på fettsyreprofilene og RMSEP-verdiene ble plottet mot antall komponenter for å se hvor mange komponenter som gav den beste metoden. Figur 3.18 viser RMSEP-plottet for PCR. Figur 3.19 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



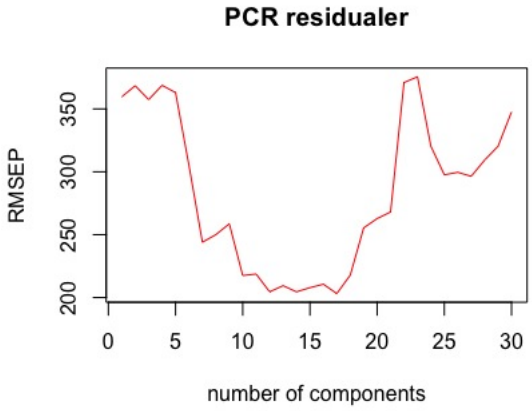
Figur 3.18: RMSEP-verdier for PCR



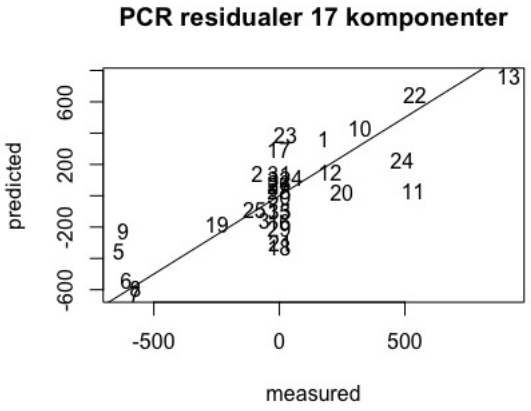
Figur 3.19: PCR 28 komponenter predikert mot sann verdi med trendlinje

Figur 3.18 viser at laveste RMSEP for denne metoden ble 159.40 på 28 komponenter. Figur 3.19 viser en gruppering av verdier rundt null og en trendlinje som passer til datapunktene. Noen skjell predikeres som negative, men generelt blir giftige skjell predikert som giftige.

Figur 3.20 viser RMSEP-plottet for PCR med residualene fra metoden i lign. 3.1. Figur 3.19 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



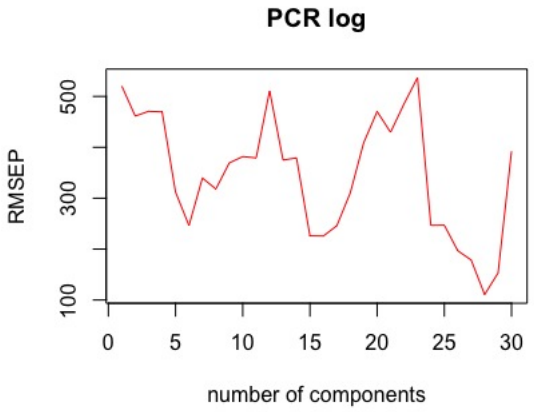
Figur 3.20: RMSEP-verdier for PCR med residualer fra modell 3.1 som respons



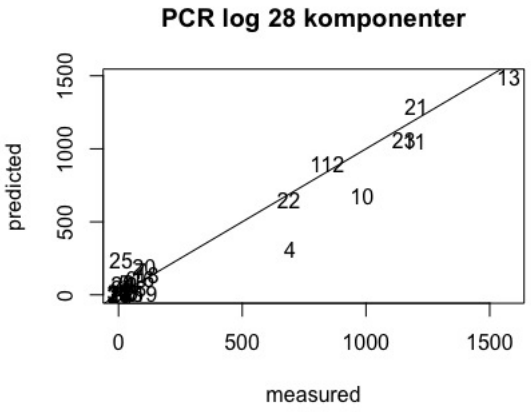
Figur 3.21: PCR residualer 17 komponenter predikert mot sann verdi med trendlinje

Figur 3.20 viser at laveste RMSEP for denne metoden ble 203.20 på 17 komponenter. Figur 3.21 viser god spredning i verdiene, men dette plottet viser større avstand mellom punktene enn plottet for den originale responsen.

Figur 3.22 viser RMSEP-plottet for PCR utført på logtransformert respons. Figur 3.23 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



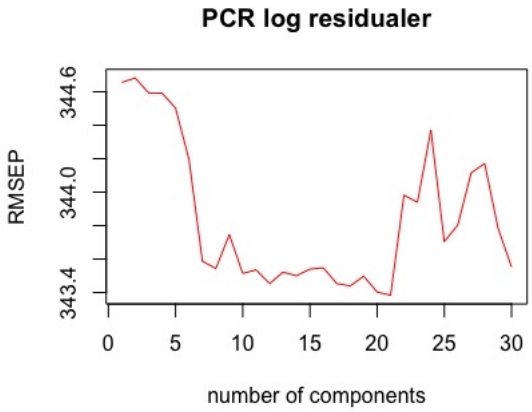
Figur 3.22: RMSEP-verdier for PCR med logtransformert respons



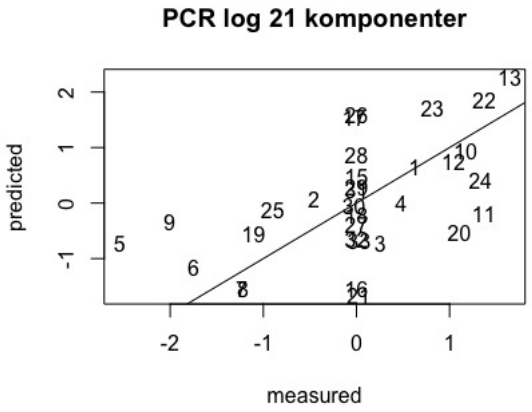
Figur 3.23: PCR logtransformert 28 komponenter predikert mot sann verdi med trendlinje

Figur 3.22 viser at laveste RMSEP for denne metoden ble 110.68 på 28 komponenter. Figur 3.23 viser mange variabler som ligger rundt null og færre variabler med høye verdier. Flere verdier ligger et stykke unna trendlinjen.

Figur 3.24 viser RMSEP-plottet for PCR med residualene fra den logtransformerte modellen gitt i lign. 3.2. Figur 3.19 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



Figur 3.24: RMSEP-verdier for PCR med logtransformerte residualer fra modell 3.2



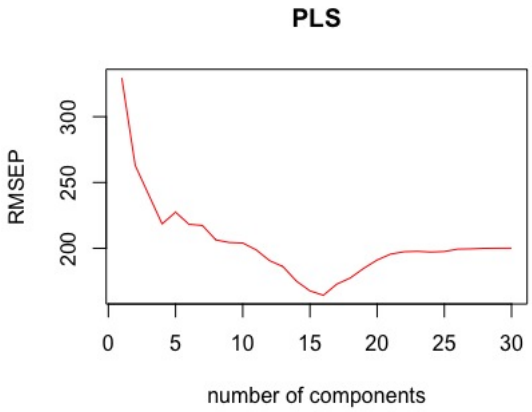
Figur 3.25: PCR logtransformerte residualer 17 komponenter predikert mot sann verdi

Figur 3.24 visert at laveste RMSEP for denne metoden ble 559.07 på 21 komponenter. Figur 3.25 viser en veldig skjev og dårlig fordeling mellom predikerte og målte verdier som ikke passer til trendlinjen.

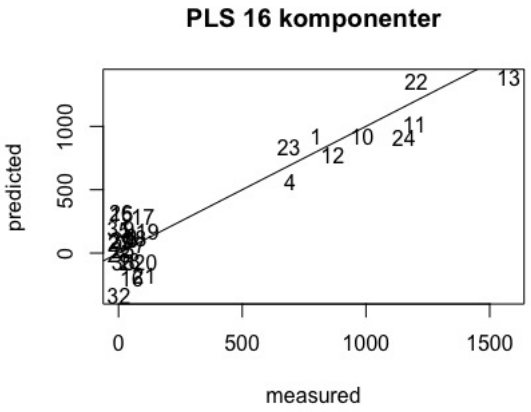
3.6.5 Partial Least Square Regresjon

PLS ble utfør på toksisiteten til skjellene mot fettsyrene. RMSEP-verdiene ble plottet mot antall komponenter for å se hvor mange komponenter som gav den beste metoden.

Figur 3.26 viser RMSEP-plottet for PLS. Figur 3.27 viser sammenhengen mellom den predikerte responsen og den målte responsen.



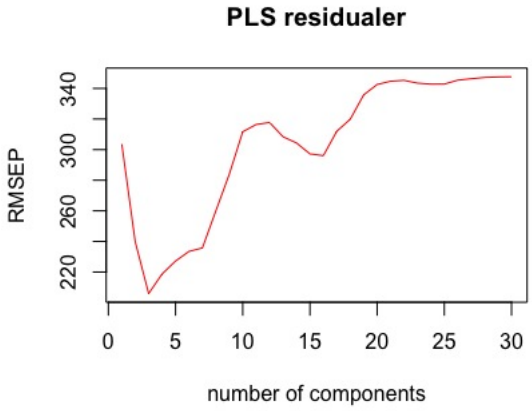
Figur 3.26: RMSEP-verdier for PLS



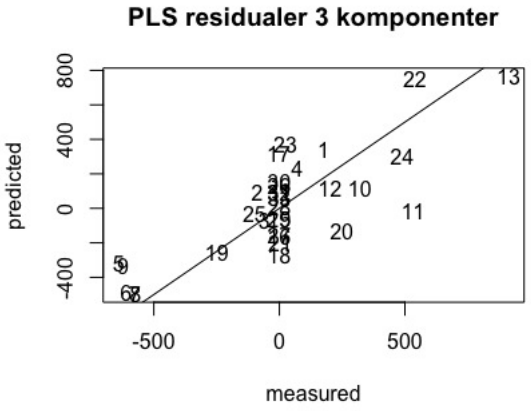
Figur 3.27: PLS 16 komponenter predikert mot sann verdi med trendlinje

For figur 3.26 ble laveste RMSEP-verdi bestemt til å være 164.3 på 16 komponenter. Figur 3.27 viser en gruppering rundt null, med en trendlinje som passer til de resterede datapunktene.

Figur 3.28 viser RMSEP-plottet for PLS med redsidualene til modell 3.1. Figur 3.29 viser sammenhengen mellom den predikerte responsen og den målte responsen for modellen.



Figur 3.28: RMSEP-verdier for PLS med residualer fra modell 3.1 som respons

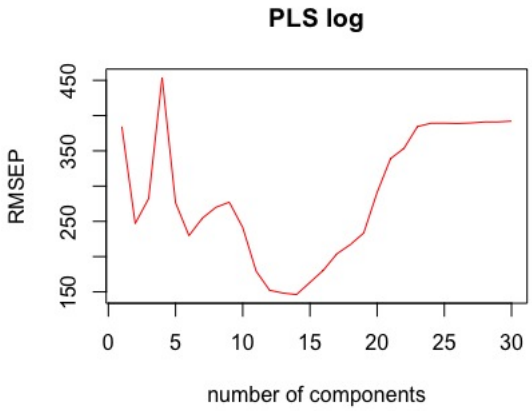


Figur 3.29: PLS for residualer predikert mot sann verdi med trendlinje

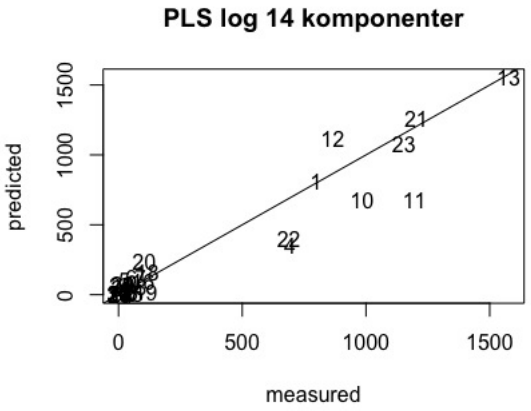
For figur 3.28 ble laveste RMSEP-verdi bestemt til å være 205.9 på 3 komponenter. Figur 3.29 viser noen negative verdier og en skjev gruppering av verdier. Verdiene viker også

en del fra trendlinjen.

Figur 3.30 viser RMSEP-plottet for PLS utført på logtransformert respons. Figur 3.31 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



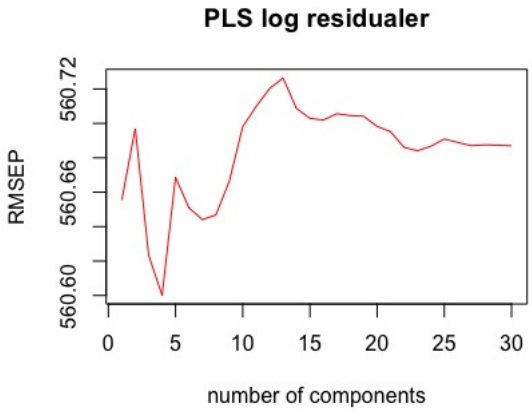
Figur 3.30: RMSEP-verdier for PLS med logtransformert respons



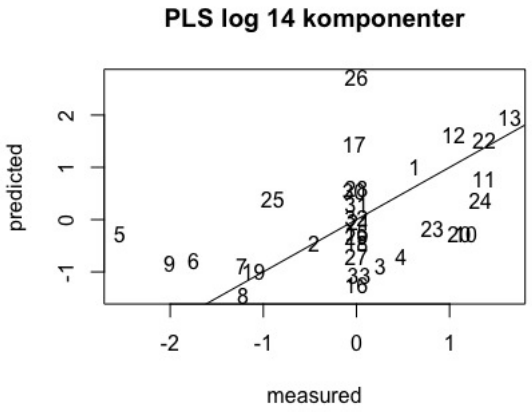
Figur 3.31: PLS logtransformasjon 14 komponenter predikert mot sann verdi med trendlinje

Figur 3.30 viser at laveste RMSEP-verdi ble bestemt til å være 146.14 på 14 komponenter. Figur 3.31 viser en gruppering rundt null og noen verdier som viker fra trendlinjen.

Figur 3.32 viser RMSEP-plottet for PCR med residualene fra den logtransformerte modellen gitt i lign. 3.2. Figur 3.33 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



Figur 3.32: RMSEP-verdier for PLS med logtransformerte residualer fra modell 3.2 som respons

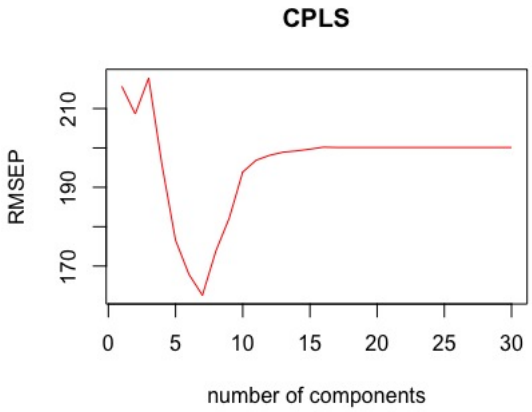


Figur 3.33: PLS for logtransformasjon 3 komponenter predikert mot sann verdi med trendlinje

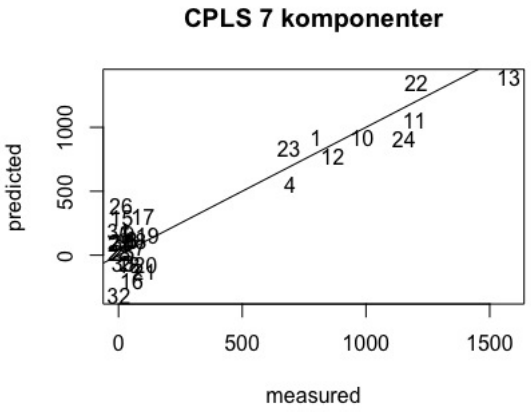
Figur 3.32 viser at laveste RMSEP-verdi ble bestemt til å være 559.11 på 4 komponenter. Figur 3.33 viser en gruppering rundt null og en trendlinje som ikke passer til datapunktene.

3.6.6 Kanonisk powered PLS

CPLS ble kjørt som total toksin mot fettsyrene med art og sted som tilleggsrespons. Figur 3.34 viser RMSEP-plottet for CPLS. Figur 3.35 viser sammenhengen mellom den predikerte responsen og den målte responsen for metoden.



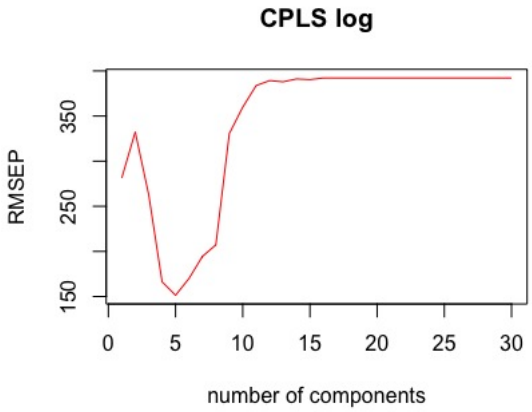
Figur 3.34: RMSEP-verdier for CPLS



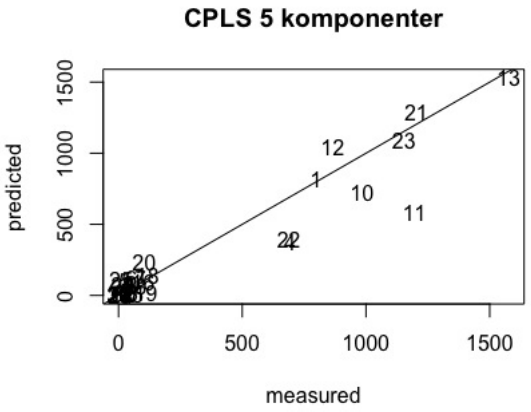
Figur 3.35: CPLS predikert mot sann verdi med trendlinje

Figur 3.34 viser at laveste RMSEP-verdi ble bestemt til å være 162.6 på 7 komponenter. Figur 3.35 viser en gruppering rundt null og en trendlinje som passer med datapunktene. Grupperingen rundt null viser større spredning enn tidligere plot.

CPLS ble også kjørt med logaritmisk transformering av responsen. Figur 3.36 viser RMSEP-plottet for metoden med logtransformering. Figur 3.37 viser y mot $\hat{y}$ på beste antall komponenter etter transformeringen.



Figur 3.36: RMSEP-verider for CPLS med logtransformert respons



Figur 3.37: CPLS log predikert mot sann verdi

Figur 3.36 viser at laveste RMSEP-verdi ble bestemt til å være 151.37 på 5 komponenter.

Figur 3.37 viser en gruppering rundt null og noen datapunkter som viker fra trendlinjen.

3.7 Metodevalidering

Den beste metoden fra PCR, PLS og CPLS ble valgt ut og forskjellen mellom predikert og sann respons (gitt som MESP) ble hentet ut for det optimale antall komponenter for hver av metodene. Modellen som brukes er gitt i 3.3.

$$z_{ij} = \mu + A_i + \tau_j + \epsilon_{ij} \tag{3.3}$$

Dette er en blandet modell hvor A angir de $i = 13$ skjellene som er en tilfeldig effekt. τ angir de $j = 3$ metodene som er en fast effekt. z_{ij} er gitt som MSEP-verdiene og ϵ_{ij} er feilleddet.

Hypotesene som testes for denne modellen er gitt i 3.4 og tester om gjennomsnittene for hver metode er like.

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \tau_3 \\ &H_1 : \text{Minst to } \tau_i \text{ er ulike} \end{aligned} \tag{3.4}$$

Modellen gav skjell en p-verdi på 0.3436 og metode en p-verdi på 0.0189. P-verdien for metode er lavere enn 5% og dermed er metode en signifikant effekt. Nullhypotesen kan dermed forkastes og det er signifikant forskjell mellom metodene.

Siden metode er en signifikant effekt er det interessant å se hvilke metoder som skiller seg ut. Derfor ble det gjennomført en Tukey par-vis test for å se på kontrastene mellom metodene. Resultatene for Tukey testen er gitt i tabell 3.2.

Tabell 3.2: Tukey par-vis test av kontraster mellom de tre beste metodene

Hypotese	Nedre	Senter	Øvre	St.avvik	t-verdi	p-verdi
CPLS - PCR	-140014	-72438	-4862	28142	-2.574	0.0329
CPLS - PLS	-70689	-3113	64463	28142	-0.111	0.9933
PCR - PLS	1749	69325	136901	28142	2.463	0.0431

Tabell 3.2 viser med signifikansnivå $\alpha = 0.05$ at det er forskjell mellom CPLS og PCR, PLS og PCR, men ikke mellom CPLS og PLS.

4 Diskusjon

Vanlige kjemometriske metoder som PCR, PLS, CPLS og variabelseleksjon ble brukt til å undersøke forskjellige måter å konstruere en prediksjonsmodell. En metode ble funnet som forhåpentligvis kan brukes til å predikere toksisiteten i skjell basert på fettsyresammensetningen.

I denne oppgaven er forklaringsvariablene ($\mathbf{X}$) ikke gjort noen endringer med. Responsen ($\mathbf{y}$) er endret for å forsterke forholdet mellom fettsyreprofilene og toksisiteten. Responsen er analysert som originaldata, logtransformert, residualer fra ANOVA-modell og logtransformerte residualer.

4.1 Generelle kommentarer om datasettet

4.1.1 Konsekvensen av $n < p$

Siden antallet prøver (n) i datasettet er mindre enn antallet forklaringsvariabler (p) har ikke $\mathbf{X}^T\mathbf{X}$ full rang. Estimatoren for minste kvadrater er gitt ved $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Denne estimatoren inneholder et ledd ($\mathbf{X}^T\mathbf{X}$) som skal inverteres. Når matrisen ikke har full rang kan den ikke inverteres. Det betyr at minste kvadrater ikke kan brukes for analyse av datasettet i denne oppgaven. Minste kvadrater er den eneste regresjonsmetoden som gir forventningsrette estimater. Siden denne ikke kan brukes blir alle estimatene forventningsskjeve.

Når $n < p$ oppstår det veldig ofte kollinearitet i forklaringsvariablene. Dette kan bety at en eller flere av variablene er avhengig av andre variabler. Når responsen predikeres vil den bli påvirket av de kollineære variablene. Prediksjonsmodellen kan da bli upresis og koeffisientene får høye standardfeiler.

4.1.2 Analyse av residualer

Det er tidligere vist at mengden toksisitet i et skjell avhenger av sted og art. ANOVA antydet også dette. Derfor er en del av metodene basert på residualene fra denne analysen. I metodene basert på residualer brukes informasjon gitt av faktorene art og sted. Metodene er avhengig av art og sted og kan dermed ikke brukes på andre steder eller andre arter. Dersom det i fremtiden da blir hentet inn et skjell fra et annet sted vil ikke prediksjonsmodellen være gjeldende for denne prøven. Det samme gjelder dersom en annen art blir analysert. Dette fører til en svakhet i modellen som hindrer den fra å være universal og bruksnyttig for fremtidige analyser. Med bakgrunn i dette er det ikke ønskelig at metodene basert på residualene skal ha den beste prediksjonsevnen.

Metodene basert på residualer gjorde det dårligst for alle regresjonsmetodene og blir derfor ikke diskutert i stor grad videre i denne oppgaven.

4.1.3 Uteligger deteksjon

Uteligger deteksjon er utført basert kun på visuelle aspekter sett ut fra plot av residualer mot tilpassede verdier og store forskjeller i toksinkonsentrasjoner i responsen. En prøve ble oppdaget med av toksininnhold i størrelsesorden fem ganger de andre prøvene. Dette er prøven som skiller seg ut i figur 3.1. Prøven ble identifisert som B-1443 Rundhaugen og definert som en uteligger før den ble tatt ut av datasettet.

B-1443 Rundhaugen er en prøve av et blåskjell tatt uke 43 i 2014 fra Rundhaugen. Dette er den eneste prøven tatt fra dette stedet. Prøven er et eksempel på en ekstrem algeoppblomstring. Denne typen algeoppblomstring er ikke vanlig, dermed blir dette en prøve

som ikke er representativ for resten av populasjonen. Beholdes denne i datasettet når metoden lages vil det føre til at prediksjonsmodellen konsekvent vil overestimere toksisiteten i de nye prøvene.

Et problem med for kritisk seleksjon av uteliggere kan bli at prøver som ikke er uteliggere tas ut. Det resulterende datasettet kan mangle viktig informasjon om prøvene eller bli tvunget til å ukorrekt passe en normalfordeling.

4.1.4 Relevante komponenter

Når n er mindre enn p blir kun de første n komponentene brukt i metoden. Egenverdiene for komponentene som holdes utenfor vil være små, men om disse komponentene har korrelasjon til responsen er vanskelig å vite. Dersom noen av komponentene har høy korrelasjon til responsen kan relevant informasjon forsvinne. Dette gjør prediksjon av nye variabler mye vanskeligere. Egenskapsplottene i figurene 3.4-3.7 viser en antydning til dette.

Disse egenskapsplottene viser de skalerte egenverdiene til komponentene og korrelasjonen mellom egenvektoren og responsen. Plottene sier noe om hvor mange komponenter som trengs for å forklare responsen. Plottet i figur 3.4 viser egenskapsplottet til originaldataene og her er de første komponentene med store egenverdier i stor grad korrelert til responsen. Plottet i figur 3.5 viser logaritmisk transformasjon av orgiginaldataene og viser at komponent 5 er mer korrelert til responsen enn komponent 2, 3 og 4. Plottet i figur 3.6 er egenskapsplottet til residualene fra modell 3.1 og viser at de tre første komponentene er lavere korrelert til responsen enn komponent 4. Plottet i figur 3.7 er egenskapsplottet til residualene fra den logtransformerte modellen 3.2 og viser lav korrelasjon for de første tre komponentene. Alle de fire plottene viser flere komponenter med små egenverdier som er høyt korrelert til responsen.

Ut fra disse fire plottene tyder det på at prediksjonsmodellene vil ha en dårlig prediksjonsevne. Helland & Almøy 1994 viste at prediksjonsmodeller basert på komponenter med små egenverdier som har høy korrelasjon til responsen resulterer i høy predik-

sjonsfeil. Utfra disse plottene vil det se ut som den beste prediksjonsmodellen vil være basert på originaldataene. Et annet alternativ er at denne store spredningen i plottene kun kommer av støy i datasettet og da vil ikke disse plottene gi mye informasjon om hvordan metodene vil prestere.

Plottene viser forskjell i korrelasjonen for originaldataene og for residualene. Dette kan tyde på at de første tre komponentene inneholder informasjon om art og sted, som kan stemme med tidligere funn. Yasumoto et al. 1978 viste at toksiner tas opp og påvirker arter forskjellig, mens Alarcan et al. 2018 viste at det er forskjell mellom toksingrupper og konsentrasjoner i skjellene avhengig av sted.

4.1.5 Nullmetoden

Dersom nullmetoden gir en RMSEP-verdi som er lavere enn RMSEP-verdiene for regresjonsmodellene vil nullmetoden være den beste prediksjonen for responsen. En høy RMSEP-verdi antyder at metoden er dårlig og ikke vil ha en veldig god prediksjonsevne. Dersom R^2_{pred} gir en negativ verdi vil også nullmetoden gi en bedre prediksjon.

Nullmetoden er uavhengig av regresjonsmetode fordi den kun presenterer $\bar{y}$ som prediksjonsmodell og setter $\hat{\beta} = 0$. Det er $\hat{\beta}$ som inneholder informasjon gitt under konstruksjonen av regresjonsmetodene og dermed avhenger av dette.

4.1.6 Standardisering av variabler

Variabler bør standardiseres dersom de er målt på en skala med veldig stor spredning i størrelse eller ikke sammenlignbare enheter (Johnson & Wichern 2002). Dersom det er stor spredning i variasjonen vil variablene med høy verdi ha en stor innvirkning på konstruksjonen av prinsipalkomponentene, da spesielt de største. Standardiseringen sørger for at alle variablene får samme varians, lik 1. Det betyr at variabler både med stor og svært liten varians vil ha samme vekt under konstruksjonen av ladningene i regresjonsmetodene. Siden datasettet i denne oppgaven har noen variabler med veldig

stor varians, er variablene standardisert i alle analyser. Dette kan derimot by på et problem dersom variablene med små varianser ikke er like relevant til responsen i forhold til variablene med store varianser. Tre av variablene i datasettet har så stor varians at det ville påvirket analysene hvis ikke standardisering ble utført.

4.1.7 Deling i kalibreringsett og testsett

Modellen med kalibreringsettet lages før den tilpasses med et testsett som ble holdt helt utenfor under konstruksjonen av metoden. Forskjellen mellom den sanne responsen og den predikerte responsen gitt av testsettet brukes til å sjekke hvor god metoden er.

Delingen inn i kalibreringsett og testsett er gjort tilfeldig. Dette betyr at hvilke prøver som er med i kalibreringsettet ikke er valgt, men trukket tilfeldig. Det betyr også at hver gang det lages et nytt kalibreringsett vil det være ulikt fra det forrige settet. Dette kan brukes til å se på feilen i metoden og forhindrer forventningskjevhet. Datasettet består bare av 32 prøver som er relativt lite. Når datasettet igjen blir delt i to baseres metoden på et enda mindre kalibreringsett. I denne oppgaven er det valgt å gi kalibreringsettet 20 prøver og testsettet 12 prøver. En metode basert på bare 20 prøver kan gi mer ustabile estimater og er ikke nødvendigvis så god.

Et problem ved bruk av kryssvalidering som valideringsmetode for variabelseleksjoner som forlengs utvelgelse, er at dette i større grad fører til overtilpasning enn ved bruk av testsettvalidering. Dette er vist i blant annet Reunanen 2003, som anbefaler bruk av testsettvalidering som vurdering av kvaliteten på prediksjonsmodellen. Kryssvalidering anbefales kun å brukes som en veiledning til å oppnå en metode, ikke til sammenligning av kvaliteten på de forskjellige metodene.

4.2 PCA

Det er viktig å notere at prinsipalkomponentene er vektede summer av de originale variablene hvor vektene er laget basert på forklaringsvariablene. Et screeplot ble laget som viser en god fordeling av egenverdiene, med to store egenverdier først mens resten av egenverdiene synker jevnt.

Når kovariansmatrisen brukes forklarer den ene komponenten 70% av variansen. Det er fordi det er tre fettsyrer som har svært stor varians. Dette er illustrert grafisk i figur 3.12. Ladningsplottet viser også tydelig at noen få fettsyrer har stor variasjon. Dette ladningsplottet gjør det veldig vanskelig å se på forskjellene mellom fettsyrene siden de fleste har verdi tilnærmet lik null både for PC1 og PC2. For å korrigere ut denne store variansforskjellen brukes korrelasjonen mellom fettsyrene i stedet for kovariansen. Det gir et mer oversiktlig plot (vist i figur 3.11) hvor fettsyrene klart skilles og kan tolkes mot hverandre selv om mindre av variansen blir forklart.

Ut fra ladningsplottet i figur 3.13 ser man at de mettede fettsyrene gir et stort positivt utslag på PC1. Fettsyrene i nedre sjikt av plottet er hovedsaklig fettsyrer med en eller flere dobbelbindinger og alle disse har cis-konfigurasjon. Alle fettsyrene med trans-konfigurasjon har stort sett stort negativt utslag på PC1. Omega-3 fettsyrene danner to separate grupperinger, en med lite utslag på PC1 og den andre med variert utslag på PC1. PC2 gav ingen klar tolkning i ladningsplottet.

Scoreplottet i figur 3.9 viser tydelig grupperinger mellom de fire skjellartene. Det viser at det er forskjell mellom artene. Fra plottet kan det tyde på at PC1 har stor positiv effekt fra Stillehavsøstersene og stor negativ effekt fra Blåskjellene. PC2 har stor negativ effekt fra O-skjellene. Scoreplottet i figur 3.10 viser noen grupperinger av steder, men veldig få prøver er fra samme sted og plottet gir derfor lite informasjon.

Sammenligninger mellom ladningsplot og scoreplot viser noen sammenhenger mellom fettsyrer og arter. For PC1 har Stillehavsøsters positivt utslag i scoreplottet, mens mettede fettsyrer har positivt utslag i ladningsplottet. Det kan tyde på at Stillehavsøs-

tersene inneholder mer mettede fettsyrer enn de andre artene. Blåskjell har negativt utslag på PC1 i scoreplottet, mens mange av trans-fettsyrene har negativt utslag i ladningsplottet. Dermed kan det tyde på at Blåskjellene inneholder mer trans-fettsyrer. I scoreplottet viser PC2 forskjell mellom Blåskjell med stort positivt utslag og O-skjell med stort negativt utslag. Siden ingen info kan tolkes fra ladningsplottet kan ikke dette settes i sammenheng til fettsyreinneholdet.

4.3 Regresjonsmetoder

Flere forskjellige regresjonsmetoder ble prøvd ut for å finne en prediksjonsmodell for toksisiteten i skjell. Den første metoden var PCR som finner et lite antall vektorer som erstatter **X** ved å finne den maksimale kovariansen for prediksjonsvariablene. Informasjonen i responsen **y** blir ikke tatt hensyn til når de nye vektorene lages. Det er dermed ikke sikkert at variasjonen i **X** har sammenheng med variasjonen i **y**. Altså kan det ikke garanteres at ladningene som lages for **X** gir noen god prediksjon av fremtidige responser. Det er vist at kovariansmatrisen til forklaringsvariablene påvirkes veldig av noen få variabler med veldig stor varians, som kan tyde på at ladninger laget på denne måten vil ha vanskeligheter med å predikere fremtidige prøver. Plottene laget for PCR viser også at det må være med veldig mange komponenter med i metodene. En av metodene for PCR som fikk en veldig lav RMSEP-verdi helt ned på 110.68 med 28 komponenter. Dette er metoden med logtransformering som er vist i figur 3.22. Dette plottet viser en veldig ustabil RMSEP-verdi som stiger drastisk etter bunnpunktet. At så mange komponenter inkluderes er et tegn på overtilpasning og kan være grunnen til at RMSEP-verdien blir så lav. Dersom metoden er overtilpasset vil den ikke kunne brukes til prediksjon i fremtiden, da vil estimatene estimeres feil og metoden vil ikke passe til de nye prøvene.

For å finne en metode som er mer relatert til **y** ble også PLS brukt til å prediksjon av toksisiteten. I PLS maksimeres kovariansen mellom responsen og forklaringsvariablene ved bruk av latente variabler. Plottene for PLS viser mer stabile RMSEP-verdier enn

PCR. PLS trenger en del færre komponenter enn PCR, men kommer likevel ned til en lav RMSEP-verdi. Den beste metoden for PLS er logtransformering som fikk en RMSEP-verdi på 146.14 på 14 komponenter. Plottet viser en stabil RMSEP-verdi i figur 3.30. Selv om RMSEP-verdien for denne metoden er noe høyere enn for PCR, er antallet komponenter så mye lavere at metoden sannsynligvis er mindre overtilpasset. Dermed vil PLS kunne predikere toksisiteten i fremtidige skjell bedre.

Det er tidligere vist at mengden toksisitet i et skjell avhenger av art og sted. Noen av metodene inneholder i tillegg denne informasjonen. CPLS er en utvidelse av PLS hvor tilleggsvariabler kan modelleres som en ekstra respons. Her brukes kun denne ekstra informasjonen når selve metoden lages og ikke når metoden brukes til prediksjon. Dette betyr at disse metodene kan brukes på helt nye ukjente prøver med ukjente arter og steder. Metoder for CPLS ble laget med dummy variabler for art og sted som tilleggsrespons. Ettersom at dette er en utvidelse av PLS har plottene samme form som PLS metodene. Siden mer informasjon er inkludert i metoden gir CPLS en mer kompakt metode enn PLS, men det betyr dermed ikke automatisk at CPLS gir en bedre prediksjonsmodell. CPLS med logtransformering (figur 3.36) blir den beste metoden, hvor laveste RMSEP-verdi etter transformering tilbake blir 151.3 på 5 komponenter. Etter det optimale antallet komponenter stiger RMSEP-verdien for denne metoden raskt og dersom et større antall komponenter velges blir modellen fort dårlig. Av alle de tre metodene har CPLS det største minimumet for RMSEP-verdien, men antallet komponenter inkludert i metoden er det laveste. Det er ønskelig med et lavest mulig antall komponenter som kan forklare all variabiliteten i metoden.

Estimatene for σ fra Bayes-PLS viser at metoden med residualene som respons gir en lavere nedre prediksjonsgrense enn med originaldata som respons. I forhold til metodene som ble analysert stemmer dette ikke overens og residualmodellene presterte dårligst i både PCR og PLS. De logtransformerte metodene av originaldatasettet slår estimater for nedre grense i tre av regresjonsmetodene; PCR, PLS og CPLS, men ikke i forlengs utvelgelse. Den nedre prediksjonsgrensen er veldig høy for dette datasettet og det kan antyde at sammenhengen mellom fettsyreprofilene til skjellene og toksisiteten ikke har et veldig sterkt forhold.

4.3.1 Forlengs utvelgelse

Valideringen av seleksjonen med forlengs utvelgelse er gjort ved bruk av testsettvalidering. Forskjellen mellom den sanne responsen og den predikerte responsen fra testsettet brukes til å sjekke hvor god metoden er. Et problem med dette er at konstruksjonen av metoden er gjort basert på et mindre antall prøver enn de andre regresjonsmetodene.

Den beste metoden for forlengs utvelgelse ble logtransformert respons ($\alpha = 0.05$) som gav en RMSEP-verdi på 190.11 med 4 variabler. Denne RMSEP-verdien er høyere enn verdiene gitt av de tre andre regresjonsmetodene; PCR, PLS og CPLS.

Forlengs utvelgelse ble forsøkt med $\alpha = 0.20$ i tillegg og da ble 17-18 variabler inkludert i metoden. I praksis beholdes da alle eller nesten alle variablene med i metoden ettersom at antallet variabler ikke kan overstige antallet prøver. Dermed gav $\alpha = 0.20$ og forlengs utvelgelse ikke en god metode for prediksjon.

4.3.2 Vurdering av prediksjonsmodellene

For å få litt mer informasjon om prediksjonsmodellene ble den målte toksisiteten plottet mot den predikerte responsen for alle metodene. En trendlinje ble lagt til alle plottene for å se hvordan datapunktene plasserte seg i forhold til den. For de fleste regresjonsmetodene viser disse plottene to grupperinger, en med lav toksisitet og en med høy toksisitet. Grupperingen kommer av at det er noen få prøver med høyere toksisitet enn det de fleste prøvene har. For prøvene med høy toksisitet er det enkelt å se hvordan sammenhengen endres når regresjonsmetode endres. Prøvene med lav toksisitet ligger tettere og gjør det vanskelig å se forskjeller mellom regresjonsmetoder.

For de tre metodene som ble valgt ut som best, viser plottene verdier som ligger seg i nærheten av trendlinjen (figurer 3.23, 3.31 og 3.37). Det er ønskelig at punktene skal ligge nærmest mulig trendlinjen for å redusere residualleddet. For alle tre metodene lå punktene noe nærmere trendlinjen ved originaldata enn ved logtransformert respons.

Denne trenden vises også i egenskapsplottene som viser relevante komponenter. Likevel gir metodene basert på logtransformering de laveste RMSEP-verdiene.

Plottene basert på originaldata i figurer 3.19, 3.27 og 3.35 viser at flere prøver predikeres som negative for disse tre metodene. Dette er metoder basert på originaldata og disse plottene vitner om en svakhet i metodene for skjell som har veldig lav toksisitet. Metoder basert på logtransformering vil ikke gi negative prediksjoner. Dette kan tyde på at denne transformeringen passer bedre til dataene som er i grupperingen med lav toksisitet og derfor vurderes modellene som de beste.

4.3.3 Bruk av prediksjonsmodellen i analyseforsøk

Prediksjonsmodellen som produseres følger modellen i lign. 2.27. Metoden som velges ut til å være den beste estimerer parametrene $\hat{\beta}$, $\bar{y}$ og $\bar{\mathbf{x}}$ i modellen. Estimatene er gitt i vedlegg A. Når nye prøver skal predikeres toksisiteten i, brukes denne modellen. Før de nye prøvene settes inn i denne ligningen må de nye forklaringsvariablene standardiseres ettersom at prediksjonsmodellen er basert på standardiserte variabler. De samme verdiene må brukes til standardisering av nye prøver og dette gjøres ved lign. 2.29 som sørger for at de nye variablene korrigeres på samme måte. Dette skal gi samme prediksjonsevne som den opprinnelige metoden gjorde, dersom dette var en god metode.

Siden metodene som er valgt ut som de beste alle bruker logtransformering av responsen, må dette tas hensyn til under prediksjon. Når et nytt datasett settes inn i prediksjonsmodellen vil den predikerte toksisiteten være på logaritmisk form. For å transformere prediksjonen tilbake til vanlig skala må det tas $e^{\hat{y}}$ av den nye responsen. Slik finnes toksisiteten i nye prøver.

4.4 Metodevalidering

Dersom vi leter etter den beste metoden kan det ende opp med at man tilfeldigvis tipper på den riktige metoden for det gitte datasettet. Det er viktig å finne ut om en kompleks metode er signifikant bedre enn en enklere metode, slik at bruken av den komplekse metoden kan forsvares. Derfor er det ønskelig med en metode som gir en svært lav RMSEP.

Resultatene fra ANOVA og Tukey par-vis test gir signifikant forskjell mellom CPLS/PLS og PCR. PCR kommer lavest ned i RMSEP, men på et høyt antall komponenter. Dette betyr mest sannsynlig at denne metoden er overtilpasset og at det er derfor den kommer så langt ned. Det er ingen signifikant forskjell mellom CPLS og PLS, disse vil prestere nesten like bra. PLS har en noe lavere RMSEP-verdi enn CPLS, men på langt flere komponenter.

Altså er det igjen to metoder som ikke har noe signifikant forskjell. Den ene med en lavere RMSEP-verdi, men et høyere antall komponenter og den andre med noe høyere RMSEP-verdi, men et lavere antall komponenter. Hvilken av disse metodene som er best må da baseres på RMSEP-plot og plot av $\mathbf{y}$ og $\bar{\mathbf{y}}$. Metoden basert på CPLS har et mye mer stabilt RMSEP-plot enn PLS metoden. Plottene over de predikerte verdiene er veldig like for de to metodene og viser ingen tydelig forskjell. Siden CPLS metoden er mer stabil og har et lavere antall komponenter blir denne valgt ut som den beste metoden.

4.5 Planleggingen av forsøket

Prøvene som er brukt i oppgaven består av prøver tatt i løpet av en lengre tidsperiode hvor veldig få av prøvene er tatt til samme tid. Prøvene kommer også fra lokasjoner spredt utover hele Norge. Fra noen av lokasjonene er det kun tatt en prøve. Innad i artene er det også et veldig forskjellig antall prøver. Kombinasjonen av dette gjør at

sammenligning av prøver, arter og steder er vanskelig.

Den ene prøven er valgt ut basert på kriteriet at den er et eksempel på en ekstrem algeoppblomstring. De ekstreme forholdene gjør at prøven er ikke sammenlignbar med resten og prøven er ikke representativ for området eller arten. Samtidig som skjellprøvene tas blir det tatt prøver av vannet som testes for giftproduserende alger. Dersom vannet inneholder ekstremt høye konsentrasjoner av giftproduserende alger er sannsynligheten for at skjellene i området er giftige veldig høy. Prediksjonsmodellen gjelder for skjell som en tilfeldig tid av året er antatt å være under normale forhold. Dersom modellen skulle ta høyde for ekstrem algeoppblomstring ville den blitt mer kompleks og sannsynligheten for overestimering ville steget betraktelig. Slike forhold representeres dermed ikke i metodene.

Prøvene er valgt ut basert på at de har et visst nivå av toksisitet. Dersom det er vanligere med prøver som er mindre giftige vil metoden som velges basert på datasettet i denne oppgaven overestimere toksisiteten når den brukes på nye prøver. Dette kan dermed bli på som et eksempel for hvordan en slik type analyse kan gjennomføres og hvordan det kan brukes til å forbedre en analyse.

I dette forsøket er det kun er sett etter fettsyrer fra OA-gruppen. Metoden som brukes for å analysere skjellene detekterer også tre andre grupper med toksiner. De andre gruppene kan også gi forgiftning i tillegg til at blandinger av toksiner fra flere grupper kan gi forgiftning (Alarcan et al. 2018). Skjellene kan altså innholde andre toksiner en de som er målt og prediksjonsmodellen vil kun gjelde for toksiner fra OA-gruppen.

5 Videre arbeid

Resultatene viser at forholdet mellom fettsyresammensetning og toksininnhold ikke er så sterkt. Noe som kan forbedre prediksjonen er å bruke en annen analysemetode til å måle toksisiteten. Et eksempel på en annen metode å måle toksisitet er NIR-Raman som er under utvikling, men kan måle DSP-toksiner ned til $75\mu g$ (Pinzaru et al. 2016). Et annet alternativ kan være å bruke NMR til deteksjon av toksisiteten.

Det enkleste alternativet vil være å øke antallet prøver slik at n blir større enn p. Da kan minste kvadrater brukes og prediksjonen vil bli forventningsrett. Flere prøver kan også føre til en mer generell prediksjonsmodell som er mer representativ for populasjonen. Dette kan gi en lavere RMSEP-verdi.

Noe annet som kan gjøres er å finne et nytt sett med forklaringsvariabler for å se om det gir en bedre prediksjon. En slik forbedringen kan gjøres med en optimalisering av GC-MS metoden for å øke sensitiviteten. Dersom det undersøkes hvilke fettsyrer som er relatert til toksisitet kan det søkes kun for disse under GC-MS analysen og fettsyrene som kun bidrar med støy vil dermed kuttes ut fra datasettet.

Et alternativ kan også være å lage en mer stedsspesifikk metode kan inneholde informasjon om et sted som algemengde, havtemperatur, vindretning og oksygenmetningen i havet. Ekstra informasjon som denne kan føre til mer nøyaktige metoder. Metoden vil få veldig mange forklaringsvariabler og forskjellen mellom n og p øker. Med dette alternativet kan forklaringsvariabler finnes som har et sterke forhold til toksisiteten enn fettsyresammensetningen har. Med flere bedre foklaringsvariabler kan nedre grense for prediksjon senkes og en lavere RMSEP-verdi blir lettere å oppnå.

Noe som ikke er prøvd i denne oppgaven er klassifisering. Det er vist at det er forskjell mellom artene, derfor kan klassifisering av art gi interessante resultater. Resultatene i denne oppgaven viste også grupperinger i toksisiteten blant prøvene. Klassifisering kan dele gruppene inn i veldig toksisk eller mindre toksisk. En metode kan utvikles som predikerer hvilken av gruppene prøven vil havne innunder og dermed graden av toksisitet.

6 Konklusjon

I denne oppgaven gis en prediksjonsmodell for toksisitet i skjell basert på fettsyreprofiler. Dersom modellen er god kan den erstatte den komplekse LC-MS/MS analyse som brukes til å måle toksisiteten idag.

Deteksjon av toksisitet i skjell har noen svakheter og istedet for optimalisering av analysen kan den erstattes av multivariat statistikk. Ulike metoder er utprøvd: PCA, PCR, PLS, CPLS og variabelseleksjon. Metodene er validert med kryssvalidering og testsettvalidering. Tre regresjonsmetoder ble valgt ut, en fra hver metode, og disse ble validert ved bruk av CVANOVA. Signifikant forskjell mellom metodene ble påvist og Tukey *post hoc*-test ble gjennomført for å sammenligne de tre. CPLS med logtransformering ble vurdert som den beste metoden etter analysen.

Metoden som anbefales for prediksjon av toksisitet i skjell basert på fettsyresammensetning ble CPLS med logtransformering av responsen på 5 komponenter og RMSEP-verdi på 151.37. Siden variablene standardiseres før analyse må de samme verdiene brukes til standardisering av nye prøver. Parametrene bestemt fra den beste modellen er gitt og må settes inn i prediksjonsmodellen når nye prøver skal predikeres.

Svakheter med metodene tyder på at forholdet mellom fettsyresammensetningen og toksisiteten ikke er veldig sterkt. Andre forklaringsvariabler eller måling av toksisitet må til for å kunne oppnå en lavere nedre grense for prediksjon.

Bibliografi

Aanrud, S. (2016), 'Fettsyreestere i toksiste skjell: fettsyreprofil, dtx-3-profil og optimalisering av hydrolyse', *Ås: Norges miljø- og biovitenskapelige universitet* . Masteroppgave.

Alarcan, J., Biré, R., Hégarat, L. L. & Fessard, V. (2018), 'Mixtures of lipophilic phycotoxins: Exposure data and toxicological assessment', *Marine Drugs* . doi: 10.3390/md16020046.

Andersen, C. M. & Bro, R. (2010), 'Variable selection in regression - a tutorial', *Journal of Chemometrics* . doi: 10.1002/cem.1360.

Bauder, A. G., Cembella, A. D., Bricelj, V. M. & Quilliam, M. A. (2001), 'Uptake and fate of diarrhetic shellfish poisoning toxins from the dinoflagellate prorocentrum lima in the bay scallop argopecten irradians', *Marine Ecology Progress Series* . doi: 10.3354/meps213039.

Cederkvist, H. R., Aastveit, A. H. & Næs, T. (2005), 'A comparison of methods for testing differences in predictive ability', *Journal of Chemometrics* . doi: 10.1002/cem.956.

de Hoffmann, E. & Stroobant, V. (2007), *Mass Spectrometry: Principles and Applications*, 3. edn, John Wiley and sons LTD, Chichester, UK.

EU-RL-MB (2015), *EU-Harmonised Standard Operating Procedure for determination of Lipophilic marine biotoxins in molluscs by LC-MS/MS*.

Fux, E., Bire, R. & Hess, P. (2008), 'Comparative accumulation and composition of li-

pophilic marine biotoxins in passive samplers and in mussels (m. edulis) on the west coast of ireland', *Harmful Algae* . doi: 10.1016/j.hal.2008.10.007.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer-Verlag, New York.

Helland, I. S. (2001), 'Some theoretical aspects of partial least squares regression', *Chemometrics and Intelligent Laboratory Systems* . doi: 10.1016/S0169-7439(01)00154-X.

Helland, I. S. & Almøy, T. (1994), 'Comparison of prediction methods when only a few components are relevant', *Journal of the American Statistical Association* . doi: 10.2307/2291191.

Helland, I. S., Sæbø, S. & Tjelmeland, H. (2012), 'Near optimal prediction from relevant components', *Scandinavian Journal of Statistics* . doi: 10.1111/j.1467-9469.2011.00770.x.

Holmes, C. F., Luu, H. A., Carrie, F. & Schmitz, F. J. (1990), 'Inhibition of protein phosphatases- 1 and -2a with acanthifolicin: Comparison with diarrhetic shellfish toxins and identification of a region on okadaic acid important for phosphatase inhibition', *Federation of European Biochemical Societies* . doi: 00!45793/90/3.50.

Höskuldsson, A. (2000), 'Variable and subset selection in pls regression', *Chemometrics and Intelligent Laboratory Systems* . doi: 10.1016/S0169-7439(00)00113-1.

Indahl, U. G., Liland, K. H. & Næs, T. (2009), 'Canonical partial least squares—a unified pls approach to classification and regression problems', *Journal of Chemometrics* . doi: 10.1002/cem.1243.

Indahl, U. G. & Næs, T. (1998), 'Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling', *Journal of Chemometrics* . doi: 10.1002/(SICI)1099-128X(199807/08)12:4<261::AID-CEM513>3.3.CO;2-Q.

Johnson, R. A. & Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, 5. edn, Prentice Hall, New Jersey.

Khanmohammadi, M. (2014), *Current Applications of Chemometrics*, Nova Science Publishers, Hauppauge, New York.

Lawrence, J., Loreal, H., Toyofuku, H., Hess, P., Iddya, K. & Ababouch, L. (2011), 'Assessment and management of biotoxin risks in bivalve molluscs', *FAO Fisheries and Aquaculture Technical Paper No. 551* .

Lay, D. C., Lay, S. R. & McDonald, J. J. (2016), *Linear Algebra and its Applications*, Pearson Education Limited, Essex.

Lee, J.-S., Igarashi, T., Fraga, S., Dahl, E., Hovgaard, P. & Yasumoto, T. (1989), 'Determination of diarrhetic shellfish toxins in various dinoflagellate species', *Journal of Applied Phycology* . doi: 10.1007/BF00003877.

Liland, K. H. (2009), 'Multivariate analysis - method development and novel applications in spectrometry', *Ås: Norges miljø- og biovitenskapelige universitet* . Doktoravhandling.

Lindegarth, S., Torgersen, T., Lundve, B. & Sandvik, M. (2009), 'Differential retention of okadaic acid (oa) group toxins and pectenotoxins (ptx) in the blue mussel, mytilus edulis (l.), and european flat oyster, ostrea edulis (l.)', *Journal of Shellfish Research* . doi: 10.2983/035.028.0213.

Mandel, J. (1982), 'Use of the singular value decomposition in regression analysis', *The American Statistician* . doi: 10.2307/2684086.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979), *Multivariate Analysis*, 3. edn, Academic Press, London.

Martens, H. & Næs, T. (1989), *Multivariate Calibration*, John Wiley and Sons, Guildford.

Mattilsynet (2018), 'Matportalen.no: Blåskjellvarsel'. Lest: 27.02.2018 09:53.
URL: *http://www.matportalen.no/verktoy/blaskjellvarsel/*

Montgomery, D. C. (2013), *Design and Analysis of Experiments*, 8. edn, John Wiley and Sons Inc., Singapore.

Næs, T. & Helland, I. S. (1993), 'Relevant components in regression', *Scandinavian Journal of Statistics* . 20 (3): 239-250.

Næs, T., Isaksson, T., Fearn, T. & Davies, T. (2002), *A user-friendly guide to: Multivariate Calibration and Classification*, 1. edn, NIR Publications, Chichester, West-Sussex.

Pinzaru, S. C., Müller, C., Tódor, I., Glamuzia, B. & V.Chis (2016), 'Nir-raman spectrum and dft calculations of okadaic acid dsp marine biotoxin microprobe', *Journal of Raman Spectroscopy* . doi: 10.1002/jrs.4870.

Reunanen, J. (2003), 'Overfitting in making comparisons between variable selection methods', *Journal of Machine Learning Research* . 3 (2003): 1371-1382.

Seisonen, S., Vene, K. & Koppel, K. (2016), 'The current practice in the application of chemometrics for correlation of sensory and gas chromatographic data', *Food Chemistry* . doi: 10.1016/j.foodchem.2016.04.134.

Steidinger, K. A. (1993), 'Some taxonomic and biologic aspects of toxic dinoflagellates', *Algal Toxins in Seafood and Drinking Water* . ISBN: 978-0-12-247990-8.

Suganuma, M., Fujiki, H., Suguri, H., Yoshizawa, S., Hirota, M., Nakayasu, M., Ojika, M., Wakamatsu, K., Yamada, K. & Sugimura, T. (1988), 'Okadaic acid: An additional non-phorbol-12-tetradecanoate-13- acetate-type tumor promoter', *Proceedings of the National Academy of Sciences of the United States of America* . doi: 10.1073/pnas.85.6.1768.

Suzuki, T., Ota, H. & Yamasaki, M. (1999), 'Direct evidence of transformation of dinophysistoxin-1 to 7-o-acyl-dinophysistoxin-1 (dinophysistoxin-3) in the scallop patinopecten yessoensis', *Toxicon* . doi: 10.1016/S0041-0101(98)00182-2.

Suzuki, T., Yoshizawa, R., Kawamura, T. & Yamasaki, M. (1996), 'Interference of free fatty acids from the hepatopancreas of mussels with the mouse bioassay for shellfish toxins', *Lipids* . doi: 10.1007/BF02523835.

Svensson, S. & Förlin, L. (2003), 'Analysis of the importance of lipid breakdown for

elimination of okadaic acid (diarrhetic shellfish toxin) in mussels, mythilus edulis: results from a field study and a laboratory experiment', *Aquatic Toxicology* . doi: 10.1016/j.aquatox.2003.11.002.

Sæbø, S., Almøy, T. & Helland, I. S. (2015), 'simrel — a versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors', *Chemometrics and Intelligent Laboratory Systems* . doi: 10.1016/j.chemolab.2015.05.012.

Séchet, V., Safran, P., Hovgaard, P. & Yasumoto, T. (1990), 'Causative species of diarrhetic shellfish poisoning (dsp) in norway', *Marine Biology* . doi: 10.1007/BF01344296.

Torgersen, T., Sandvik, M., Lundve, B. & Lidegarth, S. (2008), 'Profiles and levels of fatty acid esters of okadaic acid group toxins and pectenotoxins during toxin depuration. part ii: Blue mussels (mytilus edulis) and flat oyster (ostrea edulis)', *Toxicon* . doi: 10.1016/j.toxicon.20 08.06.011.

Tukey, J. W. (1949), 'Comparing individual means in the analysis of variance', *International Biometric Society* . doi: 10.2307/3001913.

van den Top, H. J., Gerssen, A. & van Egmond, H. P. (2011), 'Quantitative determination of marine lipophilic toxins in shellfish using lc-ms/ms - international validation study - final report', *RIKILT, Institute of Food Safety* .

Wold, S., Sjöström, M. & Eriksson, L. (2001), 'Pls-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems* . doi: 10.1016/S0169-7439(01)00155-1.

Yasumoto, T., Oshima, Y. & Yamaguchi, M. (1978), 'Occurrence of a new type of shellfish poisoning in the tohoku district', *Bulletin of the Japanese Society of Scientific Fisheries* . doi: 10.2331/suisan.44.1249.

Figurer

Vedlegg A: Regresjonskoeffisienter

Prediksjonsmodell: $e^{\hat{y}} = \bar{y} + \hat{\beta}^T \times \mathbf{x}^*$

Hvor $\mathbf{x}^* = \frac{\mathbf{x}_i - \mathbf{x}_{i,cal}}{SD(\mathbf{x}_{i,cal})}$

$\bar{y} = 4.169$

Tabell 1 viser $\hat{\beta}$, $\bar{\mathbf{x}}$ og $SD(\mathbf{x}_{i,cal})$.

Beta	Sentrert x	$SD(\mathbf{x}_{i,cal})$	Fettsyre
0.3658	0.7730	0.0423	C12:0
-0.0824	3.931	1.111	C14:0
0.8072	2.337	0.6237	C13:0 4,8,12 trimetyl
0.2331	0.9254	0.1252	C14:0 13-metyl
-0.3797	0.9577	0.0440	C14:0 12-metyl
0.1307	3.297	0.2330	C15:0
0.0643	2.133	0.0622	C15:0 14-metyl
0.0513	10.23	2.066	C16:0
0.1099	1.541	0.1013	C16:1 t7
-0.1556	0.9307	0.2775	C16:0 15-metyl
-0.1655	1.515	5.461	C16:1 c9
0.5432	4.313	0.1011	C16:1 c11
0.0056	2.013	0.4860	C17:0
-0.1707	2.724	1.403	C18:0
0.2785	1.398	0.0894	C18:1 c5 og c6

0.4575	2.879	0.9312	C18:1 c9
0.3079	2.289	1.180	C18:1 c11
-0.0899	2.209	0.0632	C18:1 c13
0.2258	0.6567	0.1805	C16:4 c6,9,12,15
0.1568	0.4560	0.0459	C19:0
-0.1585	3.994	0.5150	C18:2 c9,12
-0.0427	1.001	0.1294	C18:2 c11,14
0.6744	0.8980	0.0692	C19:1 t10
-0.5984	0.6083	0.0578	C18:3 c6,9,12
0.0318	0.9212	0.0773	C20:0
-0.5313	3.067	0.4626	C18:3 c9,12,15
0.2241	0.4157	0.0507	C20:1 c5
0.0362	2.212	0.5040	C20:1 c9
-0.2638	2.348	0.9913	C20:1 c11
0.1063	1.126	1.422	C20:1 c13
-0.2978	3.365	1.040	C18:4 c6,9,12,15
0.1472	1.374	1.123	C20:2 t8,11
-0.1676	1.785	0.2314	C20:2 t11,14
0.1617	0.3434	0.0586	C20:2 c11,13
-0.2529	3.168	0.1775	C20:2 c11,14
0.3228	1.095	0.2173	C20:3 c5,11,14
0.2747	1.003	0.0499	C20:3 c8,11,14
-0.6587	0.3157	0.0278	C22:0
-0.6515	2.755	0.6138	C20:4 c5,8,11,14
-0.5291	0.6240	0.0628	C22:1 c13
-0.4196	1.924	0.1662	C20:4 c8,11,14,17
0.1369	1.665	0.2725	C22:2 c5,13
0.1275	1.942	0.9915	C22:2 c13,16
-0.0235	3.439	3.700	C20:5 c5,8,11,14,17
0.0915	5.855	0.0966	C21:5 c6,9,12,15,18
0.0536	1.965	0.1482	C22:5 c4,7,10,13,16

-0.3804	4.260	0.1752	C22:5 c7,10,13,16,19
0.0952	2.998	6.102	C22:6 c4,7,10,13,16,19

Tabell 1: $\hat{\beta}$, $\bar{\mathbf{x}}$ og $SD(\mathbf{x}_{i,c})$ for hver fettsyre fra CPLS med 5 komponenter

Vedlegg B: R-kode

```
# Lasting av pakker
library(pls)
library(mixlm)

# Laster inn datasett
load("FAME.RData")

# Henter residualene fra modellen
res.mod <- lm(Total.toksin ~ Sted + Art, data = FAME)
kor.res <- resid(res.mod)
kor.fit <- fitted(res.mod)
plot(kor.fit, kor.res, ylab = "Residualer", xlab = "Fitted values",
    main = "Residualer vs. fitted values")

# Lager et nytt datasett uten B-1443
fame_red <- FAME[-14,]
# Henter ut fettsyreprofilene fra datasettet
fame_red2 <- fame_red[2:49]
# Henter ut toksinmengden i prøvene
toksin <- fame_red[,56]
# Henter ut residualer fra den nye modellen og plotter residualene mot
    den sanne responsen
res.mod2 <- lm(Total.toksin ~ Sted + Art, data = fame_red)
```

```
kor.res2 <- resid(res.mod2)
kor.fit2 <- fitted(res.mod2)
plot(kor.fit2, kor.res2, ylab = "Residualer", xlab = "Fitted values",
    main = "Residualer vs. fitted values")

# Log-transformering av residualmodellen
res.mod2.log <- lm(log(Total.toksin) ~ Sted + Art, data = fame_red)
kor.res2.log <- resid(res.mod2.log)
kor.fit2.log <- fitted(res.mod2.log)
plot(kor.fit2.log, kor.res2.log, ylab = "Residualer", xlab =
    "Fitted values", main = "Log transformerte residualer vs.
    fitted values")

# Propertyplot, plotter korrelasjonen til y og egenverdiene
# Kode fra Solve modifisert til å bruke korrelasjon istedet for kovarians
plotprops2 <- function(Y,X, doscaleX=FALSE, docenterX=TRUE,
ncomp=3, subset=NULL){
  n <- dim(X)[1]
  p <- dim(X)[2]
  ncomp <- min(ncomp,min(n,p))
  if(docenterX) ncomp <- ncomp-1
  if(ncomp<1)stop("Centering requires at least 2 components")
  if(is.null(subset)) subset <- 1:n
  X <- scale(X[subset,], center=docenterX, scale=doscaleX)
  Y <- matrix(Y, ncol=1)[subset,,drop=F]
  svdres <- svd(X)
  eigval <- (svdres$d^2)/(svdres$d^2)[1]
  Z <- X%*%svdres$v
  cors <- cor(Y, Z)
  cors <- abs(cors)/max(abs(cors))
  par(mar=c(5.1, 4.1, 4.1, 4.1))
```

```
  plot(1:ncomp, eigval[1:ncomp], type="h", lwd=2, xlab="Component",
  ylab="Scaled eigenvalue", axes=FALSE, main="Property plot")
  points(1:ncomp, cors[1:ncomp], type="p", pch=20, cex=2, col=2)
  axis(1)
  axis(2,at=seq(0,1,0.1), labels=as.character(seq(0,1,0.1)))
  axis(4,at=seq(0,1,0.1), labels=as.character(seq(0,1,0.1)))
  mtext("Scaled correlation",side=4, line=3)
  box()}

plotprops2(kor.res2, fame_red2, doscaleX = FALSE, docenterX = TRUE,
    ncomp = 20, subset = NULL)

# PCA
PCAModel <-
  prcomp(~C12.0+C13.0.4.8.12.trimetyl+C14.0+C14.0.12.metyl+C14.0.13.metyl
    +C15.0+C15.0.14.metyl+C16.0+C16.0.15.metyl+C16.1.c9+C16.1.c11+C16.1.t7
    +C16.4.c6.9.12.15+C17.0+C18.0+C18.1.c5.og.c6+C18.1.c9+C18.1.c11+
    C18.1.c13+C18.2.c9.12+C18.2.c11.14+C18.3.c6.9.12+C18.3.c9.12.15+
    C18.4.c6.9.12.15+C19.0+C19.1.t10+C20.0+C20.1.c5+C20.1.c9+C20.1.c11+
    C20.1.c13+C20.2.c11.13+C20.2.c11.14+C20.2.t8.11+C20.2.t11.14+
    C20.3.c5.11.14+C20.3.c8.11.14+C20.4.c5.8.11.14+C20.4.c8.11.14.17+
    C20.5.c5.8.11.14.17+C21.5.c6.9.12.15.18+C22.0+C22.1.c13+C22.2.c5.13+
    C22.2.c13.16+C22.5.c4.7.10.13.16+C22.5.c7.10.13.16.19+
    C22.6.c4.7.10.13.16.19, scale.=TRUE, data=u_flam)

# Eksempel på screeplot, scoreplot og ladningsplot
screeplot(PCAModel)
scoreplot(PCAModel, main='Scoreplot', comps=c(1,2), labels=u_flam[,'Art'],
    sub='Labels: Art')
loadingplot(PCAModel.3, main='Loadingplot', comps=c(1,2), scatter=TRUE,
    labels='names', sub='Labels: variables')
```

```
# Lager trenings- og testsett
smp_size <- floor(0.625 * nrow(fame_red))
set.seed(123)
train_ind <- sample(seq_len(nrow(fame_red)), size = smp_size)
train <- fame_red[train_ind, ]
test <- fame_red[-train_ind, ]

# Testsett validering for forlengs utvelgelse
# Kode fra Kristian Hovde Liland
fame_red2.train <- train[2:49]
v_flam <- as.data.frame(cbind(Total.toksin=train$Total.toksin,scale
    (as.matrix(fame_red2.train), center = colMeans(fame_red2.train),
    scale = apply(fame_red2.train,2,sd))))
wideF <- wideForward(log(Total.toksin) ~C12.0+C13.0.4.8.12.trimetyl+
    C14.0+C14.0.12.metyl+C14.0.13.metyl+C15.0+C15.0.14.metyl+C16.0
    +C16.0.15.metyl+C16.1.c9+C16.1.c11+C16.1.t7+C16.4.c6.9.12.15+
    C17.0+C18.0+C18.1.c5.og.c6+C18.1.c9+C18.1.c11+C18.1.c13+
    C18.2.c9.12+C18.2.c11.14+C18.3.c6.9.12+C18.3.c9.12.15+
    C18.4.c6.9.12.15+C19.0+C19.1.t10+C20.0+C20.1.c5+C20.1.c9+C20.1.c11
    +C20.1.c13+C20.2.c11.13+C20.2.c11.14+C20.2.t8.11+C20.2.t11.14+
    C20.3.c5.11.14+C20.3.c8.11.14+C20.4.c5.8.11.14+C20.4.c8.11.14.17
    +C20.5.c5.8.11.14.17+C21.5.c6.9.12.15.18+C22.0+C22.1.c13+
    C22.2.c5.13+C22.2.c13.16+C22.5.c4.7.10.13.16+C22.5.c7.10.13.16.19
    +C22.6.c4.7.10.13.16.19, data = v_flam, alpha = 0.20)
n <- nrow(fame_red2.train)
p <- length(wideF$p.values)
y_modX <- matrix(0, 32-length(train_ind), p)
for(k in 1:p){
  v_flam <- as.data.frame(cbind(Total.toksin=fame_red$Total.toksin,scale
    (as.matrix(fame_red2), center = colMeans(fame_red2[,]), scale =
```

```
    apply(fame_red2[train_ind,],2,sd))))
  eval(parse(text = paste("modX <- lm(log(Total.toksin) ~ ",
    paste(names(wideF$p.values)[1:k],collapse=" + "), ", data =
    v_flam[train_ind,])")))
  y_modX[, k] <- predict(modX, v_flam[-(train_ind), ,drop=FALSE])
}

RMSEP <- sqrt(colMeans((fame_red$Total.toksin[-train_ind] - exp(y_modX))^2))
plot(RMSEP, type = 'l', col = 'red', main = 'log(Forward selection)
    alpha = 0.05', xlab = 'number of variables')

# PCR
MVRModel.pcr <-  pcr(Total.toksin ~ X, data = u_flam, ncomp = 30,
    validation = 'LOO', scale = TRUE)

# PLS
MVRModel.pls <- plsr(Total.toksin ~ X, data = u_flam, ncomp = 30,
    validation = 'LOO', scale = TRUE)

# Eksempel på beregning av RMSEP-verdier og plotting av RMSEP-verdiene
RMSECV <- sqrt(colMeans((u_flam$Total.toksin -
    MVRModel.pcr.resid$validation$pred[,1,])^2))
plot(RMSECV, ylab = 'RMSECV', xlab = 'number of components',
    main = 'PCR log residualer', type = 'l', col = 'red')

# Eksempel på beregning av RMSEP-verdier fra logtrasformerte metoder
RMSECV <- sqrt(colMeans((u_flam$Total.toksin -
    exp(MVRModel.pcr.resid$validation$pred[,1,]))^2))

# Eksempel på plot av residualer mot tilpassede verdier
plot(MVRModel.pcr, ncomp=28, labels = rownames(u_flam))
```

```
# Lager dummy-variabler for art og sted
u_flam$X <- I(as.matrix(u_flam[,2:49]))
u_flam$dummy <- I(model.matrix(~y-1, data.frame(y = u_flam$Art)))
u_flam$dummy2 <- I(model.matrix(~y-1, data.frame(y = u_flam$Sted)))
d1 <- u_flam$dummy
d2 <- u_flam$dummy2
u_flam$Y <- cbind(as.matrix(d1), as.matrix(d2))
# CPLS med art og sted som tilleggs respons
cpls.mod <- cppls(Total.toksin ~ X, data = u_flam, Y.add = Y,
    scale = TRUE,validation = "LOO")
plot(RMSEP(cpls.mod), main = "Total toksin", xlab = "Antall komponenter")

# Eksempel på CVANOVA
pred_y_PLS <- predict(MVRModel.pls)
feil_PLS <- (exp(pred_y_PLS[,,14]) - u_flam$Total.toksin)^2
pred_y_PCR <- predict(MVRModel.pcr)
feil_PCR <- (exp(pred_y_PCR[,,28]) - u_flam$Total.toksin)^2
pred_y_CPLS <- predict(cpls.mod)
feil_CPLS <- (exp(pred_y_CPLS[,,5]) - u_flam$Total.toksin)^2
stack_PCR <- stack(feil_PCR)
stack_PLS <- stack(feil_PLS)
stack_CPLS <- stack(feil_CPLS)
stack_full <- cbind(stack_PCR, stack_PLS, stack_CPLS)
index <- cbind(stack_full[2], stack_full[4], stack_full[6])
index3 <- cbind(index, index, index)
colnames(stack_full) <- c("PCR", "skjell", "PLS", "skjell", "CPLS", "skjell")
stack_full <- stack(stack_full)
stack_full <- cbind(index[1], stack_full)
colnames(stack_full) <- c("index", "feil", "metode")
```

```
# CVANOVA-modellen kjøres i R Commander med skjell som tilfeldig effekt
    og metode som fast effekt.

# Henter ut regresjonskoeffisientene fra den beste metoden
x_strek <- cpls.mod.log$Xmeans
y_strek <- cpls.mod.log$Ymeans
beta_hatt <- cpls.mod.log$coefficients[,,5]
```

Vedlegg C: Fettsyreprofiler

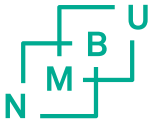
Tabell 2 Fettsyreprofiler for skjellprøver oppgitt som prosentandel for fettstren av den totale mengden fettsyrer.
Konsentrasjon av toksisitet (oppgitt i µg/kg Skjellmateriale), sted, art, uke og år prøven er tatt.

FAME	C12.0	C14.0	C13.0.4.8.12	C14.0.13.me	C14.0.12.me	C15.0	C15.0.14.me	C16.0	C16.1.t7	C16.0.15.me	C16.1.c9	C16.1.c11
B-1544 Langholmen	0,12467276	4,81769237	1,46783284	0,08209396	0,08738952	0,77904525	0,10371193	20,5631638	0,12747353	0	7,39089457	0,38383457
Ø-1542 Langholmen	0,05849427	4,93537078	0,6075451	0,39722129	0,09787943	1,21639202	0,27964416	21,7518594	0,16717449	0,68027464	2,99187555	0,65045358
Ø-1543 Langholmen	0,05227283	4,87553588	0,63617394	0,40142969	0,09013667	1,10659072	0,23932466	23,0920248	0,14747171	0,58312909	2,67864124	0,64651533
B-1543 Langholmen	0,1478046	5,20889352	1,68990728	0,08277311	0,07147614	0,76236487	0,09482115	20,8389799	0,11868585	0	7,27639417	0,3845121
B-1528 Flødevigen	0,06679741	3,00594862	0,51672865	0,06256791	0,10452222	0,74762178	0,15824374	23,5195035	0,12190519	0	9,80721144	0,41794514
B-1530 Flødevigen	0	3,45541338	1,09241991	0,08846147	0	0,74692616	0,13320444	22,3947646	0,13479465	0,44126596	6,24778239	0,43273302
B-1532 Flødevigen	0	3,2608108	1,16576489	0,07701846	0	0,72894299	0,13160888	22,3802943	0,14114298	0,49290139	7,06624089	0,36143145
B-1534 Flødevigen	0	3,25261652	1,05644165	0,07418271	0	0,80707411	0,13770906	21,6371614	0,13427393	0,47254385	6,266311	0,38749933
B-1536 Flødevigen	0	2,84997841	1,33320118	0,09742918	0	0,9337524	0,16766447	20,0941088	0,14651611	0,62338638	4,22824967	0,43127887
B-1538 Flødevigen	0	3,13169447	2,28421474	0,16515414	0	0,8514164	0,13535425	21,3571953	0,11665251	0,61359634	3,92665682	0,42619312
B-1540 Flødevigen	0	2,94977408	2,25143571	0,14947419	0	0,80117022	0,14364264	20,7963914	0,1726622	0	4,40569072	0,45932176
B-1542 Flødevigen	0	4,15489618	2,29309347	0,14382705	0	0,92266243	0,13444589	21,7051272	0,13295563	0,48476523	4,56905113	0,41236833
B-1544 Flødevigen	0,06044437	5,16021793	2,11010363	0,10226818	0,10140976	0,73113298	0,11616834	19,9271023	0,13274209	0	6,12986277	0,37786663
B-1443 Rundhaugen	0,04468501	3,24244344	1,4433925	0,14278698	0,10519866	0,76174372	0,16852418	22,0994135	0,14590643	0	8,02496816	0,8657074
K-1206 Fluarholmen	0	3,3878953	1,1521691	0,17825979	0	1,11049046	0,20610297	23,3818877	0,21087411	0,73206409	2,0413687	0,62429848
B-1423 Viganeset	0,11105021	5,70243985	1,34607962	0	0,05787308	0,47299032	0,06509647	15,6447512	0,10395809	0	17,8691381	0,34771375
O-1423 Viganeset	0,06222206	5,11306562	2,52768992	0	0	0,33693342	0,08729524	21,7892991	0,07648512	0	15,2078985	0,36717406
K-1424 Dyrøy	0	6,84584545	1,49368274	0,0836136	0	0,6200214	0	24,4595539	0,1078575	0	21,3581951	0,45592259
B-1437 Kaldvellfjorden	0	4,84985978	1,17055391	0,07256026	0,0650306	0,75767248	0,11512329	20,6901025	0,1978025	0	8,31673779	0,50785214
Ø-1437 Kaldvellfjorden	0	5,25745417	0,59754542	0,39231092	0,08703173	1,15205224	0,2385969	24,1162477	0,19864345	0,63578512	2,33423491	0,66512872
O-1535 Viganeset	0,02199772	2,78719409	1,39877687	0	0	0,41388501	0,09435668	19,0024211	0,0675432	0	11,6987073	0,35264165
B-1546 Flødevigen	0,07259395	4,63322198	2,28020105	0,1062459	0,09849309	0,76469952	0,13236035	19,4176983	0,14224565	0,32332962	5,53090649	0,40003477
B-1548 Flødevigen	0	4,10486367	2,72820892	0	0	0,76617216	0	19,6554603	0,13135889	0,32619097	4,84537	0,37441885
B-1550 Flødevigen	0,06026928	3,59610267	1,85382251	0,01865027	0,09456087	0,77454725	0,13660631	19,0828986	0,17615477	0,3615387	6,17880497	0,43988563
Ø-1541 Langholmen	0,08416989	5,27389126	0,67066997	0,38822204	0,08458481	1,1193903	0,20003282	22,5713921	0,15066668	0,57490737	3,06056518	0,63271661
Ø-1539 Rogøysund	0	4,42431975	0,51606724	0,29823463	0,08401588	1,13277036	0,23504839	21,6014906	0,67325412	0,56812557	1,83864022	0,31762873
B-1540 Møriholmen	0	3,5408708	1,27320396	0,07387786	0,05775284	0,649502	0,14489988	19,0368769	0,22401006	0	9,00153462	0,41420129
B-1540 Kjørem	0,03132615	3,30969816	1,29114423	0	0	0,52132874	0,10521156	17,8887356	0,15686623	0	8,29126023	0,30354022
B-1540 Skulgambukt	0,0496191	4,30721482	1,61689976	0	0	0,5761558	0,10718854	16,8905555	0,12586148	0	13,6490889	0,3752382
O-1418 Bergen	0	4,80595959	2,1251879	0	0	0,4147013	0,11878303	22,7776964	0,09712901	0	15,4488833	0,33864513
K-1422 Horsvær	0	7,221108	1,0232704	0,07100462	0,08188163	0,65486856	0,06695932	22,3796066	0,10481622	0	21,1780591	0,41328612
K-1422 Svinøya	0	5,01954066	1,29239	0,10176644	0,08428713	0,77596792	0,10450653	22,8113459	0,16533992	0,35264262	8,9265024	0,47582392
O-1423 Bergen	0,04155331	4,56355902	1,78852953	0	0	0,43411714	0,11392224	23,032676	0,09063947	0	15,02868	0,37177737

C18.0	C18.1.c5.og.	C18.1.c9	C18.1.c11	C18.1.c13	C16.4.c6.9.1	C19.0	C18.2.c9.12	C18.2.c11.14	C19.1.t10	C18.3.c6.9.1	C20.0	C18.3.c9.12.	C20.1.c5
2,31908602	0,24228132	3,12488663	2,06573972	0,13628559	0	0	2,43641288	0,09829781	0,10139466	0	0	1,44145321	0,11033642
5,16382647	0	3,69225815	2,03213065	0,2001037	0	0,1180792	1,82291454	0,13207153	0	0,06607344	0,22798246	1,19176707	0
4,92436359	0,06151302	3,55079392	1,67886597	0,19916592	0	0,10290982	1,88322415	0,10863793	0	0	0,21881936	1,20418712	0
2,39781034	0,19092794	2,82569084	1,88171266	0,1203736	0,11224924	0	2,23558432	0	0,08081834	0	0,08725411	1,31232611	0
4,10156907	0,14217971	2,58924477	2,16789077	0,16751127	0	0	1,48971593	0,15970321	0,13005293	0	0	1,30905721	0
3,48207954	0,20752991	2,50716821	2,28083866	0,17869702	0	0	1,9152923	0,14855952	0,09539385	0	0	1,55338969	0
3,40230524	0,22373174	2,39661243	2,28455944	0,15360551	0	0	1,58733604	0,14116723	0,09930833	0	0	1,2107732	0
3,55037847	0,22781603	2,2524762	2,33421147	0,17174356	0	0	1,70317909	0,15856855	0,11997628	0	0	1,45210633	0
4,10748721	0,17681272	2,12012807	2,15824109	0,1500954	0	0	1,73570081	0	0,16718538	0	0,14455838	1,45148058	0
3,46752913	0,12452642	2,35652599	2,19794102	0,09067326	0	0	1,53851637	0	0,22432564	0	0	1,64994287	0,15081789
3,0888844	0,15371435	2,65531425	2,30780821	0,10856756	0	0	1,73007938	0	0,19254061	0	0,11118439	1,89226847	0
2,81556314	0,17718501	2,95610644	2,02102172	0	0	0	2,1273368	0	0,14819384	0	0	1,65007812	0
2,32816389	0,1945981	3,58194242	1,90782006	0,09642824	0	0	2,73278966	0	0,1106047	0	0,10527425	1,35928077	0,17165406
2,31978795	0,23212522	2,80374432	2,59993641	0,26717048	0	0	2,0617064	0,09836984	0,11606415	0	0	1,24110475	0
8,84436838	0	4,13856744	2,71815099	0	0	0	1,28734824	0	0	0	0,20794753	1,41872718	0
2,95392411	0,13503548	1,90831465	2,8490799	0,14521819	0,48780015	0	1,23913644	0,22556682	0,09270366	0,09216854	0	0,76310266	0
3,77431509	0	1,04619811	4,69747581	0,13231516	0,52153645	0	2,01096567	0,36004832	0	0,12942821	0,06904644	1,04930662	0
4,64569526	0	1,77927703	5,20010433	0,10311804	0,48375531	0	2,77917761	0,22285619	0	0,16815175	0	0,79387664	0
3,24567124	0,27555754	3,11857404	2,50166808	0,20479146	0	0	3,05327346	0,09931324	0,1048899	0	0	1,87656035	0
5,38091266	0	4,23715357	1,71013922	0,18170034	0	0,13896664	2,51168544	0	0	0	0,15899925	2,04299611	0
3,80614048	0,1072995	1,07744197	6,89015469	0,17917125	0,20715682	0	1,36214873	0,51965084	0	0	0,09421425	0,99432649	0
2,47692842	0,17262982	3,27218749	1,90544458	0,09493101	0	0	2,50584899	0	0,10243208	0	0,08967145	1,29850372	0,13306311
2,68327104	0,13868053	3,17566378	1,77737569	0	0	0	2,35581943	0	0	0	0	1,15313041	0
2,25683441	0,20449846	3,32620324	2,0848533	0,13480334	0	0	2,49093804	0,08544963	0,13508877	0	0,06816283	1,40656831	0,10810077
5,02210909	0,06462331	3,46101293	1,85858407	0,19343663	0,07009335	0,09907236	1,86489786	0,11957442	0	0,06282377	0,20281654	1,34853583	0
6,02667071	0	4,58056641	1,36190212	0,28283326	0	0,14683663	1,76886988	0	0	0	0,19002889	1,53624815	0
2,7736514	0,25652089	2,90725357	2,8906652	0,21358784	0,0826404	0	2,21415645	0,18242275	0,08302057	0,06516525	0	2,90669982	0
2,69981184	0,18394927	1,97028814	2,87006983	0,13588023	0,08269147	0	1,73020528	0,1845521	0	0	0	2,58417011	0
2,4798296	0,17257869	1,88860888	2,73644561	0,14851275	0,37625355	0	1,28340701	0,18832379	0	0	0	1,25863445	0
3,82773673	0,07600461	1,14343987	3,65547035	0,11434098	0,47821626	0	2,64401951	0,31504519	0	0,14669627	0,07080389	0,91782047	0
4,2543107	0	2,21369863	3,67753868	0,08230739	0,27983617	0	3,02025485	0,12253922	0	0,16173877	0,06835889	1,00235929	0
5,66532088	0	2,9380052	3,52318097	0,2100139	0,24883289	0,06365183	2,21852157	0,21295423	0	0,08301641	0,06336597	1,44443594	0
4,2961007	0,09055461	1,00130223	4,19905673	0,13669626	0,36274374	0	2,54560187	0,36180638	0	0,15045535	0,10159224	0,91867464	0

C20.1.c9	C20.1.c11	C20.1.c13	C18.4.c6.9.1	C20.2.t8.11	C20.2.t11.14	C20.2.c11.13	C20.2.c11.14	C20.3.c5.11.	C20.3.c8.11.	C22.0	C20.4.c5.8.1	C22.1.c13	C20.4.c8.11.
1,18134862	2,56443075	0,69809474	4,48148416	2,16440182	0,43659964	0	0,74299342	0,45871479	0,05722378	0	1,20054726	0	0,20001082
1,09422291	0,5533908	4,64617289	2,52877976	0,08378906	0,15029068	0	0,36575067	0	0,13471167	0,10262168	1,87781826	0,08878165	0,49679795
1,05341728	0,57244939	4,15782461	2,51331602	0,08519031	0,15647328	0	0,39214083	0	0,10706058	0,09144856	1,5007629	0,08094068	0,4977439
1,07979939	2,3188113	0,64554878	4,37109793	1,873631	0,38908965	0	0,65642567	0,36739082	0	0	1,19625335	0	0,2004036
1,1738076	3,60786407	0,98543905	3,05928969	2,16938157	0,4317483	0	0,70902508	0,39311657	0,06118487	0	1,78733607	0,04805558	0,26300775
1,49044122	3,12708248	0,63326588	2,392385	2,44937664	0,9413169	0	0,73478452	0,41193733	0	0	1,58094445	0	0,26183622
1,46471978	3,150075	0,89234242	2,49452788	2,16207785	0,44372814	0	0,64761081	0,34044977	0	0	1,49305558	0	0,22054669
1,73946858	3,14235503	0,91339173	2,82601892	2,46501689	0,53182683	0	0,70683025	0,408299	0	0	1,56252769	0	0,24705598
2,42547839	4,16124874	0,94244175	1,99521549	3,06641875	0,67434932	0	0,74494127	0,31468626	0	0	2,842162	0	0,17587197
1,91454501	3,63686032	0,64434619	2,91408535	2,50948537	0,57471837	0	0,51013255	0,27234001	0	0	2,79225903	0	0,14228741
1,55495966	3,17477611	0,63747217	3,35088107	2,54236524	0,61403351	0	0,54512739	0,32284951	0	0	2,28856109	0	0
1,4095565	2,96730544	0,48771329	3,87322124	2,61547129	0,53034274	0	0,51296127	0,38381132	0	0	1,93224212	0	0,1507688
1,21177436	2,62795711	0,48562017	5,30796898	2,69509803	0,48966811	0	0,63025614	0,58455445	0	0	1,37720627	0	0,16348772
1,26271717	3,01380719	0,95848933	3,556652	2,21842292	0,57263265	0	0,82912676	0,47797084	0,05298634	0	1,36214758	0	0,21616686
0,24845911	1,91522936	0,43113432	3,46798045	0,3177158	0,04299157	0	0,55897595	0	0	0	4,05285742	0	0,43108195
0,81731338	2,58810014	1,52695276	2,618497	1,76364909	0,65320031	0	0,53411202	0,18376963	0,11725327	0	1,3734693	0,09484708	0,3544183
0,52032319	1,87997366	2,74416536	3,78360161	0,39612647	0,25281811	0,11079	0,22945064	0	0,08971287	0	1,01042747	0,09076733	0,33002774
0,19035976	0,80943303	1,15229741	3,09944266	0,23667798	0,18003243	0	0,65554992	0	0,10323695	0	0,8866621	0	0,38528476
1,61618318	3,34723846	0,85588908	3,28284457	2,58890195	0,56237575	0	0,92148805	0,49395469	0	0	1,19304874	0	0,21954334
1,46412577	0,74425772	4,05757551	2,5614505	0,09799539	0,12417888	0	0,40026576	0	0,10574448	0	1,91445403	0	0,49121531
1,32886439	2,25839269	3,81973749	2,02592394	0,57297469	0,61169581	0,26826892	0,23711871	0	0,0864383	0	1,26966769	0,10396797	0,35477396
1,19994321	2,81173637	0,50945468	5,31042112	2,74586914	0,46944632	0	0,63624085	0,52982613	0	0	1,49920755	0	0,18097118
1,14662373	3,11713531	0,4293866	4,19286571	3,09823798	0,49372041	0	0,58795797	0,56327366	0	0	1,52342712	0	0,15485522
1,16668484	2,65043714	0,54258467	4,60018204	2,79622558	0,5213911	0	0,68560191	0,64184057	0	0	1,2439033	0	0,20124788
1,19922911	0,61496204	4,21172648	2,49286256	0,10545228	0,16777371	0	0,38116982	0	0,10496043	0,08685146	1,51954366	0,08442566	0,49431413
0,9361863	1,22472421	4,66461913	2,90188828	0,12471896	0,11768577	0	0,50555382	0	0,10080174	0	1,72231869	0,11600048	0,81440854
1,09053819	2,83878237	1,04709287	4,57986719	2,03198033	0,6508945	0	0,7672558	0,34603132	0,08797297	0	1,88381953	0	0,46544627
0,89458626	2,41026238	1,23093814	5,15531763	1,78922787	0,64803477	0	0,72854653	0,30726795	0,07116776	0	1,76084924	0	0,34685788
1,07806262	2,63223362	0,95744178	3,70215036	2,45131943	0,68176916	0	0,45135939	0,23027512	0,08688652	0	1,59113594	0,28931851	0,32572822
0,82754837	2,37508243	2,29892394	3,26018899	0,4364315	0,27957223	0,12495402	0,2467499	0	0	0	1,27539833	0	0,28537543
0,14174659	1,00941718	0,67503878	3,77640055	0,23470716	0,09427215	0	0,53098642	0	0,10025802	0	1,69649491	0,08082086	0,39173121
0,16860372	1,46949771	0,76203015	5,9853755	0,27074347	0	0	0,76483231	0	0,11092038	0	1,94413558	0,10252837	0,65598309
0,85281819	2,17572012	2,54550139	3,0318895	0,41533517	0,30052713	0,14017031	0,27436136	0,0629334	0,07596746	0	1,31065501	0,07443004	0,32821315

C22.2.c5.13	C22.2.c13.16	C20.5.c5.8.1	C21.5.c6.9.1	C22.5.c4.7.1	C22.5.c7.10.	C22.6.c4.7.10.1	Total.toksin	Sted	Art	Aar	Uke	
0,38537647	1,63291045	11,4055154	0,53028554	0,28641909	0,73331793	22,22343549	798	Langholmen	Blaaskjell		15	44
0,69838417	3,75035392	13,2737825	0,62464505	0,43932082	0,37182123	18,40939461	28	Langholmen	Stillehavsoes		15	42
0,73700285	3,30746647	12,9804484	0,5764364	0,39294542	0,7372809	20,00351116	57	Langholmen	Stillehavsoes		15	43
0,31831261	1,56847218	12,4644164	0,51791948	0,27124371	0,69946269	22,52133094	693	Langholmen	Blaaskjell		15	43
0,5043958	1,71165566	12,6923594	0,47850447	0,27223125	0,88425007	17,00252843	24	Floedevigen	Blaaskjell		15	28
0,45657949	1,71583376	11,8809478	0,48905844	0,2636946	0,88800961	21,91260892	53	Floedevigen	Blaaskjell		15	30
0,34797391	1,71519094	12,5254427	0,45537056	0,24896874	0,84524472	22,46502449	89	Floedevigen	Blaaskjell		15	32
0,4837416	2,02139587	11,6757754	0,49274912	0,27550625	0,90713317	22,50790708	90	Floedevigen	Blaaskjell		15	34
0,55865823	2,64549284	10,4371583	0,48077997	0,40547257	1,03678262	22,71750955	41	Floedevigen	Blaaskjell		15	36
0,38996087	2,27344072	10,1883511	0,52700278	0,40049303	0,84760475	23,56627077	984	Floedevigen	Blaaskjell		15	38
0,42227032	2,06350753	10,3345628	0,56812951	0,37830786	0,8325749	25,05244849	1196	Floedevigen	Blaaskjell		15	40
0,41789589	1,96455192	9,25150826	0,53259507	0,29430136	0,71897424	24,28891022	865	Floedevigen	Blaaskjell		15	42
0,50463405	1,78292124	8,16752323	0,57993515	0,23092633	0,71044153	24,31969508	1576	Floedevigen	Blaaskjell		15	44
0,34654626	1,75815607	10,1712112	0,43826925	0,30603196	0,6666579	22,40454252	5037	Rundhaugen	Blaaskjell		14	43
0	0	9,42351941	0,82472998	0,78685742	0,66033898	23,04072108	12	Fluarholmen	Kamskjell		12	6
0,27466775	2,06774644	20,6502109	0,51794707	0,21308015	0,71319136	9,82288004	54	Viganeset	Blaaskjell		14	23
0,35363448	1,33231136	18,6451424	0,57110975	0,16760528	0,61813863	7,008945014	99	Viganeset	O-skjell		14	23
0	0	10,735796	0,54046434	0,1135199	0,35855787	7,835122306	42	Dyroey	Kamskjell		14	24
0,46019571	1,8279388	8,97073945	0,3738628	0,37303197	0,63220171	20,06467533	116	Kaldvellfjord	Blaaskjell		14	37
0,88865715	3,54972648	9,04136682	0,5203221	0,53936845	0,8452686	18,53357563	109	Kaldvellfjord	Stillehavsoes		14	37
0,76508962	2,71749618	21,410448	0,6712492	0,14864383	1,08327303	8,540990047	104	Viganeset	O-skjell		15	35
0,4824246	1,79508436	8,98927768	0,59527422	0,23304573	0,68313639	24,75417841	1203	Floedevigen	Blaaskjell		15	46
0,53035593	1,9541213	10,5364721	0,57539591	0,23015173	0,72144177	25,28516543	688	Floedevigen	Blaaskjell		15	48
0,50311565	1,68213558	9,93680717	0,43845987	0,25895749	0,73704479	25,64290929	1153	Floedevigen	Blaaskjell		15	50
0,76304421	3,4617584	12,8572639	0,55554232	0,38680499	0,71254293	19,06345577	18	Langholmen	Stillehavsoes		15	41
1,28137617	3,72146551	11,0788583	0,63423698	0,49480219	0,78708232	19,21216847	10	Rogoeysund	Stillehavsoes		15	39
0,25562628	1,93837753	15,3778096	0,4518055	0,28912913	0,65082312	15,47281573	4	Moeriholmen	Blaaskjell		15	40
0,20154373	1,81738415	17,6708314	0,61552951	0,24331372	0,91765949	18,2214242	2	Kjoerem	Blaaskjell		15	40
0,33181029	1,78494765	19,6615331	0,66883418	0,13423182	0,78810308	13,40534034	9	Skulgambukt	Blaaskjell		15	40
0,62132049	2,02070786	16,9974164	0,67402567	0,08820885	0,94176149	7,095791422	18	Bergen	O-skjell		14	18
0	0	9,73974069	0,54382794	0,19173592	0,41967278	11,49402999	1	Horsvaer	Kamskjell		14	22
0	0	11,3960139	0,76244482	0,27015422	0,45120957	16,65089326	1	Svinoeya	Kamskjell		14	22
0,5841012	1,79613913	16,7838776	0,71505478	0	0,9418683	7,28451523	19	Bergen	O-skjell		14	23



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

Postboks 5003
NO-1432 Ås
Norway