Måling af enzymatiske effekter med spektral billedanalyse

Kristian Ryder Thomsen

Kongens Lyngby 2013
B.Sc.-2013-10

Danmarks Tekniske Universitet
Institut for Matematik og Computer Science
Bygning 303B, 2800 Kgs. Lyngby, Danmark
Telefon 4525 3031, Fax 4588 1399
compute@compute.dtu.dk
www.compute.dtu.dk B.Sc.-2013-10

Resumé

Målet for dette bachelorprojekt er at undersøge, om multispektral billedanalyse kan benyttes til måling og dokumentation af enzymatiske effekter ifm. vask af lipidplettede stoflapper. Et multispektralt billede består af billeder taget ved mange forskellige bølgelængder.
Der er udviklet en metode til måling og dokumentation af enzymatiske effekter med spektral billedanalyse. Metoden er primært baseret på normalized canonical discriminant analysis, sekventiel fremadgående udvælgelse af attributter samt multiple lineær regression baseret på basisekspansioner. Den udviklede metode giver relative værdier for renheden af de målte emner og det er foreslået, hvordan metoden måske kan udvides til at give absolutte værdier.
Der er endvidere opnået lovende resultater ved at anvende metoden på glastallerkner.
Projektet er udført i samarbejde med Novozymes.

Forord

Dette bachelorprojekt er udarbejdet ved Institut for Matematik og Computer Science på Danmarks Tekniske Universitet i henhold til kravene for at erhverve en Bachelor of Science in Engineering, Mathematics and Technology.

Lyngby, 1. juli 2013

Kristian Ryder Thomsen

Tak

Tak til Novozymes, Technical Service, Household Care, afdeling 264. Det har været en fornøjelse at arbejde i en virksomhed med imødekommende og hjælpsomme kollegaer og et behageligt arbejdsmiljø.

Speciel tak til min eksterne vejleder Peter Klindt Mogensen for forklaring af Novozymes konkrete problemstilling og nuværende metoder til måling af renheden af stoflapper og tallerkner. Stor tak skal endvidere lyde til Connie Westergaard for den store hjælp ved vask af stoflapper og for at prioritere sin arbejdsdag efter mit bachelorprojekt, samt besvare utallige spørgsmål. Også tak til Dorte Steen Eskebæk for hjælp i forbindelse med mit arbejde med glastallerknerne. Sluttelig takkes Lone Poulsen fra R&D i Bagsværd for hjælp med intensitetsmålinger.

Tak til mine vejledere Jens Michael Carstensen (DTU), Anders Lindbjerg Dahl (DTU) og censor for at fremskynde mit bachelorforsvar så dette kan gennemføres inden sommerferien.

Forkortelser

CDA	Canonical discriminant analysis
CDF	Canonical discriminant function
FTIR	Fourier transform infrared spectroscopy
LOM	Launder-O-Meter
LOOKV	Leave-one-out krydsvalidering
MNF	Minimum noise fraction
MLR	Multiple lineær regression
nCDA	Normalized canonical discriminant analysis
PCA	Principal component analysis
SUA	Sekventiel (fremadgående) udvælgelse af attributter
SOP	Standard operating procedure

Indhold

Kapitel 1

Introduktion

Novozymes er en bioteknologisk virksomhed der handler med, samt forsker i udvikling og fremstilling af, enzymer og mikroorganismer til industriel anvendelse. I afdelingen Technical Service foregår bl.a. udvikling og fremstilling af demomateriale der benyttes ifm. salg og marketing. Dette demomateriale kunne f.eks. være stoflapper med kontrolleret påførte fedtpletter, som derefter er vasket med forskellige enzymer. Se f.eks. figur 1.2. Enzymkoncentrationen eller andre vaskeparametre varieres og renheden af det vaskede stof bedømmes visuelt eller maskinelt. Ved fremstilling af demomateriale har Novozymes mulighed for visuelt at lade en potentiel kunde se effekten af f.eks. bestemte enzymer ved tøjvask. Demomateriale kan således virke salgsfremmende. Ulempen ved visuel bedømmelse af demomateriale er, at bedømmelsen er subjektiv. Forskellige personer kan have forskellige holdninger til, hvornår noget er "rent" eller være uenige om, *hvor rent* f.eks. et stykke vasket stof er. For opnå større troværdighed er det derfor nødvendigt at ledsage demomaterialet med objektive, uafhængige og reproducerbare målinger.

Triglycerid (ofte blot kaldet fedtstof) er farveløst hvilket yderligere besværliggør visuel bedømmelse. At triglycerid er farveløst umuliggør ikke visuel bestemmelse, da stof (tekstil) mættet med triglycerid har andre fysiske egenskaber, hvilket gør, at der kan ses en farveforskel mellem det rene stof og stoffet mættet med triglycerid. Tilsvarende er f.eks. vand en farveløs væske, men hvis vand hældes ud over et stykke stof, kan der ses farveforskel mellem det tørre og det våde stof. Dette til trods for at selve vandet er farveløst.

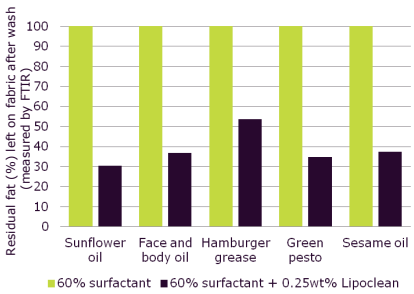
Hvis fedtpletter ikke fjernes fuldstændigt i vask, kan de via vaskevandet fordele sig til ellers rent tøj. Fedtrester i tøjet kan fungere som en slags lim og fastholde skidtpartikler, hvilket resulterer i at tøjet får gråligt udseende og bliver hurtigere

beskidt. Det er derfor også interessant at kunne måle forskel på renheden af f.eks. stoflapper, selvom det menneskelige øje ikke nødvendigvis kan se forskel i renheden af pågældende stoflapper. Figur 1.1 og 1.2 illustrerer hvor svært det kan være for det menneskelige øje at skelne renheden af fedtpletter.

Novozymes benytter i dag bl.a. farveintensitet og remission som mål for renheden af stoflapper. Remissionen måles med et spektrofotometer og farveintensiteten beregnes fra indscannede farvebilleder. På f.eks. hvidt stof giver måling af remission ofte gode resultater, der tillader f.eks. at separere stoflapper efter enzymkoncentrationer fra et dosis-respons forsøg. Ved måling af blå stoflapper benyttes farveintensitet. Dette giver brugbare resultater i nogle situationer og i andre overstiger måleusikkerheden forskellen på de enkelte målinger, hvorfor måleresultaterne bør forkastes som ikke konkluderbare.

Specielt ved blå stoflapper med lipidpletter[1] er det ofte ikke muligt at opnå statistisk signifikant brugbare måleresultater ved de før nævnte eksisterende metoder, dvs. remission og farveintensitet. Dette betyder at dokumentationen af enzymatiske effekter besværliggøres for blå stoflapper med lipidpletter.

Hvis man ønsker at sælge et produkt, er det altid ønskeligt videnskabeligt at kunne dokumentere effekten af produktet over for den potentielle køber. Multispektral billedanalyse har ikke tidligere været anvendt til måling af renheden efter vask. Jeg har derfor i dette bachelorprojekt i samarbejde med Novozymes undersøgt om multispektral billedanalyse kan benyttes som en let og praktisk udførlig metode til måling og dokumentation af enzymatiske vaskeeffekter.



Figur 1.1: *Målinger af renheden af demomaterialet vist på figur 1.2. Målingerne er foretaget med fourier transformeret infrarød spektrometri (FTIR). Novozymes kan ved FTIR dokumentere at enzymerne i Lipoclean har fjernet cirka 62% mere fedtstof end vasken uden Lipoclean. Tilsvarende observation kan ikke opnås visuelt.*
Figuren er udlånt af Novozymes.

[1]Lipid er en samlet betegnelse for fedtstof og fedtlignende kemiske forbindelser, der er hydrofobe, men opløselige i ikke-polære opløsningsmidler så som f.eks. benzen, æter og kloroform. Bl.a. triglycerider (populært kaldet fedtstoffer) og visse voksarter er lipider.

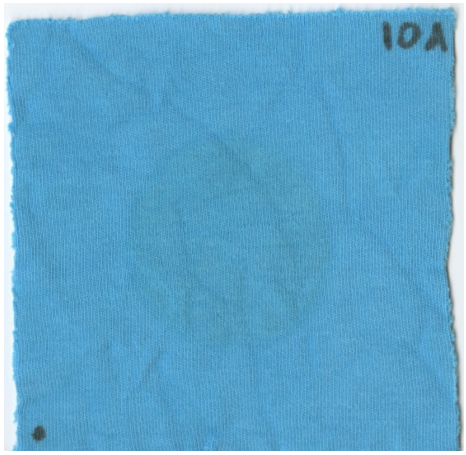
Figur 1.2: *Eksempel på demomateriale. En blå t-shirt er blevet klippet op i tre dele. Hver del er påført fem forskellige fedtstoffer. Hver af de tre dele af t-shirten er herefter vasket separat. Den første er vasket helt uden enzym og kun med surfactant. (Surfactant er overfladeaktive stoffer der sænker overfladespændingen af vaskevandet.) Den anden også uden enzym og kun med 60% surfactant. Den tredje del af t-shirten er vasket med 60% surfactant og med 0,25 vægtprocent af Novozymes enzymblanding Lipoclean. Visuelt er det meget svært at vurdere hvor meget af de påførte pletter, der er vasket af. Ved en FTIR udført af Novozymes er den enzymatiske effekt påvist at have fjernet cirka 62% mere fedtstof end vasken uden Lipoclean. Se figur 1.1.*
Figuren er udlånt af Novozymes.

1.1 Datamaterialet

Datamaterialet denne afhandling primært er baseret på, er multispektrale billeder[2] af stoflapper[3] af vævet bomuld påført kontrollerede mængder fedtstof. Stoflapperne er fremstillet af firmaet Warwick Equest og er cirka 10 gange 10 centimeter. Centreret på stoflapperne er en cirkulær fedtplet med en diameter på cirka 5 centimeter. For fedtpletter af typen lipid benyttes blå stoflapper. På figur 1.3 ses et eksempel en sådan stoflap. Disse stoflapper vaskes i et Launder-O-Meter (LOM) som simulerer en europæisk vaskemaskine. Parametre som vandtemperaturen, enzymkoncentrationen, detergentkoncentrationen og vandets hårdhedsgrad mm. kontrolleres og kan varieres afhængig af det konkrete forsøg. Efter vask lufttørrer stoflapperne og der foretages målinger af renheden senest 24 timer efter endt vask. At målingen foretages senest 24 timer efter vask er for at sikre, at målingerne er mest muligt reproducerbare, dog skal stoflapperne have haft tid til at tørre.

Tidligere blev benyttet grønne stoflapper, men med lanceringen af et nyt produkt blev der skiftet til blå stoflapper for at sikre at kunder ikke skulle tro at målingerne eller demomaterialet fra det gamle produkt blev genbrugt.

Der anvendes mørkt stof, frem for hvidt, da farveløse lipidpletter (fedtpletter) bliver mest synlige på mørkt farvet stof.



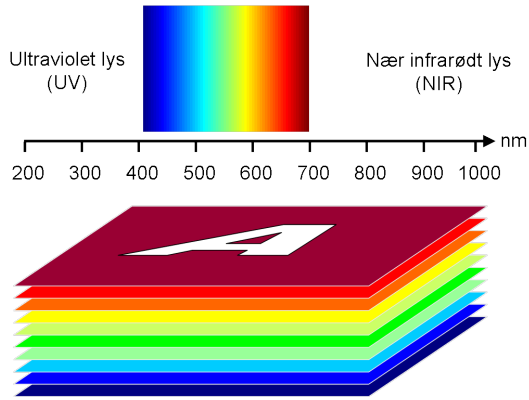
Figur 1.3: *Eksempel på vasket blå stoflap med lipidplet.*
Pletten kan stadig svagt ses cirkulært midt på tekstilet.

[2]Et multispektralt billede består af billeder taget ved mange forskellige bølgelængder. Dette er illustreret på figur 1.5.

[3]Datamaterialet udvides senere i projektet til også at inkludere multispektrale billeder af glastallerkner der benyttes til opvaskeforsøg. Se afsnit 4.9 på side 95.

Der er taget multispektrale billeder af stoflapperne ved brug af VideometerLab. VideometerLab er udviklet på DTU Informatik og kommercialiseret af Videometer. VideometerLab er bygget op af en Ulbrichtkugle, hvor der er monteret et CCD-kamera samt lysdioder langs ækvator, der gør det muligt at belyse prøven med 20 forskellige bølgelængder, benævnt spektrale bånd (herefter blot kaldet bånd). Ulbrichtkuglen er indvendig belagt med en hvid diffuserende belægning, der gør at den betragtede prøves belysning altovervejende kommer fra refleksioner på den indre kugleoverflade og ikke direkte fra lyskilderne. Dette giver en meget jævn og diffus belysning som er grundlaget for præcis farvemåling.
Pixelbredden er 72,5 μm og de 20 bånd er i intervallet 385nm til 1050nm: 385nm, 430nm, 450nm, 470nm, 505nm, 565nm, 590nm, 630nm, 645nm, 660nm, 700nm, 850nm, 870nm, 890nm, 910nm, 920nm, 940nm, 950nm, 970nm, 1050nm.

Figur 1.4: *VideometerLab, www.videometer.com.*



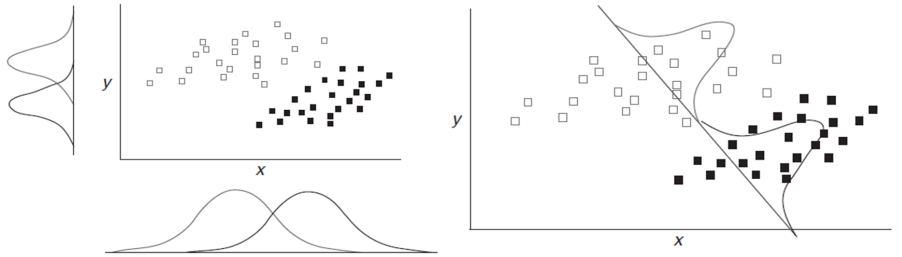
Figur 1.5: *Illustration af et multispektralt billede med 10 bånd i det synlige spektrum. www.videometer.com.*

Kapitel 2

Anvendt teori

2.1 Canonical discriminant analysis

Canonical discriminant analysis (CDA) er en måde at foretage klassifikation af observationer som hørende til en af k grupper. Princippet er, at man indlægger en ret linje (i to dimensioner), et plan (i tre dimensioner) eller et hyperplan (i fire eller flere dimensioner) og projekterer data ind på linjen/planet/hyperplanet, hvorved det bliver muligt at klassificere en ny observation blot ved at betragte de k marginale fordelinger. Figur 2.1 illustrerer princippet.
Figur 2.1 er en modificeret udgave af to figurer fra [BB09].



Figur 2.1: *Princippet i CDA illustreret for $k = 2$ grupper i 2 dimensioner. De marginale fordelinger af data projekteret ind på x- og y-aksen giver en meget dårlig diskrimination mellem de to grupper, da der er stort overlap mellem de marginale fordelinger. Hvis data i stedet projekteres ind på linjen, indlagt i plottet til højre, opnås en tydelig diskrimination mellem de to grupper, da de marginale fordelinger ikke overlapper.*

Vi betragter k grupper med $n_1, ..., n_k$ observationer i hver. Gennemsnittet af hver gruppe kaldes $\bar{X}_1, ..., \bar{X}_k$. De k grupper er malede, eller på anden måde udvalgte, repræsentative udsnit af de overordnede grupper.
Vi definerer between groups matricen

$$B = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})', \tag{2.1}$$

og within groups matricen

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)', \tag{2.2}$$

og total matricen

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})'.$$

Der gælder da fundamentalt, at $T = B + W$, hvilket betyder at variationen mellem de k grupper og variationen inden for de k grupper summerer til den samlede variation i data.
Med definitionen af B og W på plads, er vi nu klar til at gå i gang med selve klassifikationen. Vi søger den bedste diskriminant funktion, hvor bedste betyder at diskriminant funktionen skal maksimere forholdet mellem variationen mellem grupperne og variationen inden for grupperne.
Vi søger altså en funktion $y = d'x$ så

$$\varphi(d) = \frac{d'Bd}{d'Wd} \quad (d \text{ vælges så } d'Wd = 1)$$

bliver maksimeret. $\varphi(d) = \frac{d'Bd}{d'Wd}$ kaldes Rayleigh koefficienten. Fra lineær algebra, se f.eks. teorem 1.23 i [EC12], vides at den maksimale værdi af $\varphi(d)$ opnås når $d = d_1$, hvor d_1 er egenvektoren hørende til den største egenværdi λ_1 for det generaliserede egenværdiproblem

$$det(B - \lambda W) = 0 \quad \text{eller} \quad det(W^{-1}B - \lambda I) = 0.$$

Hvis vi har $k = 2$ grupper er vi færdige, og d_1 er linjen som data skal projekteres ind på for at opnå den bedste diskrimination ift. førnævnte betingelser. Har vi $k > 2$ fortsættes og der søges en ny diskriminant funktion d_2 så

$$\varphi(d_2) = \frac{d_2'Bd_2}{d_2'Wd_2}$$

maksimeres under bibetingelserne

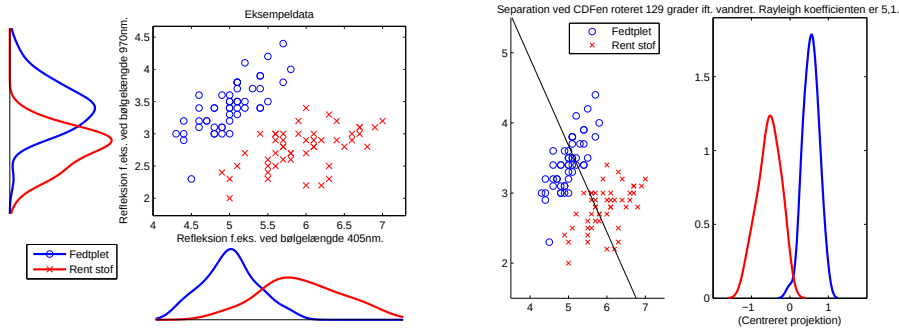
$$d_2'Wd_1 = 0 \quad \text{og} \quad d_2'Wd_2 = 1.$$

Dette svarer til den næststørste egenværdi for $W^{-1}B$ og den tilhørende egenvektor.
På denne måde kan fortættes indtil man enten opnår en egenværdi for $W^{-1}B$ som er nul eller indtil $W^{-1}B$ er udtømt.

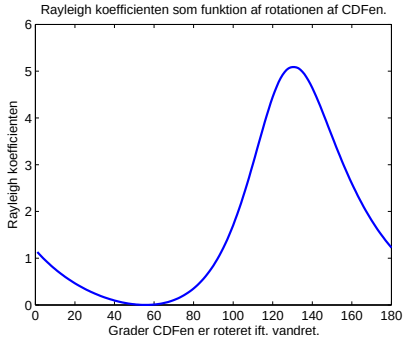
Funktionerne $d'x$ kaldes canonical discriminant functions (CDFs) og selve analysen hvor de beregnes kaldes canonical discriminant analysis (CDA).

Den første CDF defineret af d_1 er den afine transformation af de oprindelige variable, der giver den bedste diskrimination mellem de k grupper. En højere ordens CDF er den afine transformation af de oprindelige variable der giver den bedste diskrimination mellem de k grupper, under bibetingelsen at transformationen skal være ortogonal (ift. B og W) på alle lavere ordens CDFs. Bemærk, at det maksimale antal af CDFs er givet ved betragtninger af rangen af B og W. Hvis B og W har fuld rang er det maksimale antal af CDFs givet ved $min(k - 1, p)$, hvor p er dimensionen af data.

Et plot af værdierne $(d'_r(x_{ij} - \bar{x})\ ,\ d'_s(x_{ij} - \bar{x}))$ er i nogle situationer en nyttig måde at visualisere data. Disse plots separerer data bedst i betydningen forklaret herover, hvor variationen mellem grupperne er maksimeret og variansen inden for grupperne minimeret.
Et andet nyttigt plot består af vektorerne $(d_{11}, d_{21}), ..., (d_{1(k-1)}, d_{2(k-1)})$. Dette plot viser med hvilken vægt værdierne af hver enkelt variabel bidrager til plottet i (d_1, d_2)-planet.



Figur 2.2: *Til venstre ses et eksempel på to-dimensionalt data inddelt i $k = 2$ grupper. Der ønskes diskrimineret mellem disse to grupper. Det ses at både 1. og 2. aksen er dårlige CDFs, da de marginale fordelinger af data projekteret herpå, overlapper kraftigt.*
Til højre ses et plot af det samme data, den optimale CDF og de marginale fordelinger af data, projekteret på denne CDF. Bemærk, at de marginale fordelinger næsten ikke overlapper, og de 2 grupper derfor diskrimineres godt.



Figur 2.3: *Rayleigh koefficienten for en ret linje roteret i intervallet $\theta = [0, 180]$ grader ift. vandret. Data er det samme som i figur 2.2. Bemærk, at Rayleigh koefficienten antager maksimum på 5,1 ved 129 grader. Denne situation er illustreret til højre på figur 2.2.*

2.1.1 Normalized canonical discriminant analysis

Normalized canonical discriminant analysis (nCDA) er en udvidelse af CDA, der er givet ved følgende tilføjelser [CS]:

- Ved brutal force beregnes Rayleigh koefficienten for de malede (eller anden måde udvalgte) områder for de oprindelige bånd, logaritmen af de oprindelige bånd, og alle mulige bånd-normaliseringer. Den transformation af data der giver den højeste Rayleigh koefficient benyttes så fremadrettet.

- Gennemsnittet c af projektionen af på de malede områder $d'x_{malet}$ beregnes og $d'x_{malet}$ centreres.

- Ved beregningen af B vægtes grupperne ligeligt. Formel (2.1) bliver således $B = \sum_{i=1}^{k}(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})'$. I den oprindelige form af (2.1) vægtes med antallet af observationer i hver malet gruppe. Dette betyder, at hvis man vil have en gruppe til at vægte højere i analysen, så kan man blot male/markere flere pixels som tilhørende denne gruppe a priori.

- Projektionen $d'x_{malet}$ skaleres så det maksimale absolutte gruppegennemsnit er 1. Skaleringsfaktoren s gemmes.

- Egenvektoren d' orienteres så gennemsnittet af den første gruppe altid er positivt.

- På figur 2.1 og 2.2 på forrige side er CDA illustreret anvendt på en punktsky. Hvis CDA eller nCDA i stedet anvendes på et billede (hvilket principielt er det samme), er resultatet $d'x$ også et billede, som kaldes scorebilledet. For at holde fortolkningen af dette scorebillede konstant vises scorebilledet altid med fast skalering fra -2 til 2 med en blå-grøn-rød farveafbildning.

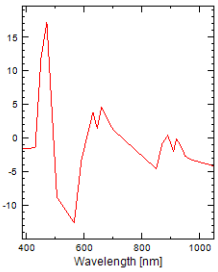
- Når CDA anvendes beregnes blot $d'x$, når nCDA anvendes beregnes $\frac{d'x-c}{s}$. CDFen centreres med konstanten c og normaliseres med konstanten s.

Fordelen ved at bruge nCDA frem for CDA er, at der ved nCDA ofte opnås en højere Rayleigh koefficient (grundet brutal force metoden beskrevet i første punkt herover), samt at orienteringen af egenvektoren er kontrolleret, hvilket betyder at f.eks. fedtpletten altid er gruppen med positivt gennemsnit.

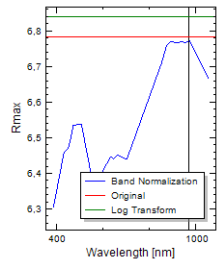
Figur 2.4 til 2.7 på side 13 viser nCDA anvendt på et billede af en lipidplettet (fedtplettet) stoflap. Figur 2.6 viser det multispektrale billede (ved bånd 11, 700nm) hvor der er malet et udsnit af det rene stof (rødt) og et udsnit af den

del af stoffet, som indeholder fedtpletten[1] (grønt). At projektere et multispektralt billede ind på et hyperplan svarer til at beregne en vægtet sum af båndene. De pågældende vægte er elementerne i egenvektoren, hørende til den største egenværdi for Rayleigh koefficienten. På figur 2.4 ses disse vægte. Figur 2.4 kan også tænkes på som værende CDFen med parallelle koordinater. Figur 2.5 viser Rayleigh koefficienten af de malede områder, for de oprindelige bånd, logaritmen af de oprindelige bånd, og alle mulige bånd-normaliseringer. Det ses at logaritmen af de oprindelige bånd giver den største Rayleigh koefficient. Figur 2.7 er derfor genereret ved at anvende CDFen på logaritmen af de oprindelige bånd. Det ses af figur 2.7 at det ved nCDA er muligt tydeligt at adskille fedtplet og rent stof. Bemærk, at dette blot viser at nCDA tydeligt kan *adskille* fedtpletten fra det rene stof, og ikke om nCDA kan benyttes til at skelne mellem forskellige mængder af residualfedt efter vask.

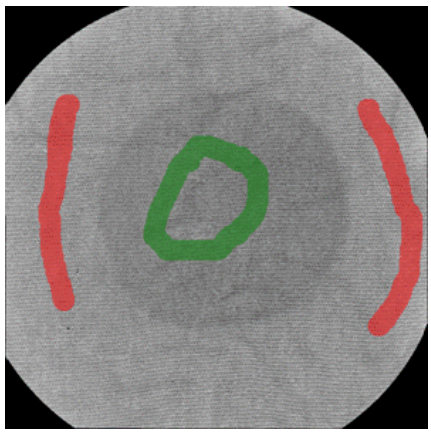
[1] Stoflappen er cirka 10 gange 10 centimeter og fedtpletten er cirkulær med en diameter på cirka 5 centimeter.



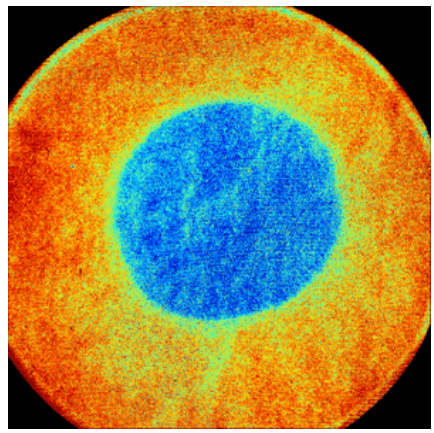
Figur 2.4: *CDFen.*



Figur 2.5: *Rayleigh koefficienten.*



Figur 2.6: *Det multispektrale billede (ved bånd 11, 700nm) hvor der er malet et udsnit af det rene stof (rødt) og et udsnit af den del af stoffet som indeholder fedtplet (grønt).*



Figur 2.7: *Scorebilledet $d'x$, hvor det multispektrale billede er projekteret ind på CDFen. De blå pixels er klassificeret som fedtplet, og de rød-orange som rent stof.*

2.2 Hotelling's T^2 test

Tilsvarende t-testen[2] fra univariat statistik kan Hotelling's T^2 test bruges til at undersøge om prøver fra to normalfordelinger, med samme varians-kovarians struktur, kan antages at have samme middelværdi. Vi betragter to uafhængige stokastiske variable $X_1, ..., X_n$ og $Y_1, ..., Y_m$, hvor $X_i \in N_p(\mu, \Sigma)$ og $Y_i \in N_p(\nu, \Sigma)$, og vi ønsker at teste

$$H_0 : \mu = \nu \quad \text{imod} \quad H_1 : \mu \neq \nu.$$

Vi benytter notationen [EC12]:

$$\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\bar{Y} &= \frac{1}{m}\sum_{i=1}^{m} Y_i \\
S_1 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' \\
S_2 &= \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})(Y_i - \bar{Y})' \\
S &= \frac{(n-1)S_1 + (m-1)S_2}{n+m-2}
\end{aligned}$$

Derved kan T^2 defineres til

$$T^2 = \frac{nm}{n+m}(\bar{X} - \bar{Y})' S^{-1} (\bar{X} - \bar{Y})$$

Den kritiske region for test af H_0 imod H_1 ved signifikansniveau α bliver da

$$C = \{x_1, ..., x_n, y_1, ..., y_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > F(p, n+m-p-1)_{1-\alpha}\}$$

hvor t^2 er den observerede værdi af T^2.
Nulhypotesen afvises hvis $\frac{n+m-p-1}{(n+m-2)p} t^2$ er større end $F(p, n+m-p-1)_{1-\alpha}$.

Hotelling's T^2 test kan f.eks. benyttes ifm. CDA for at teste om to malede, eller på anden måde udvalgte, områder er repræsentative for deres respektive grupper. Se evt. afsnittet om CDA på side 8.

[2]Teorien bag t-testen kan findes i næsten alle bøger om indledende statistik. Se f.eks. [JFM11].

2.3 Multiple lineær regression

Regression er en prædiktiv modelleringsteknik hvor den afhængige responsvariabel der søges estimeret er kontinuert. Eksempler på anvendelser af regression er forudsigelse af aktieindekset ud fra økonomiske indikatorer, forudsigelse af nedbør over et landområde baseret på karakteristika af luftstrømme og estimation af alderen af et fossil baseret på mængden af kulstof 14 der er tilbage i det organiske materiale.

Den generelle form af en lineær model i matrix notation er

$$y = X\beta + \epsilon,$$

hvor y (den afhængige responsvariabel) og ϵ er n-dimensionelle vilkårlige vektorer, X er en $n \times p$ regressionsmatrix af kendte konstanter (X er ofte også kaldet designmatricen) og β er en p-dimensionel parametervektor. Her er n antallet af observationer og p er antallet af parameter.
Spørgsmålet er, hvordan vi beregner et godt estimat for β, og når vi har beregnet dette estimat er næste spørgsmål, hvor godt estimatet endelig er.

Det optimale β skal approximere værdierne af den afhængige responsvariabel y så godt som muligt. Med andre ord, så skal residualerne

$$e_i(\beta) = y_i - (X\beta)_i$$

være mindst mulige.
Derfor søges efter det β der minimerer *summen af de kvadrerede residualer (SKR)*. Det er ikke tilstrækkeligt kun at minimere summen af residualerne, da to residualer med modsat fortegn da vil kunne udligne hinanden.

$$SKR(\beta) = \sum_{i=1}^{n}(e_i(\beta))^2 = \sum_{i=1}^{n}(y_i - (X\beta)_i)^2 = ||y - X\beta||_2^2$$

For at minimere summen differentieres $SKR(\beta)$ i forhold til β og differentialkvotienten $\frac{\partial SKR(\beta)}{\partial \beta}$ sættes lig 0.

$$SKR(\beta) = ||y - X\beta||_2^2 = (y - X\beta)'(y - X\beta) = y'y - 2\beta'X'y + \beta'X'X\beta$$

$$\frac{\partial SKR(\beta)}{\partial \beta} = -X'y + (X'X)\beta = 0$$

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{2.3}$$

Dette $\hat{\beta}$ kaldes mindste kvadraters estimat, det kan dog nemt vises at dette estimat er identisk med maximum likelihood estimatet, når fejlene ϵ er multivariat

normalfordelte med middel 0 og variansmatrix $\sigma^2 I$.

Likelihood for den lineære model er givet ved:

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi)^{(n/2)}} \frac{1}{\sigma^n} exp\left(-\frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma^2}\right).$$

Ved at maksimere denne likelihood opnås maksimum likelihood estimatet. Da det er nemmere at maksimere log-likelihood og dette giver det samme resultat [Mij10], arbejdes videre med

$$log(L(\beta, \sigma^2; y)) = l(\beta, \sigma^2; y) = -\frac{n}{2} ln(2\pi) - nln(\sigma) - \frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma^2}.$$

Differentieres $l(\beta, \sigma^2; y)$ i forhold til β opnås maximum likelihood estimatet (MLE) for β:

$$\frac{\partial l(\beta, \sigma^2; y)}{\partial \beta} = -\frac{1}{2}\frac{(-2X'y + 2X'X\beta)}{\sigma^2} = 0 \Rightarrow \hat{\beta} = (X'X)^{-1}X'y. \tag{2.4}$$

Det ses klart at MLE (2.4) og mindste kvadraters estimat (2.3) er identiske.

Vi mangler dog at gøre en antagelse. For at sikre at den inverse af $X'X$ eksisterer, antages at søjlerne i designmatricen X er lineært uafhængige og at antallet af variable p er mindre end eller lig med antallet af observationer n, så X har rang p. På baggrund af denne antagelse ved vi at:

$$rang(X'X) = rang(X).$$

Bevis. Da $X'X$ er en $p \times p$ matrix og derfor har samme antal søjler som X, vides at:

$$rang(X) + dim(ker(X)) = p = rang(X'X) + dim(ker(X'X)),$$

så hvis vi kan vise at dimensionen af kernen af $X'X$ er lig med dimensionen af kernen af X, så må der gælde at $rang(X'X) = rang(X)$. Lad b være i kernen af X, så $Xb = 0$. Så vides at $X'Xb = X'0 = 0$, så b er også i kernen af $X'X$. Derudover gælder at for alle vektorer b for hvilke $X'Xb = 0$, gælder også at $b'X'Xb = 0 = ||Xb||_2^2$, så b er i kernen af både X og $X'X$. Hvorved det ønskede er vist [Mij10]. □

Da $rang(X) = p$ ved vi at $rang(X'X) = p$ hvilket medfører at $(X'X)^{-1}$ eksisterer. Bemærk, at dette ikke holder hvis p bliver større end n. Altså hvis antallet af variable bliver større end antallet af observationer. I så fald skal benyttes en metode der anvender regularisering, f.eks. ridge regression, lasso eller the elastic net. Se f.eks. [HTF09].

Bemærk, at hvis der ved formel (2.3) beregnes et mindste kvadraters estimat af $\hat{\beta}$, hvor designmatricen X benyttes med de ovenfor givende dimensioner $n \times p$ så tvinges regressionsligningen gennem origo, i det regressionsligningen ikke indeholder et konstantled. Et konstantled kan inkluderes i modellen hvis der først tilføjes en søjle med 1-taller som første søjle i X, hvorefter mindste kvadraters estimat af $\hat{\beta}$ kan beregnes ved (2.3).

2.4 Krydsvalidering

Statistiske modeller benyttes inden for mange forskningsområder. Uanset hvilken konkret model der benyttes, vil det sidste skridt i modelleringen dog altid være en valideringsproces. I denne valideringsproces ønskes undersøgt om den genererede prædiktive model prædikterer ønskeligt, ikke kun på træningsdata, men også på et uafhængigt testsæt. Træningsfejlen for en model kan altid opnås vilkårligt lille, ved blot at inkludere flere variable i modellen. Derfor er det vigtigt at benytte et uafhængigt testsæt, da resultaterne opnået ved at træne og validere modellen på samme datasæt vil være for optimistiske.

Når uafhængig testdata ikke er opnåelig (hvilket oftest er tilfældet) og et estimat af den prædiktive nøjagtighed ønskes, benyttes ofte resampling af det originale data for at skabe et uafhængigt testsæt. Dette kan gøres på mange måder, men en af de mest brugte metoder kaldes krydsvalidering. De vigtigste krydsvalideringsmetoder kaldes leave-one-out krydsvalidering og k-fold krydsvalidering.
Ved leave-one-out krydsvalidering (LOOKV) benyttes hver observation én gang som testsæt. Modellen beregnes så n gange på et træningssæt bestående af $n-1$ observationer og testes hver gang på det éne datapunkt, der udgør testsættet. Der kan så benyttes et gennemsnit af resultaterne af de n tests som det endelige estimat af modellens prædiktive nøjagtighed. En fordel ved denne metode er, at hvert af de n træningsæt kun indeholder én observation mindre end det potentielle træningssæt (det fulde datasæt) og derfor vil det opnåede estimat af modellens prædiktive nøjagtighed ligge tæt op ad faktiske nøjagtighed. En ulempe ved denne metode er, at fremgangsmåden er tidskrævende, da der skal beregnes $n-1$ modeller.
En mere generel udgave af LOOKV er k-fold krydsvalidering. Ved denne metode inddeles data i k (næsten) lige store dele. Efterfølgende beregnes modellen ved at benytte $k-1$ dele af datasættet og testes på den sidste del. For små værdier for k vil store dele af det potentielle træningsdata (det fulde datasæt) ikke blive benyttet til at beregne modellen og den estimerede prædiktive nøjagtighed vil derfor formentlig være dårligere end den faktiske prædiktive nøjagtighed. Dette problem mindskes når k vælges større.

2.5 Effektiv leave-one-out krydsvalidering

Ved leave-one-out krydsvalidering, beregnes en model på baggrund af et træningssæt bestående af alle på nær ét datapunkt. Modellens testfejl beregnes så på baggrund af det datapunkt der blev udlagt af træningen. Dette gentages for alle n datapunkter. Dette betyder at der skal beregnes n separate modeller. Én model for hvert af de n datapunkter. For lineær regression er det dog muligt at opnå alle leave-one-out testfejlene kun ud fra én modelberegning. Uafhængigt af n. I dette afsnit gennemgås teorien bag.

Sherman-Morrison-Woodbury teoremet (som givet i [Hag89]) er nødvendigt i de efterfølgende udledninger, så lad os begynde med at opskrive teoremet og dettes bevis.

Teorem 1 *Sherman-Morrison-Woodbury teoremet*
Lad A være en ikke-singulær $p \times p$ matrix, og u og v være to p-dimensionale søjlevektorer. Så gælder at

$$(A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 + v'A^{-1}u}.$$

For at bevise dette teorem skal det vises, at $(A+uv')Y = Y(A+uv') = I$, hvor Y er højresiden af Sherman-Morrison-Woodbury teoremet. I beviset udnyttes det faktum at $v'A^{-1}u$ blot er en skalar.

Bevis.

$$\begin{aligned}(A+uv')\left(A^{-1}-\frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}\right) &= AA^{-1}+uv'A^{-1}-\frac{AA^{-1}uv'A^{-1}+uv'A^{-1}uv'A^{-1}}{1+v'A^{-1}u}\\ &= I+uv'A^{-1}-\frac{uv'A^{-1}+uv'A^{-1}uv'A^{-1}}{1+v'A^{-1}u}\\ &= I+uv'A^{-1}-\frac{(1+v'A^{-1}u)uv'A^{-1}}{1+v'A^{-1}u}\\ &= I+uv'A^{-1}-uv'A^{-1}\\ &= I\end{aligned}$$

og

$$\begin{aligned}\left(A^{-1}-\frac{A^{-1}uv'A^{-1}}{1+v'A^{-1}u}\right)(A+uv') &= A^{-1}A+A^{-1}uv'-\frac{A^{-1}uv'A^{-1}A+A^{-1}uv'A^{-1}uv'}{1+v'A^{-1}u}\\ &= I+A^{-1}uv'-\frac{A^{-1}uv'+A^{-1}uv'A^{-1}uv'}{1+v'A^{-1}u}\\ &= I+A^{-1}uv'-\frac{(1+v'A^{-1}u)A^{-1}uv'}{1+v'A^{-1}u}\\ &= I+A^{-1}uv'-A^{-1}uv'\\ &= I\end{aligned}$$

□

Når vi med formel (2.3) fra side 15 har fundet værdien af $\hat{\beta}$ baseret på et træningssæt, vil vi gerne teste modellen på et uafhængigt testsæt, men desværre er et uafhængigt testsæt ofte ikke tilgængeligt. Et estimat af den faktiske prædiktive nøjagtighed kan opnås ved leave-one-out krydsvalidering. Vi beregner modellen n gange på $n-1$ observationer og tester modellen på den éne observation, der er udeladt. Den dertilhørende sum af kvadrerede krydsvaliderede residualer er givet ved:

$$SKR_K = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i})^2, \tag{2.5}$$

hvor $\hat{\beta}_{-i}$ er givet ved (2.3), blot hvor X nu er en $(n-1)\times p$ matrix (fra nu af refereret til som X_{-i}) og element i af y er udeladt (y_{-i}). Bemærk K'et med sænket skrift i SKR_K.
For at beregne SKR_K er det nødvendigt at foretage n inverteringer, nemlig alle $(X_{-i}'X_{-i})^{-1}$ hvor i løber fra 1 til n. For at spare tid kan vi udnytte at alle disse inverse er meget ens, hvilket kan ses af Sherman-Morrison-Woodbury teoremet.

Lad A (i teorem 1) være $X'X$, lad u' være række i af X-matricen (herefter kaldet x_i) og lad $v = -u$. Så kan teorem 1 skrives om

$$(X'_{-i}X_{-i})^{-1} = (X'X - x_i x_i')^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1 - x_i'(X'X)^{-1}x_i}, \quad (2.6)$$

hvor $X'_{-i}X_{-i}$ er $X'X$ matricen hvor observation i er udeladt. For at se at $(X'X)_{-i}$ reelt er givet ved $X'X - x_i x_i'$ kan følgende omskrivning benyttes

$$X'X = \sum_{i=1}^{n} x_i x_i'.$$

Ved at gange ligning (2.6) med $X'y - x_i y_i$ fås:

$$(X'_{-i}X_{-i})^{-1}(X'y - x_i y_i) = (X'X)^{-1}(X'y - x_i y_i) + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}(X'y - x_i y_i)}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta}_{-i} = \hat{\beta} - (X'X)^{-1}x_i y_i + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}(X'y - x_i y_i)}{1 - x_i'(X'X)^{-1}x_i} \quad (2.7)$$

Som det er gjort i [Mij10] kan højresiden omskrives til

$$\hat{\beta} - \frac{(X'X)^{-1}x_i y_i(1 - x_i'(X'X)^{-1}x_i) - (X'X)^{-1}x_i x_i'(X'X)^{-1}(X'y - x_i y_i)}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i y_i - (X'X)^{-1}x_i y_i x_i'(X'X)^{-1}x_i - (X'X)^{-1}x_i x_i'(X'X)^{-1}(X'y - x_i y_i 1)}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i y_i - (X'X)^{-1}x_i x_i'(X'X)^{-1}(x_i y_i + X'y - x_i y_i)}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i y_i - (X'X)^{-1}x_i x_i'(X'X)^{-1}(X'y)}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i y_i - (X'X)^{-1}x_i x_i'\hat{\beta}}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i(y_i - x_i'\hat{\beta})}{1 - x_i'(X'X)^{-1}x_i}$$

$$\hat{\beta} - \frac{(X'X)^{-1}x_i e_i}{1 - h_{ii}}$$

Med denne omskrivning af højresiden af (2.7) har vi nu at

$$\hat{\beta}_{-i} = \hat{\beta} - \frac{(X'X)^{-1}x_i e_i}{1 - h_{ii}}, \quad (2.8)$$

hvor $e_i = y_i - x_i'\hat{\beta}$ og h_{ii} er diagonalelement i i hatmatricen H givet ved $H = X(X'X)^{-1}X'$.

Dette udtryk (2.8) kan substitueres ind i udtrykket for summen af kvadrerede krydsvaliderede residualer givet ved (2.5):

$$\begin{aligned} SKR_K &= \sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i})^2 \\ &= \sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta} + \frac{x_i'(X'X)^{-1}x_i e_i}{1-h_{ii}}\right)^2 \\ &= \sum_{i=1}^{n}\left(e_i + \frac{h_{ii}e_i}{1-h_{ii}}\right)^2, \end{aligned}$$

hvilket kan omskrives så vi i sidste ende får

$$SKR_K = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2. \tag{2.9}$$

Et logisk spørgsmål er nu om der er en måde at beregne denne sum (2.9) ved brug af matrix multiplikation.

Hvis vi kalder vektoren med prædikterede y-værdier baseret på den krydsvaliderede $\hat{\beta}$ for $\hat{y}_K$, så kan ligning (2.9) omskrives til

$$y - \hat{y}_K = (diag(I_n - H))^{-1}(y - \hat{y}).$$

som ved at benytte det faktum at $\hat{y} = Hy$ kan omskrives til

$$y - \hat{y}_K = (diag(I_n - H))^{-1}(I_n - H)y.$$

Da vi i sidste ende ønsker en ligning for summen af *kvadrerede* krydsvaliderede residualer, tages det indre produkt af den netop udledte vektor med sig selv.

$$\begin{aligned} SKR_K &= ((diag(I_n - H))^{-1}(I_n - H)y)'((diag(I_n - H))^{-1}(I_n - H)y) \\ &= y'(I_n - H)(diag(I_n - H))^{-2}(I_n - H)y \end{aligned} \tag{2.10}$$

Denne omskrivning gælder da $(diag(I_n - H))^{-1}$ og $(I_n - H)$ begge er symmetriske matricer. (Den første er en diagonal matrix og $(I_n - H)$ er symmetrisk da H er symmetrisk[3].)

[3] $H' = (X(X'X)^{-1}X')' = (X')'((X'X)^{-1})'X' = X((X'X)')^{-1}X' = X(X'X)^{-1}X' = H$

Vi har nu opnået en formel (2.10) hvormed vi momentant kan beregne summen af de kvadrerede leave-one-out krydsvaliderede residualer, når vi har et datasæt (og dermed hatmatricen), uden at skulle beregne modellen om og om igen, n gange.

Der opstår dog et problem når nævneren i (2.8) bliver 0. Altså når et eller flere af diagonalelementerne i hatmatricen $H = X(X'X)^{-1}X'$ bliver 1, for da bliver nævner i (2.8) $1 - h_{ii} = 1 - 1 = 0$. En situation, hvor dette vil opstå er, når designmatricen X er en invertibel kvadratisk matrix, altså når der er lige mange observationer n som parametre p. I det tilfælde er hatmatricen $H = X(X'X)^{-1}X'$ eksagt lig med enhedsmatricen.
Dette kan nemt bevises ved at vise at $(X'X)^{-1}X' = X^{-1}$. Da X er invertibel, så er X' det også, så der gælder $(X'X)^{-1}X' = X^{-1}(X')^{-1}X' = X^{-1}$ hvorved hatmatricen bliver $H = X(X'X)^{-1}X' = XX^{-1} = I$.
Fortolkningen heraf er at de prædikterede værdier for y er eksakt ens med de faktiske værdier for y. Dette giver mening da den oprindelige ligning $y = X\beta$ kan løses eksakt da X er invertibel.

Ovenstående metode til effektiv leave-one-out krydsvalidering er implementeret i Matlab og Matlabkoden er vedlagt i A på side 107.

2.6 Basisekspansioner

Mindste kvadraters estimat som det er gennemgået i afsnit 2.3 er et *lineært* mindste kvadraters estimat. For et lineært mindste kvadraters estimat gælder, at regressionsligningen der beregnes ikke behøver at være lineær i argumentet x, men kun skal være lineær i parametrene i β. Dette leder til ideen bag basisekspansioner: Udskift variablerne (kolonnerne) i designmatricen X (se evt. afsnit 2.3) med ulineære transformationer $h_i(X)$. Den lineære model bliver da

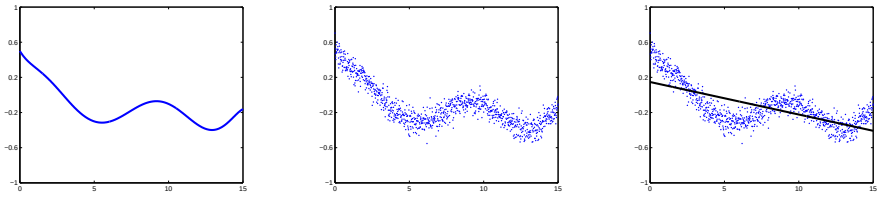
$$y = X\beta = \sum_{i=1}^{p} \beta_i x_i \longrightarrow y = h(X)\beta = \sum_{i=1}^{M} \beta_i h_i(X)$$

og det bemærkes at modellen stadig er lineær i parametrene β, hvorfor modellen kan beregnes ved et lineært mindste kvadraters estimatet. Det smarte ved basisekspansioner er, at det muliggør ulineær modellering af data, kun ved brug af lineære metoder som f.eks. mindste kvadraters estimat. Ved brug af basisekspansioner kan f.eks. en MLR model (husk at L'et i MLR står for lineær), benyttes til at modellere ulineært data. Det øgede antal parametre $(M - p)$ gør dog, at det ofte er nødvendigt at benytte variabel udvælgelse til at finde de bedst beskrivende variable, og derved minimere risikoen for at overtilpasse modellen. Se evt. afsnit 2.7.

Basisekspansioner kan f.eks. være:

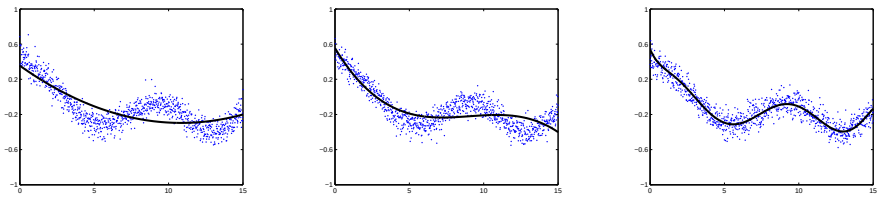
- Lineære
$$h_i(X) = \alpha + X_i, \quad i = 1, ..., p, \qquad \alpha \in \mathbb{R}$$
- Polynomier
$$h_i(X) = X_i^2 \quad \text{eller} \quad h_i(X) = X_i X_k$$
- Ikke lineære transformationer af enkelte variable
$$h_i(X) = log(X_i), \quad \sqrt{X_i}, \quad ...$$
- Ikke lineære transformationer af flere variable
$$h_i(X) = ||X||$$
- Indikatorfunktioner
$$h_i(X) = ind(L_i \leq X_i \leq U_i)$$

Figur 2.8 illustrerer en situation hvor en MLR model ikke kan beskrive data godt, uden indførslen af basisekspansioner. Figur 2.9 illustrerer brugbarheden af polynomier som basisekspansioner på det samme datasæt som i figur 2.8.



(a) (b) (c)

Figur 2.8: *Figur 2.8a viser den bekvemme sandhed - den bagvedliggende sandhed som vi prøver at modellere.*
Figur 2.8b viser den ubekvemme virkelighed - der vil altid være støj på vores målinger.
Figur 2.8c viser en MLR model af data beregnet ved et mindste kvadraters estimat.



(a) (b) (c)

Figur 2.9: *MLR modellen vist på figur 2.9a er en lineær model ift. modellens argumenter. Inden modellen blev beregnet er foretaget basisekspansionen $h_i(X) = X_i^2$, hvorved det blev muligt lineært at beregne en 2. ordens regressionslinje.*
Figur 2.9b og 2.9c viser tilsvarende situation, blot hvor der er benyttet basisekspansioner til henholdsvis et 3. ordens og 8. ordens polynomier.

2.7 Sekventiel udvælgelse af attributter

En attribut er et kendetegn for et objekt. Hvis man for eksempel har målt højden, vægten og hårlængden af to personer, siges at man har foretaget 2 observationer af 3 attributter. Har man målt mange attributter for hver observation kan det blive nødvendigt at vælge et udsnit af disse attributter inden man beregner f.eks. en MLR model over det indsamlede data. Det er dog afgørende hvor mange attributter der medtages i den endelige model. Medtages for få bliver modellen for simpel, medtages for mange bliver modellen for kompleks. Begge tilfælde er utilsigtede, da det medfører dårligere prædiktiv nøjagtighed.

Udvælgelse af attributter reducerer dimensionalitet af data ved at udvælge en delmængde af de målte attributter (prediktorvariable) som derefter benyttes til at beregne f.eks. en MLR model. Kriteriet for at udvælge en attribut er som regel minimering af en kostfunktion eller maksimering af et mål for modellens prædiktive nøjagtighed. Algoritmer søger efter en delmængde af attributter, der optimerer kriteriet under bibetingelser om f.eks. visse attributter enten skal indgå i modellen, ikke må indgå i modellen, eller krav til antallet af attributter udvalgt.

Udvælgelse af attributter er at foretrække frem for attribut transformationer der kombinerer attributterne (som f.eks. PCA[4]), i situationer hvor enhederne af de originale attributter er vigtige og målet med analysen er at finde frem til en delmænge af de *oprindelige* attributter der er beskrivende for data. Hvis man har kategoriske attributter og numeriske transformationer derfor er uhensigtsmæssige, er udvælgelse af attributter den primære måde at reducere dimensionen af data.

Det er almindeligt brugt at foretage udvælgelse af attributter *sekventielt*. Sekventiel udvælgelse af attributter (SUA) består af to dele:

- En objektfunktion, kaldet kriteriet, som SUA-metoden forsøger at minimere over alle realisable delmængder af attributter. Almindeligt brugte kriterier er den middel kvadrerede fejl (for regression) og fejlklassificeringsraten (for klassifikation).

- En sekventiel søgealgoritme, som tilføjer eller fjerner attributter fra en kandidat-delmængde, på baggrund af evalueringer af objektfunktionen. Da en udtømmende sammenligning af værdien af kriteriet for alle 2^n delmængder af et datasæt bestående af n attributter typisk ikke er en

[4]Principal component analysis, se f.eks. [EC12] eller [TSK06].

praktisk mulighed[5], benyttes en sekventiel søgning, der kun bevæger sig i én retning ved altid at tilføje (eller fjerne) én attribut og derved øge (eller mindske) den udvalgte delmængde.

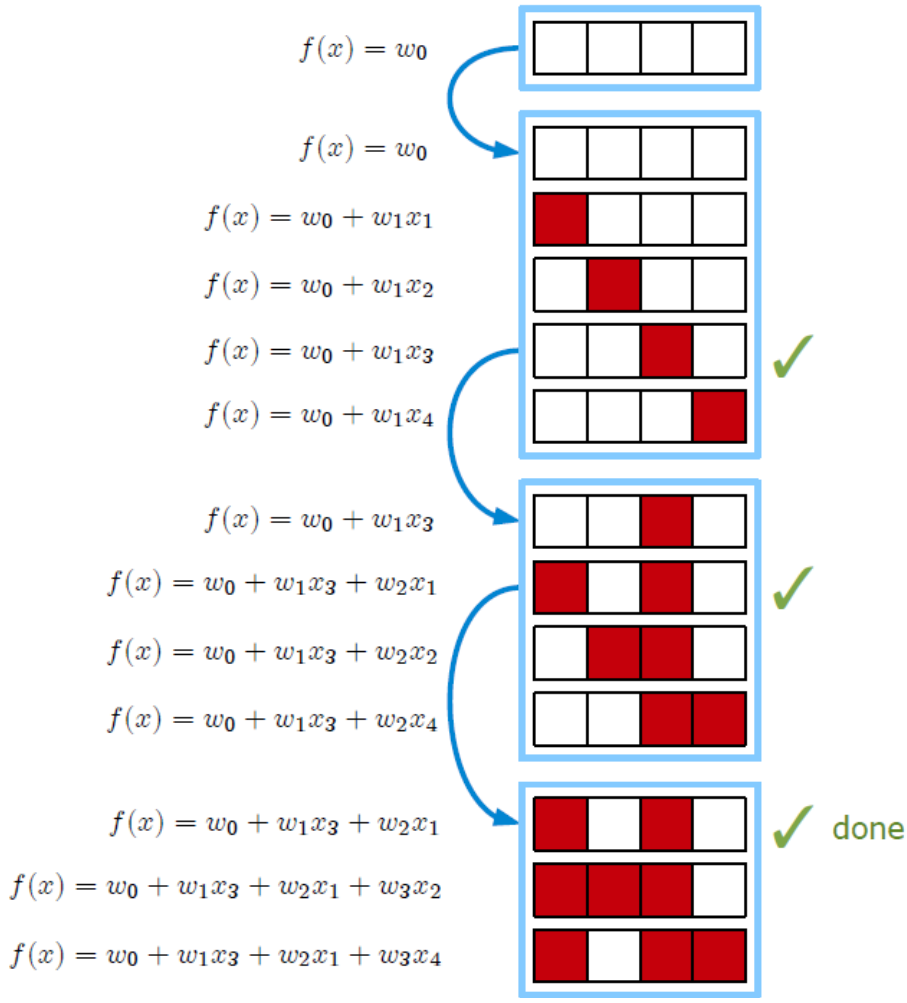
SUA findes i to varianter:

- *Sekventiel fremadgående udvælgelse,* hvor attributter sekventielt tilføjes til en fra starten tom delmængde, indtil tilføjelsen af flere attributter ikke længere mindsker kriteriet.
- *Sekventiel bagudgående udvælgelse,* hvor attributter sekventielt fjernes fra den fulde mængde af attributter, indtil fjernelse af flere attributter ikke længere mindsker kriteriet.

Figur 2.10 på den følgende side illustrerer princippet i sekventiel fremadgående udvælgelse af attributter.

Med mindre andet fremgår refererer SUA og formuleringen "sekventiel udvælgelse af attributter", fra dette punkt og fremad, altid til sekventiel *fremadgående* udvælgelse af attributter.

[5] Afhængig af størrelsen af n og omkostningerne forbundet ved evaluering af objektfunktionen. Bemærk, at ved blot $n = 20$ attributter er 2^n over en million kombinationsmuligheder.



Figur 2.10: *Princippet i sekventiel fremadgående udvælgelse af attributter. Her vist for en situation med fire mulige attributter, kaldet $x_1, ..., x_4$.*
Trin 1: Der tages udgangspunkt i den tomme mængde af attributter, dvs. at modellen kun indeholder et konstantled w_0.
Trin 2: Kriteriet evalueres for alle mulige kombinationer der kun indeholder kontantledet w_0 og én attribut. Hvis det f.eks. viser sig, at det at tilføje attributten x_3 giver det laveste kriterium, så vælges modellen i dette trin til $f(x) = w_0 + w_1x_3$.
Trin 3: Hvis det f.eks. viser sig, at tilføjelsen af attributten x_1 mindsker kriteriet yderligere, og mere end tilføjelsen af attributten x_2 eller x_4, så bliver modellen i dette trin $f(x) = w_0 + w_1x_3 + w_2x_1$.
Trin 4: Hvis kriteriet ikke kan mindskes yderligere, ved enten at tilføje attributten x_2 eller x_4, stoppes processen. I dette eksempel bliver den endelige model så modellen fundet i trin 3. Der er således fremadrettet udvalgt attributterne x_3 og x_1 som værende de attributter der bedst beskriver data ift. kriteriet.
Figuren, men ikke forklaringen, er fra [Sch12].

2.8 Hypotesetests for multiple lineær regression

Der findes tre typer af hypotesetests der kan udføres på multiple lineær regression (MLR) modeller også kaldet regressionsligninger.

- F-test for signifikansen af hele regressionsligningen.
- t-test for signifikansen af individuelle koefficienter i regressionsligningen.
- Partiel F-test for signifikansen af en gruppe af individuelle koefficienter i regressionsligningen.

Herunder vil t-testen blive gennemgået.

2.8.1 t-test for signifikansen af individuelle regressionskoefficienter

En t-test kan benyttes til at kontrollere signifikansen af de enkelte koefficienter i regressionsligningen i en MLR model. Tilføjes en betydelig variabel til en MLR model øger det MLR modellens prædiktive nøjagtighed, tilsvarende hvis der tilføjes en ubetydelig variabel kan det, og oftest vil det, mindske MLR modellens prædiktive nøjagtighed. Det er derfor interessant at kunne teste signifikansen af individuelle koefficienter i regressionsligningen, for at tjekke om alle variable bidrager signifikant til MLR modellen.
Nulhypotesen og den alternative hypotese for at teste signifikansen af koefficient β_j er:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Teststørrelsen for denne test er baseret på t-fordelingen [She09]:

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{2.11}$$

hvor $se(\hat{\beta}_j)$ er standardfejlen af $\hat{\beta}_j$ givet ved $\hat{\beta}_j = \sqrt{C_{jj}}$ hvor $C = \hat{\sigma}^2(X'X)^{-1}$. Kovariansmatricen C af de estimerede koefficienter i regressionsligningen er en symmetrisk matrix, hvis (hoved)diagonalelementer C_{jj} er variansen af den estimerede koefficient $\hat{\beta}_j$. Elementerne uden for (hoved)diagonalen er kovariansen

mellem de estimerede koefficienter $\hat{\beta}_i$ og $\hat{\beta}_j$. Værdien af $\hat{\sigma}^2$ er den middel kvadrerede fejl givet ved $MS_E = \frac{SS_E}{dof(SS_E)}$, hvor SS_E er summen af kvadrerede fejl, $SS_E = y'(I_n - H)y$, hvor y er vektoren med de afhængige variable, I_n er en $n \times n$ enhedsmatrix og $H = X(X'X)^{-1}X'$ er hatmatricen. Antallet af frihedsgrader $dof(SS_E)$ i SS_E, er givet ved $n - (p+1)$, hvor n er antallet af observationer og p er antallet af parametre i MLR modellen.

Nulhypotesen kan ikke afvises hvis teststørrelsen T_0 givet ved (2.11) ligger inden for det acceptable interval [She09]:

$$-t_{(1-\alpha/2,n-(p+1))} < T_0 < t_{(1-\alpha/2,n-(p+1))} \tag{2.12}$$

Hvis teststørrelsen T_0 givet ved (2.11) ligger uden for det acceptable interval (2.12), kan nulhypotesen afvises til fordel for den alternative hypotese H_1. Værdierne $-t_{(\alpha/2,n-(p+1))}$ og $t_{(\alpha/2,n-(p+1))}$ kaldes også for de kritiske værdier.
I naturvidenskabelige sammenhænge sættes signifikansniveauet α typisk til 5%, dvs. $\alpha = 0,05$. Nulhypotesen forkastes altså hvis kun hvis der er mindre end 5% risiko for at den er sand. I medicinske sammenhænge anvendes typisk et signifikansniveau på 1%, dvs. $\alpha = 0,01$ [Dal12]. Da falske negativer ifm. diagnosticering af sygdom oftest har langt større konsekvenser end falske positiver. Ved at benytte $\alpha = 0,01$ mindskes risikoen for falske negativer, tilgengæld for en øget risiko for falske positiver.
Denne t-test tjekker signifikansen af én variabel mens de ørige variable forbliver inkluderet i MLR modellen. Hvis der f.eks. i MLR modellen $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ testes for signifikansen af $\hat{\beta}_1$ så vil testproceduren tjekke signifikansen af at inkluderer variablen x_1 i MLR modellen der indeholder x_2 og x_3, altså at inkluder variablen x_1 i MLR modellen $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$.

2.8.2 Signifikansniveau og fejltyper

Bemærk, at selvom teststørrelsen T_0 (2.11) for nulhypotesen H_0 ligger uden for det acceptable interval (2.12), ved f.eks. $\alpha = 0,05$, betyder dette ikke nødvendigvis at nulhypotesen rent faktisk ér forkert. Der er to muligheder:

- Nulhypotesen er faktisk sand og vi har blot været "uheldige". Hvis nulhypotesen rent faktisk ér sand, vil vi jo stadig i $\alpha = 0,05$ af tilfældende få teststørrelser som ligger uden for 0,025- og 0,975-fraktilerne. (Hvilket jo også netop er definitionen af en fraktil.) I den virkelige verden sker der jo også usandsynlige ting, f.eks. kan det jo lade sig gøre at blive lottomillionær, til trods for at det ikke er særlig sandsynligt.
- Nulhypotesen er rent faktisk forkert og teststørrelsen (2.11) er derfor beregnet på et forkert grundlag, hvilket forklarer at teststørrelsen (2.11) ligger (evt. langt) uden for det acceptable interval (2.12).

Dilemmaet ved en signifikanstest er, at det ikke er muligt at afgøre, hvilken af de ovenstående to grunde der er den korrekte [JFM11]. Vi er derfor nødsaget til blot at antage at det er sidstnævnte forklaring, der er sand.

2.8.2.1 Type I fejl - forkastelsesfejl

Det valgte signifikansniveau er sandsynligheden for - fejlagtigt - at forkaste nulhypotesen, i de tilfælde hvor nulhypotesen reelt er sand. En sådan fejl kaldes en type I fejl eller forkastelsesfejl [JFM11].
Ved at justere signifikansniveauet kan det frit fastsættes hvor stor en risiko der kan accepteres for at lave en type I fejl. Dette leder til det naturlige spørgsmål: Hvorfor ikke bare altid vælge et meget lavt signifikansniveau, for at sikre at vi ikke - fejlagtigt - forkaster nulhypotesen?

2.8.2.2 Type II fejl - acceptfejl

Problemet ved at vælge et meget lavt signifikansniveau, og derved sikre en meget lille risiko for at begå type I fejl, er, at vi i stedet får øget risiko for at begå type II fejl. At begå en type II fejl betyder at acceptere nulhypotesen selvom nulhypotesen i virkeligheden er forkert. Type II fejl kaldes også for β-fejlen [JFM11].
Risikoen for at begå type I fejl er blot signifikansniveauet α, men det er ikke muligt at give et simpelt udtryk for risikoen for at begå type II fejl[JFM11].

I praksis er valget af signifikansniveau altså en afvejning af hvor stor en risiko der kan accepteres for at begå type I fejl henholdsvis type II fejl.

2.9 Eksempel

Det ønskes undersøgt om udbyttet y af en bestemt kemisk proces afhænger signifikant af de to variable x_1 og x_2. Der er foretaget samtidige observationer af udbyttet y og de to variable x_1 og x_2. På baggrund af kemisk teori og tidligere forsøg vides at der ikke er interaktion mellem de to variable x_1 og x_2, det blandede led x_1x_2 undersøges derfor ikke. En lineær regressionsmodel (se afsnittet om MLR på side 15) beregnes og herefter tjekkes ved t-tests for signifikansen af de individuelle koefficienter i regressionsligningen. De kemiske enhederne for x_1, x_2 og y ignoreres i dette eksempel og x_1, x_2 og y betragtes som enhedsløse tal. De målte værdier for x_1, x_2 og y kan ses i tabel B.1 i bilag på side 110.

Den ønskede regressionsligning, skrevet på vektorform, er

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Mindste kvadraters estimatet $\hat{\beta}$ for koefficienterne i regressionsligningen beregnes med formel (2.3), fra side 15, til

$$\hat{\beta} = \begin{pmatrix} -144,22 \\ 1,25 \\ 16,58 \end{pmatrix}$$

Matricerne X og y kan ses i bilag på side 109.
Den beregnede regressionsmodel er altså

$$\hat{y} = -144,22 + 1,25x_1 + 16,58x_2$$

Først undersøges signifikansen af *variablen* x_1, ved at teste om *koefficienten* ($\hat{\beta}_2$) foran x_1 er signifikant forskellig for nul. Nulhypotesen og den alternative hypotese er

$$H_0 : \ \hat{\beta}_2 = 0$$
$$H_1 : \ \hat{\beta}_2 \neq 0$$

For at beregne teststørrelsen T_0, givet ved (2.11), skal vi kende standardfejlen af $\hat{\beta}_2$, kaldet $se(\hat{\beta}_2)$, som er estimeret[6] ved $se(\hat{\beta}_2) = \sqrt{C_{22}}$ hvor $C = \hat{\sigma}^2(X'X)^{-1}$.

$$\hat{\sigma}^2 = \frac{SS_E}{dof(SS_E)} = \frac{y'(I_n - H)y}{n-(p+1)} = \frac{1214,5}{20-(2+1)} = 71,4$$
$$(X'X)^{-1} = \begin{pmatrix} 152,212 & 0,642 & -9,203 \\ 0,642 & 0,003 & -0,040 \\ -9,203 & -0,040 & 0,561 \end{pmatrix}$$
$$se(\hat{\beta}_2) = \sqrt{C_{22}} = \sqrt{71,4 \cdot 0,003} = 0,46$$

[6] Da $\hat{\sigma}^2$ er et estimat bliver $se(\hat{\beta}_2)$ også et estimat, da $se(\hat{\beta}_2)$ afhænger af $\hat{\sigma}^2$.

Teststørrelsen (2.11) kan nu beregnes

$$(T_0)_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{1,25}{0,46} = 2,72 \tag{2.13}$$

De kritiske værdier, med signifikansniveau $\alpha = 0,05$, udregnes til $\pm t_{(1-\alpha/2,n-(p+1))}$ $= \pm t_{(1-0,05/2,20-(2+1))} = \pm 2,11$ så det acceptable interval (2.12) bliver

$$-2,11 < (T_0)_{\hat{\beta}_2} < 2,11 \tag{2.14}$$

Da teststørrelsen (2.13) ligger uden for det acceptable interval (2.14), kan nulhypotesen H_0 afvises til fordel for den alternative hypotese H_1 ved signifikansniveauet $\alpha = 0,05$. *Koefficienten* $\hat{\beta}_2$ foran *variablen* x_1 er altså signifikant forskellig fra 0 ved signifikansniveauet $\alpha = 0,05$. Dette betyder, at *variablen* x_1 har signifikant betydning for udbyttet y ved signifikansniveauet $\alpha = 0,05$.

Signifikansen af variablen x_2 eller konstantleddet $\hat{\beta}_0$ kan tjekkes på tilsvarende måde.

2.10 Anscombe's kvartet

Anscombe's kvartet består af fire datasæt, der har næsten identiske simple statistiske egenskaber. Når de fire datasæt plottes ser de dog meget forskellige ud. Se figure 2.11 (a-d). Hvert datasæt består af elleve (x, y) punkter. Datasættet blev fremstillet i 1973 af statistiker Francis Anscombe for at demonstrere dels vigtigheden af plotte data, før dataet analyseres, og dels effekten af outliers på statistiske egenskaber [Ans73].
Hvis data er mangedimensionelt, er det ofte ikke meningsfyldt muligt grafisk at illustrere data før analyse. I stedet kan benyttes f.eks. residualanalyse.

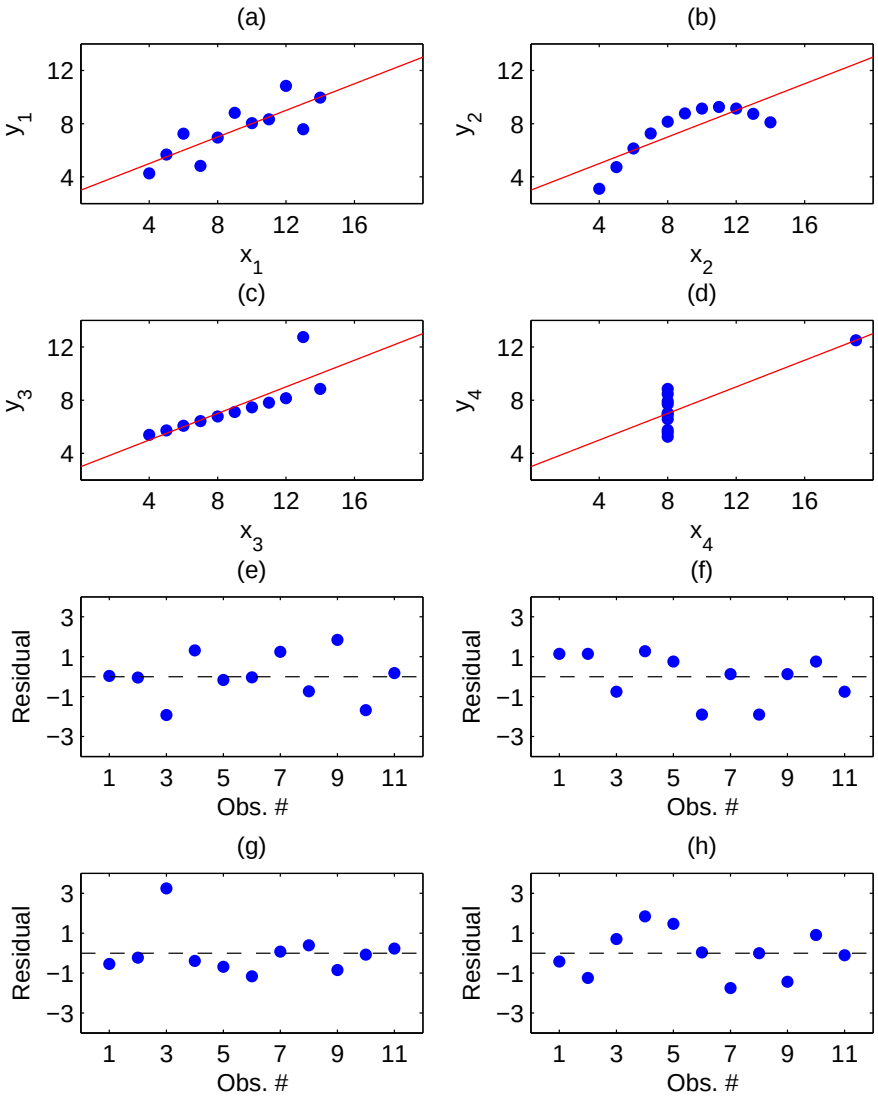
Tabel 2.1: *Alle fire datasæt i Anscombe's kvartet har samme simple statistiske egenskaber.*

Egenskab	Værdi
Gennemsnittet af x	9 (eksakt)
Variansen af x	11 (eksakt)
Gennemsnittet af y	7,50 (med 2 decimalers nøjagtighed)
Variansen af y	4,122 eller 4,127 (med 3 decimalers nøjagtighed)
Korrelationen mellem x og y	0,816 (med 3 decimalers nøjagtighed)
Lineær regressionslinjen	$y = 3{,}00 + 0{,}500x$ (til 2 henholdsvis 3 decimalers nøjagtighed)

På figur 2.11 (a-d) ses spredningsdiagrammer (eng: scatter plots) af de fire datasæt i Anscombe's kvartet. Ved det første spredningsdiagram (a) ser der ud til at være en simpel lineær sammenhæng mellem x og y. Ved det andet spredningsdiagram (b) ses en klar ikke-lineær sammenhæng mellem x og y. Det tredje spredningsdiagram (c) viser en situation hvor én outlier har nok indflydelse til at den beregnede lineære regressionslinje bliver løftet og drejet, og derved sænker korrelationskoefficienten fra 1 til $0,816$. Det fjerde spredningsdiagram (d) viser et eksempel, hvor blot én outlier er nok til at producere en høj korrelationskoefficient mellem x og y, selvom der tydeligvis ikke er en lineær sammenhæng mellem x og y.
Anscombe's kvartet benyttes ofte til at illustrere vigtigheden af at betragte data grafisk før der foretages analyse baseret på en antagelse om en given sammenhæng i data, samt til at illustrere unøjagtigheden af simple statistiske egenskaber til at beskrive realistiske datasæt [SW91]. Anscombe's kvartet datasættet findes bl.a. i [Ans73].
Da korrelation mellem x og y er ens for alle fire datasæt og vi arbejder med en lineær regressionsmodel, så er R^2 [7] (eng: coefficient of determination) også ens, og givet ved $\sqrt{0,816} = 0,666$, for alle fire datasæt [Rod]. Det betyder, at

[7] I statistik er R^2 et mål for hvor godt en model tilnærmer data [SW91].

den lineære regressionslinje, i følge R^2, tilnærmer alle fire datasæt lige godt. Ved grafisk at betragte data, se figur 2.11 (a-d), er det dog indlysende at den lineære regressionslinje *ikke* udgør en lige god beskrivelse af de fire datasæt. Det er derfor ikke nok kun at betragte R^2 når en regressionsmodel skal evalueres.



Figur 2.11: *Spredningsdiagrammer (a-d) (eng: scatter plots) af de fire datasæt i Anscombe's kvartet, samt plot af residualerne (e-h), $r = y - x'\hat{\beta}$, for den lineære regressionslinje $y = 3.00 + 0.500x$.*

2.10.1 Residualanalyse

Ved at plotte og analysere residualerne $r = y - x'\hat{\beta}$ kan man undgå at løbe risikoen for at blive snydt af en misvisende R^2 til at tro, at man har en retvisende model. På figur 2.11 (e-h) ses plot af residualerne for den lineære regressionslinje og de fire datasæt. I (e) ser residualerne ud til at være tilfældigt fordelt omkring 0 og der ses ingen trend i residualerne. Dette indikerer, at modellen er velegnet til at beskrive pågældende data. I (f) ses en anden ordens tendens i residualerne, hvilket indikere at modellen *ikke* er velegnet til at beskrive pågældende data, da modellen (som minimum) mangler et anden ordens led. I (g) ser residualerne ud til at være tilfældigt fordelt omkring 0, med undtagelse af observation nr. 3, og der ses ingen trend i residualerne. Dette tyder på at observation nr. 3 er en outlier og det er derfor anbefalelsesværdigt at teste for outliers, fjerne eventuelle outliers, og beregne modellen igen. I (h) ser residualerne *ikke* ud til at være tilfældigt fordelt omkring 0 og der ses noget der ligner en tredje ordens trend i residualerne. Dette indikerer, at modellen mangler et eller flere led for at beskrive pågældende data godt. Situationen i (d) (med residualerne (h)) er et eksempel på det statistiske begreb dominans, hvor ét punkt (det helt ude til højre i (d)) dominerer bestemmelsen af hældningen af den rette linje. Uden dette punkt ville det dog ikke være muligt at beregne en regressionslinje til pågældende data, da alle de resterende punkter har samme x-værdi.

2.10.1.1 Test af residualer

I det følgende gennemgås to statistiske tests til at undersøge om residualerne opfører sig tilfældigt eller indeholder en eller flere tendenser.

- **Test for tilfældige fortegn:** Tjekker for tilfældige skift af fortegn for residualerne.
- **Test for korrelation:** Tjekker om residualerne er ukorrelerede.

Disse og mange andre tests benyttes ofte i signalbehandling og tidsrækkeanalyse [HPS13].

2.10.1.2 Test for tilfældige fortegn

Den måske simpleste analyse af residualer er baseret på det statistiske spørgsmål: Kan vi betragte fortegnene for residualerne som tilfældige? Dette spørgsmål kan besvares ved en såkaldt *run test* fra tidsrækkeanalyse. Se afsnit 14.5 i [JFM11]. Givet en sekvens af to symboler - i vores tilfælde plustegn for positive og minustegn for negative residualer r_i - er et *run* defineret som en sekvens af samme symbol omringet af andre symboler. F.eks. har sekvensen "+ + + - - - - + + - - - - - + + +" $m = 17$ elementer, $n_+ = 8$ plustegn, $n_- = 9$ minustegn og $u = 5$ *runs*: + + +, - - - -, + +, - - - - - og + + +. Fordelingen af *runs* u (ikke residualerne) kan approximeres med en normalfordeling med middelværdi μ_u og standardafvigelse ς givet ved

$$\mu_u = \frac{2n_+n_-}{m} + 1, \quad \varsigma_u = \sqrt{\frac{(\mu_u - 1)(\mu_u - 2)}{m - 1}}.$$

Med et 5% signifikansniveau accepteres fortegnsrækkefølgen som tilfældig hvis

$$z_\pm = \frac{|u - \mu_u|}{\varsigma_u} < 1,96 \tag{2.15}$$

(andre værdier for den kritiske værdi, for andre signifikansniveauer, kan findes i enhver statistikbog. Se f.eks. [JFM11].) Hvis fortegnene ikke er tilfældige, så er der formentlig tendenser i residualerne. I eksemplet herover med 5 *runs* har vi $z_\pm = 2,25$ så i følge (2.15) er sekvensen af tegn ikke tilfældig. Fortegnede for residualerne som sekvensen repræsenterer, kan altså ikke antages at være tilfældige, hvilket indikere at der er trend i residualerne.

2.10.1.3 Test for korrelation

Et andet spørgsmål vi kan stille er, om en kort sekvens af residualer er korrelerede, hvilket ville være en klar indikation af trend i residualerne. Autokorrelationen[8] af residualerne er et statistisk værktøj til at analysere dette [HPS13]. Vi definerer autokorrelationen ϱ af residualerne og en grænse for trend T_ϱ som

$$\varrho = \sum_{i=1}^{m-1} r_i r_{i+1}, \quad T_\varrho = \frac{1}{\sqrt{m-1}} \sum_{i=1}^{m} r_i^2.$$

Da ϱ er summen af produkter af naboresidualer, så er ϱ faktisk enheds-forsinkelses autokorrelationen. Autokorrelationer med større forsinkelse, eller afstand i index, kan også betragtes. Vi siger, at trends sandsynligvis er til stede i residualerne, hvis absolutværdien af autokorrelationen overstiger grænsen for autokorrelation, dvs. hvis $|\varrho| > T_\varrho$.

[8] Autokorrelation er et mål for udviklingen i en tidsrække. Den k'te ordens autokorrelation er defineret som korrelationskoefficienten mellem værdier i tidsrækken med en tidsforskel på k tidsperioder.

2.11 Angivelse af måleusikkerhed

Når man efter udførelsen af et forsøg angiver resultatet, er det også vigtigt at angive hvilken usikkerhed der er på resultatet.
Ved et forsøg udført $n = 10$ gange har man f.eks. observeret resultaterne $\kappa_{obs} = \{14, 13, 16, 16, 15, 17, 15, 15, 14, 15\}$. Gennemsnittet $\bar{\kappa}_{obs} = 15$ benyttes som estimat for den sande værdi af κ, og der ønskes angivet et estimat af usikkerheden på $\bar{\kappa}_{obs} = 15$. Som estimat af usikkerheden kan benyttes enten standardafvigelsen af κ_{obs}, standardfejlen af $\bar{\kappa}_{obs}$ eller et konfidensinterval omkring $\bar{\kappa}_{obs}$.
I denne rapport benyttes primært standardfejl og konfidensintervaller ved angivelser af forsøgsresultater.

2.11.1 Standardafvigelse og standardfejl

I videnskabelig og teknisk litteratur er eksperimentielle data som oftest enten angivet med gennemsnit og standardafvigelse eller med gennemsnit og standardfejl. Det er vigtigt at huske forskellen. Gennemsnittet og standardafvigelsen er beskrivende statistik, der fortæller noget om den aktuelle stikprøve κ_{obs}, hvorimod gennemsnittet og standardfejlen beskriver grænser for den bagvedliggende fordeling af κ.

De to udtryk er defineret ved

$$\text{Standardafvigelsen:} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \Rightarrow s_{\kappa_{obs}} = 1,15$$

$$\text{Standardfejlen:} \qquad SF = \frac{s}{\sqrt{n}} \qquad \Rightarrow SF_{\bar{\kappa}_{obs}} = 0,37$$

hvor s er standardafvigelsen af de observerede værdier κ_{obs} og n er antallet af observerede værdier.

Standardfejlen, SF, af et gennemsnit ($\bar{\kappa}_{obs}$) er standardafvigelsen af det gennemsnits ($\bar{\kappa}_{obs}$) estimat af det sande gennemsnit ($\bar{\kappa}$) af den bagvedliggende fordeling. Der gælder således, at hvis forsøget gentages mange gange, og gennemsnittet af $n = 10$ nye observationer beregnes mange gange, så vil $68,27\%$[9] af de gennemsnit forventes at ligge i intervallet $\bar{\kappa}_{obs} \pm SF$.

[9] Da der for en normalfordeling gælder, at $\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0,6827$, hvor μ er normalfordelings gennemsnit og σ er normalfordelings standardafvigelse.

Formuleret simpelt, så er standardfejlen et estimat af hvor tæt gennemsnittet af observationerne $\bar{\kappa}_{obs}$ er på det sande gennemsnit $\bar{\kappa}$ af den bagvedliggende fordeling. Hvorimod standardafvigelsen er et mål for hvor meget de individuelle observationer varierer omkring deres gennemsnit $\bar{\kappa}_{obs}$.

Standardfejlen aftager når antallet af observationer n øges, da estimatet af det sande gennemsnit κ forbedres. Standardafvigelsen aftager *ikke* når antallet af observationer n øges.

2.11.2 Konfidensintervaller

Et konfidensinterval er en ofte benyttet måde at angive den statistiske usikkerhed forbundet med f.eks. et gennemsnit. Hvis det antages at den bagvedliggende sande fordeling er en normalfordeling benyttes [JFM11]:

$$\bar{\kappa}_{obs} - t_{(\alpha/2,n-1)} \cdot \frac{s_{\kappa_{obs}}}{\sqrt{n}} < \kappa < \bar{\kappa}_{obs} + t_{(\alpha/2,n-1)} \cdot \frac{s_{\kappa_{obs}}}{\sqrt{n}}$$

Et 95% konfidensinterval for κ kan derved beregnes til

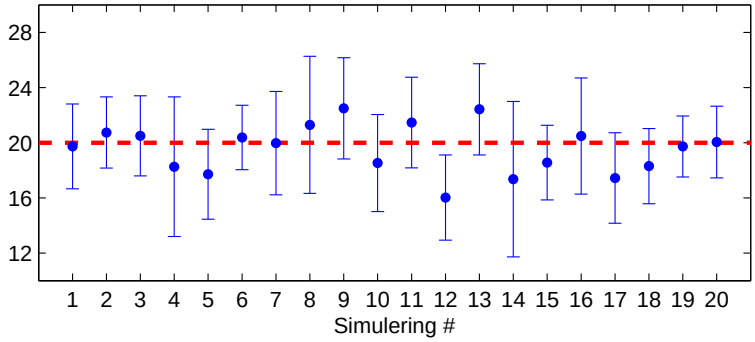
$$\begin{aligned} 15 - t_{(0{,}05/2,10-1)} \cdot \frac{1{,}15}{\sqrt{10}} < \kappa &< 15 + t_{(0{,}05/2,10-1)} \cdot \frac{1{,}15}{\sqrt{10}} \\ 14{,}18 < \kappa &< 15{,}82 \end{aligned}$$

Vi er altså 95% overbevist om at den sande værdi κ er i mellem 14,18 og 15,82.

Det er vigtigt at bemærke [Arv13]:

- Ethvert 95% konfidensinterval har **ikke** 95% sandsynlighed for at indeholde den faktiske sande værdi.
- I stedet siges at vi er 95% overbevist om at den faktiske sande værdi er indeholdt i konfidensintervallet.
- Da den faktiske sande værdi ikke er stokastisk, kan vi **ikke** sige at den faktiske sande værdi er i et bestemt interval med en given sandsynlighed.
- Konfidensintervaller er til gengæld stokastiske, da de afhænger af stokastisk data, dvs. at 95% af 95% konfidensintervaller vil indeholde den faktiske sande værdi.
- Konfidensintervaller er en praktisk måde at beskrive hvad udfaldet af et forsøg, *formentlig*, vil være hvis forsøget gentages.

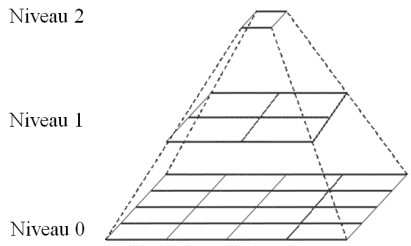
For at tydeliggøre ovenstående pointer er simuleret $n = 10$ udtræk fra en normalfordeling med $\mu = 20$ og $\sigma = 5$ og beregnet 95% konfidensintervallet for μ. Dette er gentaget 20 gange og de 20 konfidensintervaller ses på figur 2.12. De forskellige udtræk har forskellige gennemsnit og derfor er de 20 konfidensintervaller centeret om forskellige punkter. De forskellige udtræk har ligeledes forskellige standardafvigelser og derfor varierer bredden af de 20 konfidensintervaller.
I modsætning til en virkelig anvendelse, kender vi det sande faktiske gennemsnit ($\mu = 20$) af den bagvedliggende fordeling. Andelen af konfidensintervaller der indeholder den sande værdi $\mu = 20$ vil altid være $\approx (1 - \alpha)$, og i dette tilfælde ses at vi har netop den andel $\frac{19}{20} = 0,95$.



Figur 2.12: *Her ses 20 forskellige 95% konfidensintervaller for μ, baseret på 20 simuleringer af $n = 10$ udtræk fra en normalfordeling med $\mu = 20$ og $\sigma = 5$. Bemærk, at konfidensintervallet fra simulering nummer 12 ikke indeholder det faktiske sande gennemsnit $\mu = 20$. Andelen af konfidensintervallerne der indeholder den sande værdi $\mu = 20$, er $\frac{19}{20} = 0,95 = (1 - \alpha)$, hvilket illustrerer, at konfidensintervaller er stokastiske og ikke altid indeholder den faktiske sande værdi.*

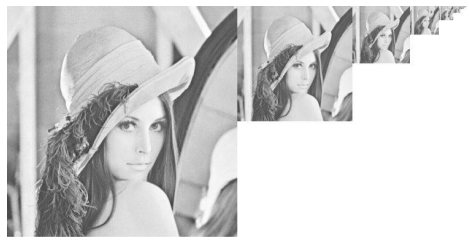
2.12 Gauss pyramider

Opløsningen af et digitalt billede er ofte valgt på baggrund af øvre begrænsninger i hardware og ikke efter hvad der er passende for den applikation, billedet skal bruges i. I de fleste applikationer skal desuden bruges information fra flere rumlige opløsninger af billedet [Car02]. Pyramider er en effektiv måde at præsentere billeder på, når der arbejdes med flere opløsninger.
Niveau 0 i en pyramide er en matrix med samme opløsning som det oprindelige billede. Ved at gå et niveau op i pyramiden reduceres antallet af søjler og rækker i matricen. Denne proces kan gentages indtil der nås et niveau, hvor billedet kun indeholder én pixel. Opbygningen af en pyramide er illustreret på figur 2.13.



Figur 2.13: *Pyramide. Niveau 0 har samme opløsning som det oprindelige billede. Opløsningen mindskes når niveauet øges.*

En type pyramide er den såkaldte Gauss pyramide. I en Gauss pyramide bliver hvert niveau lowpass filtreret (udglattet) og derefter subsampled. Subsampling foretages ved at udtrække hver anden række og søjle og slette resten. Resultatet er det næste niveau i Gauss pyramiden. Denne proces gentages til alle niveau er genereret. Figur 2.14 viser et eksempel på en Gauss pyramide. Både figur 2.13 og 2.14 er fra [Car02].



Figur 2.14: *Gauss pyramide.*

Statistik beregnet på de enkelte niveauer i en Gauss pyramide skal senere vise sig praktisk ifm. vurdering af effekten af forskellige koncentrationer af lipase ved vask af lipidplettede stoflapper. Se evt. afsnit 4.5.

Kapitel 3

Indledende arbejde

I starten af projektperioden fik jeg af Novozymes forklaret, hvordan Novozymes foretager forsøg med tøjvask. Efter denne gennemgang udarbejdede vi sammen en forsøgsplan. Forsøgsplanen beskrev i detaljer hvordan 40 stoflapper skulle vaskes i et dosis response forsøg. Multispektrale billeder af disse 40 stoflapper skulle så udgøre projektets datasæt.
Disse 40 stoflapper blev dog aldrig fremstillet. I stedet blev 24 stoflapper udleveret som stammede fra et større tidligere forsøg[1]. Disse 24 stoflapper var vasket under tilsvarende betingelser som i forsøgsplanen, hvorfor Novozymes vurdering var, at disse stoflapper var lige så anvendelige. At vaske nye stoflapper blot for at vaske nye stoflapper gav ikke mening, når der allerede eksisterede stoflapper vasket tilsvarende forsøgsplanen.
Tidligere havde jeg fået at vide, at Novozymes har en regel om, at målinger på stoflapper foretages senest 24 timer efter at de er vasket, dels fordi stoffet skal kunne nå at tørre og dels fordi målingen skal være mest mulig reproducerbar.
De 24 stoflapper fra det tidligere forsøg er vasket i august 2012. Dvs. der gik seks måneder fra stoflapperne blev vasket og til jeg fik dem udleveret og tog multispektrale billeder heraf. I den periode har stoflapperne været opbevaret i kølerum ved 5°C.
Jeg undrede mig over og spurgte derfor ind til hvorfor stoflapperne stadig kunne bruges efter seks måneder, når standardproceduren som nævnt er at måle efter senest 24 timer. Novozymes vurderede at stoflapperne stadig kunne bruges til dette projekt og ikke var påvirket af de seks måneders opbevaring. Hverken fedtpletterne eller stoflapperne oversteg Warwick Equests (producentens) fastsatte holdbarhedsdato og opbevaringen på køl stemte overens med Warwick Equests anbefalinger.
Konklusionen var derfor at de seks måneder gamle stoflapper godt kunne anvendes til dette projekt.

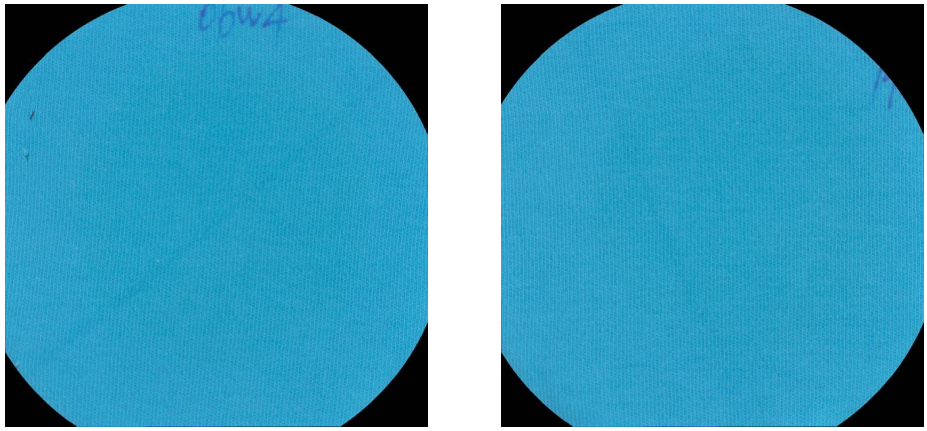
[1] Novozymes task no. H244-12.

Stoflapperne er med fedtpletter af typen olive oil spread[2] henholdsvis hamburger grease[3]. Stoflapperne er vasket i et dosis response forsøg med enzymkoncentrationerne 0,00% (også kaldet "blank"), 0,05%, 0,10%, 0,15% og 0,25% Lipex, som er et af Novozymes lipaseprodukter[4]. Procentangivelserne er i vægtprocent af detergentet. Der er foretaget en firedobbelt bestemmelse dvs. der er vasket fire af hver stoflap ved hver enzymkoncentration. Der er taget multispektrale billeder af to af de fire bestemmelser ved hver dosis samt af uvaskede stoflapper. Dette er gjort for både olive oil spread og hamburger grease stoflapperne. På figur 3.1 ses et eksempel på stoflapper med olive oil spread. Figur 3.2 viser en stoflap set ved forskellige bølgelængder.

[2]Identifikationsnummer hos Warwick Equest: 089BKC Olive Oil Spread.

[3]Identifikationsnummer hos Warwick Equest: 064BKC Hamburger Grease.

[4]Lipase er betegnelsen for vandopløselige enzymer, der nedbryder lipid. Lipid er en samlet betegnelse for fedtstof og fedtlignende kemiske forbindelser, der er hydrofobe, men opløselige i ikke-polære opløsningsmidler så som f.eks. benzen, æter og kloroform. Bl.a. triglycerider (populært kaldet fedtstoffer) og visse voksarter er lipider.



(a) (b)

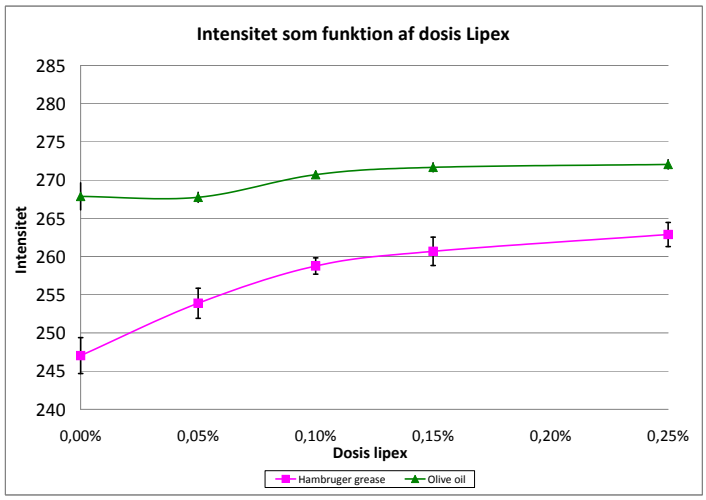
Figur 3.1: *Stoflapper påført fedtplet af typen olive oil spread. Stoflap (a) er vasket med 0,05% enzym og stoflap (b) er vasket med 0,25% enzym. Udover forskellen i enzymdosis er de to stoflapper behandlet ens. Det er meget svært visuelt at se forskel på renheden af de to stykker stof. Resultater fra kemisk FTIR analyse af tilsvarende stoflapper med tilsvarende fedtpletter viser dog, at det kan forventes at (b) kun indeholder cirka 30% af fedtmængden i (a). Se evt. figur 1.1 på side 2.*

(a) (b) (c)

Figur 3.2: *Stoflap med fedtplet af typen olive oil spread, set ved ultraviolet lys med bølgelængden 385nm (a), gult lys med bølgelængden 590nm (b) og nær infrarødt lys med bølgelængden 920nm (c). Det bemærkes at fedtplettens synlighed varierer med bølgelængden. Billederne er vist med grå farveafbildning (colormap).*

3.1 Novozymes egne målinger

Stoflapperne, hvis multispektrale billeder udgør det første af to stoflap-datasæt, er et uddrag af et større forsøg. Et døgn efter vask af stoflapperne er foretaget målinger af farveintensitet af fedtpletterne. Novozymes benytter bl.a. farveintensiteten som et udtryk for renheden af stof. Jo højere farveintensitet desto renere anses stoffet. På figur 3.3 ses Novozymes egne intensitetsmålinger af hamburger grease og olive oil spread pletterne. Ved hver dosis er vasket fire ens stoflapper. Middelværdien samt standardfejlen af disse fire ens stoflappers intensitet er vist i figur 3.3. Det bemærkes, at intensitetsmålingerne for olive oil spread pletterne ikke afspejler at renheden stiger med dosis, som det ellers må forventes ved et dosis respons forsøg. Det bemærkes også, at variationen i målingerne for hamburger grease pletterne reelt betyder at størstedelen af de enkelte dosis statistisk ikke kan adskilles unikt fra hinanden.
Dette bekræfter at der er behov for en ny målemetode.
Novozymes mener dog det er rimeligt at antage at der ér forskel i renheden af stoflapperne og at jo større dosis Lipex der er brugt jo mere af fedtpletten er vasket af[5]. Der arbejdes derfor videre under denne antagelse.

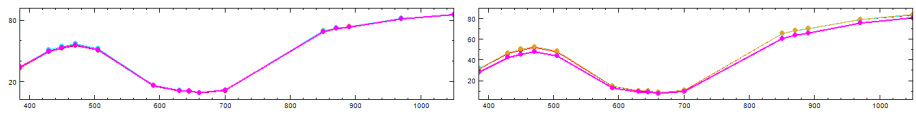


Figur 3.3: *Novozymes egne intensitetsmålinger af hamburger grease og olive oil pletterne. Der er angivet en standardfejl over og under hvert gennemsnit. Det bemærkes, at overlappende standardfejl betyder at pågældende stoflapper statistisk ikke kan adskilles fra hinanden.*

[5]Dog kun op til en naturlig grænse, hvor enzymets plateau niveau opnås. Når enzymkoncentrationen overstiger enzymets plateau niveau ses ikke længere en stigning i den enzymatiske effekt når der tilføjes mere enzym.

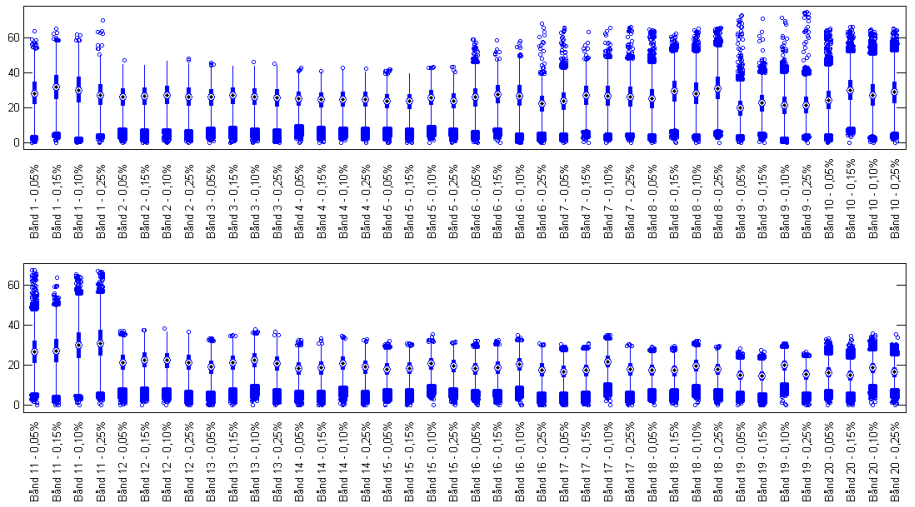
3.2 Signifikant forskellighed

På figur 3.4 ses et spektrum af middelværdierne af et repræsentativt udsnit af olive oil spread henholdsvis hamburger grease pletterne ved de fire enzymkoncentrationer. Spektrene overlapper meget, da der er meget lidt forskel i middelværdierne.

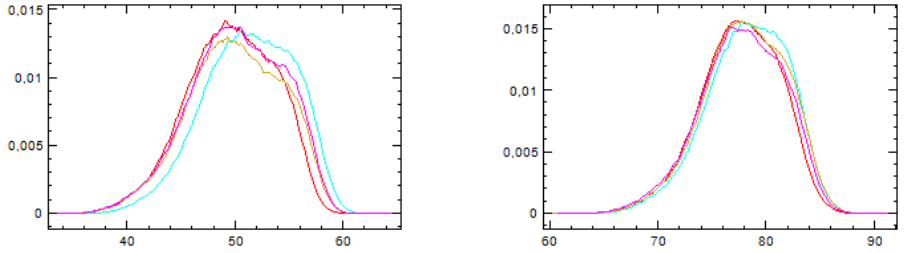


Figur 3.4: *Spektrum af middelværdierne af et repræsentativt udsnit af olive oil spread henholdsvis hamburger grease pletterne ved de fire enzymkoncentrationer. Rød 0,05%. Turkis 0,10%. Orange 0,15%. Lilla 0,25%.*

Motiveret af de meget ens spektrums er foretaget en række t-tests. Nulhypotesen er, at pixelværdierne i de respektive bånd, for de respektive pletter, kan antages at være uafhængige tilfældige stikprøver fra normalfordelinger, med samme middelværdi og samme men ukendte varians. Ved et $\alpha = 5\%$ signifikansniveau kan nulhypotesen forkastes for alle tests på nær seks. For olive oil kan nulhypotesen ikke afvises for bånd 4 i 0,15% og 0,25% samt bånd 5 i 0,05% og 0,25%. For hamburger grease kan nulhypotesen ikke afvises for bånd 11 i 0,05% og 0,15%, bånd 2 i 0,10% og 0,25%, bånd 4 i 0,10% og 0,25% samt bånd 6 i 0,15% og 0,25%. Der er altså signifikant forskel mellem størstedelen af de respektive bånd. På figur 3.5 på næste side og 3.7 på side 49 ses boxplots af olive oil spread og hamburger grease pletterne. På disse boxplots ses også en stor intern variation i pixelværdierne i de enkelte plettede stofområder.

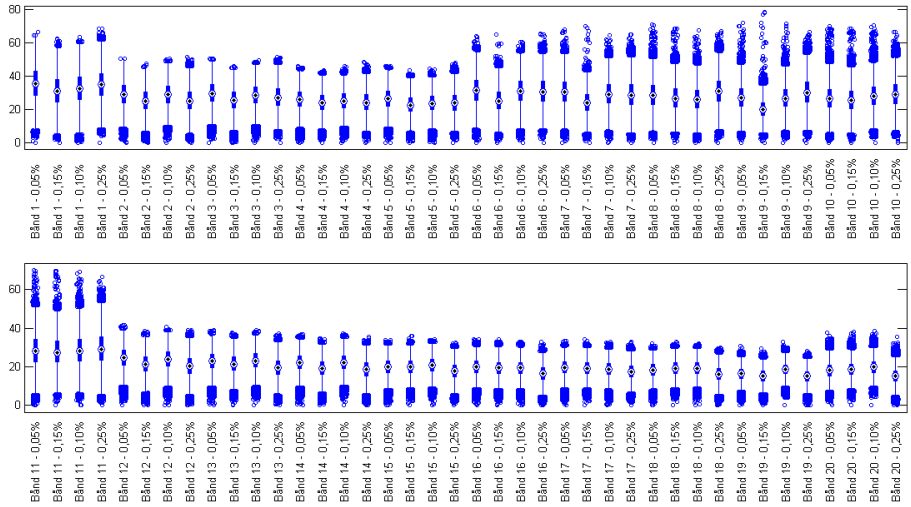


Figur 3.5: *Som illustration til de mange t-tests er genereret ovenstående boxplots af olive oil spread datasættet. Det ses at der er stor intern variation i pixelværdierne inden for de enkelte plettede stofområder.*

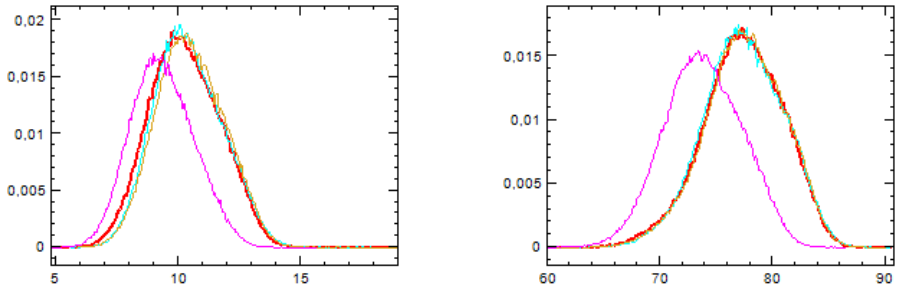


(a) **(b)**

Figur 3.6: *Histogrammerne for pixelværdierne i olive oil spread pletterne er meget ens og overlapper meget. Histogram (a) er af olive oil spread ved 430nm og (b) er af olive oil spread ved 920nm.*
Rød: 0,05%. Turkis: 0,10%. Orange: 0,15%. Lilla: 0,25%.



Figur 3.7: *Som illustration til de mange t-tests er genereret ovenstående boxplots af hamburger grease datasættet. Det ses at der er stor intern variation i pixelværdierne inden for de enkelte plettede stofområder.*



(a) **(b)**

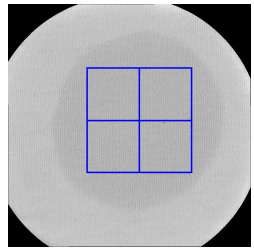
Figur 3.8: *Histogrammerne for pixelværdierne i hamburger grease pletterne er meget ens og overlapper meget. Histogram (a) er af hamburger grease ved 630nm og (b) er af hamburger grease ved 950nm. Dog skiller histogrammet for 0,25% sig markant ud fra de andre tre, ved både 630nm og 950nm.*
Rød: 0,05%. Turkis: 0,10%. Orange: 0,15%. Lilla: 0,25%.

3.3 Tekstur sammenligning

Det at påføre et stykke stof en fedtplet påvirker stoffets tekstur. Det var derfor naturligt at starte med at undersøge, om ændringer i forskellige statistikker, der beskriver tekstur, kunne sammenkædes med ændringer i dosis.
Af boxplotsne i figur 3.5 og 3.7 kan ses at der er stor intern variation i fedtpletterne. For at kunne undersøge denne variations indflydelse på resultaterne af tekstur statistikker er hver fedtplet inddelt i fire lige store repræsentative dele. På figur 3.9 ses et eksempel på denne inddeling. Hver af de fire dele er betragtet som en række pixelværdier uden spatial struktur og nedenstående statistikker beregnet.

Gennemsnit	$\mu = \frac{1}{N}\sum_{i=0}^{N-1} x_i$
Varians	$\sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(x_i - \mu)^2$
Varianskoefficient	$cv = \frac{\mu}{\sigma}$
Skewness	$\gamma_1 = \frac{1}{(N-1)\sigma^3}\sum_{i=0}^{N-1}(x_i - \mu)^3$
Kurtosis	$\gamma_2 = \frac{1}{(N-1)\sigma^4}\sum_{i=0}^{N-1}(x_i - \mu)^4 - 3$
Energi	$e = \sum_i p_i^2$
Entropi	$s = -\sum_i p_i log(p_i)$

For energien og entropien gælder at p_i er andelen af pixels med værdi i.



Figur 3.9: *Et repræsentativt udsnit af fedtpletten er valgt og inddelt i fire lige store dele.*

Hver statistik er beskrivende for billedets tekstur:

- Gennemsnittet kan ses som den gennemsnitlige intensitet i billedet.
- Variansen af et, som her, gråskala billede kan betragtes som et mål for billedets globale kontrast.
- Varianskoefficienten er et skalauafhængigt mål for den globale kontrast i billedet.
- Skewness beskriver til hvilken udstrækning outliers befinder sig på den ene side af gennemsnittet.
- Kurtosis beskriver forholdet mellem midten af fordelingen og halerne. En leptokurtic fordeling (positiv kurtosis) har meget vægt nær gennemsnittet og er slank i halerne. En platykurtic fordeling (negativ kurtosis) har meget vægt mellem gennemsnittet og halerne.
- Energien måler hvor ikke-uniform histogrammet af billedet er.
- Entropien måler uniformiteten af histogrammet af billedet.

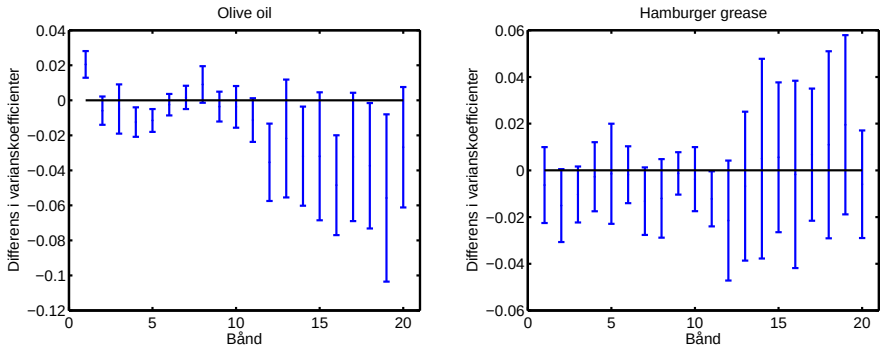
Først undersøges om der kan registreres forskel i dosisyderpunkterne 0,05% og 0,25%. Hvis tekstur statistikken ikke kan adskille dosisyderpunkterne, kan metoden forkastes, da der søges efter unik adskillelse af alle dosis. Kan statistikken adskille dosisyderpunkterne er det undersøgt, om de øvrige dosis, 0,10% og 0,15%, giver værdier, der stemmer overens med dosis.

3.3.1 Gennemsnit og varians

Af boxplotsne i figur 3.5 på side 48 og 3.7 på side 49 kan ses at gennemsnittet og variansen af de enkelte bånd overlapper for alle dosis for både olive oil spread og hamburger grease. Derfor kan gennemsnit og varians ikke benyttes til at adskille de forskellige dosis, hvorfor dette ikke er undersøgt nærmere.

3.3.2 Varianskoefficient

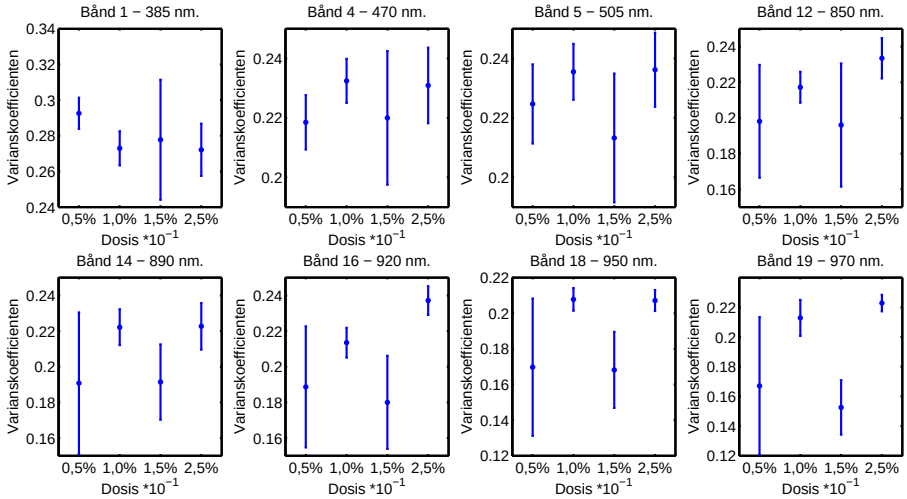
For hvert bånd er varianskoefficienten beregnet for hver af de fire inddelinger af olive oil spread og hamburger grease pletterne ved henholdsvis dosis 0,05% og 0,25%. Differensen mellem de to dosis er beregnet og plottet på figur 3.10. Standardafvigelsen af varianskoefficienterne for de fire dele af pletterne er betragtet som mål for metodens usikkerhed og plottet som usikkerhedsbarer på figuren. Hvis metodens usikkerhed overstiger differensen mellem de to dosis forkastes metoden for pågældende bånd. Metoden forkastes altså for et givent bånd, hvis usikkerhedsbaren for givende bånd overlapper nul.



Figur 3.10: *Differens i varianskoefficienter som funktion af bånd for olive oil spread henholdsvis hamburger grease.*
For olive oil apread forkastes metoden for alle bånd på nær bånd 1, 4, 5, 12, 14, 16, 18 og 19.
For hamburger grease er det kun bånd 11 hvor metoden ikke forkastes.

De bånd hvor usikkerheden på varianskoefficienten ikke overstiger differensen mellem dosis 0,05% og 0,25% undersøges nærmere. Hvis varianskoefficienten ikke kan adskille yderpunkterne i dosis er der ingen grund til at undersøge om varianskoefficienten kan adskille øvrige dosis.

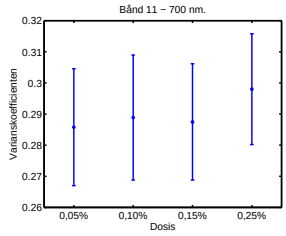
Nærmere undersøgelse af olive oil spread
På figur 3.11 ses varianskoefficienten for hver dosis i dosis response forsøget. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9. Som det fremgår af figur 3.11 kan de enkelte dosis ikke adskilles i rækkefølge af varianskoefficienten ved nogen af de pågældende bånd. For olive oil spread må metoden derfor forkastes.



Figur 3.11: *Varianskoefficienten beregnet for hver dosis for de otte bånd aflæst af figur 3.10. Det ses at varianskoefficienten ikke unikt kan adskille de enkelte dosis i rækkefølge for nogen af de pågældende bånd.*

Nærmere undersøgelse af hamburger grease

På figur 3.10 ses at for hamburger grease er det kun bånd 11 hvor usikkerheden ikke overstiger differensen mellem dosisyderpunkterne 0,05% og 0,25%. Bånd 11 er derfor undersøgt nærmere. På figur 3.12 ses varianskoefficienterne for bånd 11 for alle dosis i dosis response forsøget. Det ses at usikkerheden er for stor til at metoden i rækkefølge kan adskille de enkelte dosis. Metoden forkastes derfor for hamburger grease.



Figur 3.12: *Varianskoefficienten af bånd 11 beregnet for hver dosis. De enkelte dosis kan ikke unikt adskilles i rækkefølge.*

3.3.3 Øvrige statistikker

De øvrige statistikker, skewness, kurtosis, energi og entropi, er på tilsvarende måde som varianskoefficienten tjekket for korrelation med dosis. Beregningerne og argumentationen er tilsvarende ovenstående, og kan ses i bilag C på side 111.

3.3.4 Delkonklusion

Som tidligere skrevet ændres stofs struktur, når stoffet tilføres en fedtplet. Det var derfor naturligt at starte med at undersøge, om ændringer i forskellige statistikker, der beskriver tekstur, kunne sammenkædes med ændringer i dosis. Dette er for de enkelte bånd undersøgt for gennemsnittet, variansen, varianskoefficienten, skewness, kurtosis, energien og entropien. Det har dog ikke været muligt unikt at adskille de enkelte dosis ved ændringer i førnævnte statistikker.

Såfremt det havde været muligt unikt at adskille de enkelte dosis ved statistik beregnet på de enkelte bånd separat, ville det, afhængig af bølgelængden for og antallet af pågældende bånd, have været muligt at benytte en simplere kameraopsætning end VideometerLab. Såfremt teksturen af fedtpletten var markant nok til at kunne beregnes ud fra et gråskala billede af stoflappen, og den spektrale information i et RGB billede er tilstrækkelig til at adskille fedtpletten fra det rene stof, ville en opstilling bestående af et almindeligt kamera samt kamerablitz, kunne benyttes til at måle renheden af de vaskede stoflapper. Dette ville være en langt billigere løsning end at benytte af multispektrale billeder.
Da dosis ikke kan adskilles ved statistik beregnet på de enkelte bånd separat kan en sådan opstilling dog ikke benyttes.

3.4 Linearkombinationer og Mahalanobis afstand

I forrige afsnit om tekstur sammenligning blev det undersøgt om teksturstatistikker beregnet på de enkelte bånd kunne sammenkædes med dosis. Der blev regnet på de enkelte bånd separat og den multidimensionelle struktur i de multispektrale billeder blev således ikke udnyttet. For at udnytte den multidimensionelle struktur arbejdes i dette afsnit med at udnytte linearkombinationer af de enkelte bånd for at undersøge om dosis på denne måde kan adskilles unikt i rækkefølgen forventet fra et dosis respons forsøg.

Der gennemgås forsøg med normalized canonical discriminant analysis (nCDA), principal component analysis (PCA), minimum noise fractions (MNF) og Mahalanobis afstandsmål.
Teorien bag nCDA er gennemgået i afsnit 2.1 på side 8. Teorien bag PCA kan bl.a. findes i [TSK06] og [EC12] og teorien bag MNF kan f.eks. findes i [Nie99].

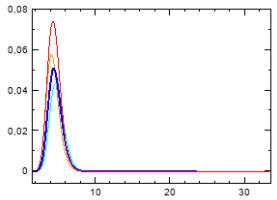
3.4.1 Mahalanobis afstand

Mahalanobis afstand er et afstandsmål, som tager hensyn til kovariansen mellem variable. For intuitivt at forstå dette, kan tænkes på, at for det euklidiske afstandsmål, er en punktmængde, hvor alle punkter er lige langt fra et andet givent punkt, en perfekt kugle. Målt i Mahalanobis afstand vil det ikke nødvendigvis være en perfekt kugle, da Mahalanobis afstanden strækker i kuglen for at korrigere for forskellige skalaer i de forskellige variable og for at redegøre for korrelationen mellem de enkelte variable. Derfor er Mahalanobis afstanden også afhængig af skala[6].

Mahalanobis afstanden fra en multivariat vektor $x = (x_1, x_2, x_3, ..., x_N)'$ fra en fordeling med gennemsnit $\mu = (\mu_1, \mu_2, \mu_3, ..., \mu_N)'$ og kovariansmatricen S er defineret således [EC12]:

$$D_M(x) = \sqrt{(x-\mu)'S^{-1}(x-\mu)}$$

I VideometerLab er Mahalanobis afstanden fra gennemsnittet af fedtpletten i olive oil spread dosis 0,05% og til alle pixelværdier af fedtpletterne i samtlige dosis beregnet. Figur 3.13 viser histogrammet af disse afstande.



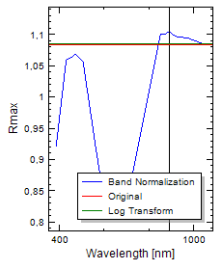
Figur 3.13: *Histogram over Mahalanobis afstanden fra gennemsnittet af pixelværdierne i olive oil spread dosis 0,05% og til alle pixelværdier af samtlige dosis. Det ses at Mahalanobis afstanden benyttet på denne måde ikke kan adskille de enkelte dosis, da histogrammerne overlapper kraftigt.*

[6] Yderligere information om Mahalanobis afstanden kan findes i f.eks. [TSK06] og [EC12].

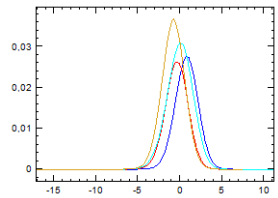
3.4.2 Anvendt nCDA

Der er beregnet en nCDA transformation (evt. se teorien på side 11) mellem det rene stof og olive oil spread fedtpletten for dosis 0,05%. Denne transformation er herefter anvendt på billeder af alle dosis. De malede områder kan ses på figur D.1 i bilag D på side 121. Den største Rayleigh koefficient er opnået med båndnormalisering som vist på figur 3.14. Resultatet er vist ved histogrammet på figur 3.15. Som det ses kan der ikke diskrimineres mellem de enkelte dosis. At Rayleigh koefficienten til trods herfor er høj skyldes, at Rayleigh koefficienten i dette tilfælde er et udtryk for hvor godt det rene stof og fedtpletten for dosis 0,05% kan adskilles ved nCDA.

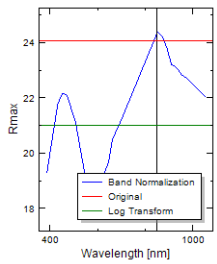
Hvis nCDA i stedet beregnes mellem dosisyderpunkterne for olive oil spread pletterne (dosis 0,05% og dosis 0,25%) og herefter anvendes på alle dosis opnås Rayleigh koefficienten og diskriminationen vist på figur 3.16 og 3.17. De malede områder kan ses på figur D.2 i bilag D. Som det ses kan der ikke diskrimineres mellem de enkelte dosis. Som det forventes fra teorien på side 11 har dosisyderpunkternes histogrammer gennemsnit på -1 og 1.



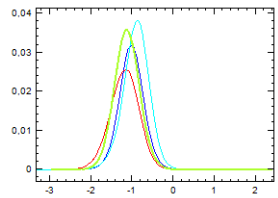
Figur 3.14: *Rayleigh koefficienten.*



Figur 3.15: *Histogrammer af scorebillederne. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, grøn: 0,25%.*



Figur 3.16: *Rayleigh koefficienten.*

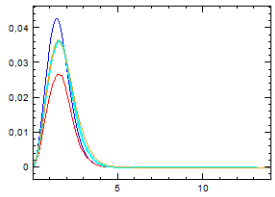


Figur 3.17: *Histogrammer over scorebillederne. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, brun: 0,25%.*

3.4.3 Anvendt PCA

Ved at beregningen og anvendelsen af PCA på hele billedet[7] af olive oil spread dosis 0,05%, observeres det at fedtpletten kun er synlig i principal komponent 1, 3 og 6. Dette er gældende for alle dosis. Transformationen til principal komponent 1, 3 og 6 er beregnet for olive oil spread dosis 0,05% og denne transformation er anvendt på alle dosis.
Herefter er gennemsnittet, medianen, standardafvigelsen, skewness, kurtosis, korrelationen og diagonal korrelationen[8] af de første 6-7 niveauer af en Gauss pyramide og en Laplace pyramide[Car02] beregnet på den første principal komponent. De 14 plots af statistikkerne regnet på disse pyramider ses i bilag D på figur D.3 og figur D.4. Hvis de enkelte dosis kan adskilles i rækkefølge ved ét eller flere niveauer i disse pyramider, ville pågældende niveauer kunne benyttes som model for dosis og dermed renheden af stoflapperne. Det ses dog at ingen af niveauerne kan adskille dosis i rækkefølge.

Mahalanobis afstanden fra gennemsnittet af principal komponent 1, 3 og 6 af olive oil spread dosis 0,05% og til alle pixelværdier af principal komponent 1, 3 og 6 af samtlige dosis er beregnet. Figur 3.18 viser histogrammet af disse afstande.



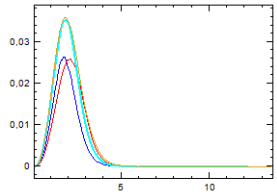
Figur 3.18: *Histogram over Mahalanobis afstanden fra gennemsnittet af principal komponent 1, 3 og 6 af olive oil spread dosis 0,05% og til alle pixelværdier af principal komponent 1, 3 og 6 af samtlige dosis. Det ses kan Mahalanobis afstanden benyttet på denne måde ikke kan adskille de enkelte dosis, da histogrammerne overlapper kraftigt.*

Hvis fedtpletterne i de enkelte dosis klippes ud af de respektive billeder og der beregnes en PCA kun på fedtpletten i dosis 0,05% og de første fire principal komponenter benyttes, repræsenteres 94,5% af variationen i fedtpletten for dosis 0,05%. For udklippet af fedtpletten gælder for alle dosis, at de første fire principal komponenter repræsenterer cirka 94% af variationen, og at ingen af de resterende principal komponenter repræsenterer mere end 1% af variationen.

[7] Fedtpletten er *ikke* klippet ud af billedet, så både det rene stof og fedtpletten er synlig.
[8] Korrelationen og diagonal korrelationen af en gray-level co-occurrence matrix (GLCM). Se f.eks. [Car02] side 227 formel 4.

Da fedtpletterne er klippet ud, kan kontrasten mellem fedtplet og det rene stof ikke benyttes til at vurdere et rimeligt antal principal komponenter. Derfor er antallet af principal komponenter valgt udfra hvor meget af variation i data de repræsenterer.
PCA transformationen er beregnet kun på fedtpletten i olive oil spread dosis 0,05% og de første fire principal komponenter er anvendt på de alle dosis. Der er beregnet en række statistikker på de enkelte niveauer i Gauss pyramider og Laplace pyramider af billederne af den første principal komponent. De 14 plots af statistikkerne regnet på disse pyramider ses i bilag D på figur D.5 og figur D.6. Ingen af niveauerne kan adskille dosis i rækkefølge.

Mahalanobis afstanden fra gennemsnittet af principal komponent 1-4 af olive oil spread dosis 0,05% og til alle pixelværdier af principal komponent 1-4 af samtlige dosis er beregnet. Figur 3.19 viser histogrammet af disse afstande.

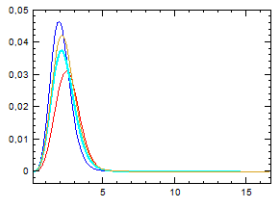


Figur 3.19: *Histogram over Mahalanobis afstanden fra gennemsnittet af principal komponent 1-4 af olive oil spread dosis 0,05% og til alle pixelværdier af principal komponent 1-4 af samtlige dosis. Det ses at Mahalanobis afstanden benyttet på denne måde ikke kan adskille de enkelte dosis, da histogrammerne overlapper kraftigt.*

3.4.4 Anvendt MNF

Ved at beregningen og anvendelsen af MNF på hele billedet[9] af olive oil spread dosis 0,05% observeres det, at fedtpletten kun er synlig i komponent 1-5. Dette er gældende for alle dosis. Transformationen til MNF komponent 1-5 er beregnet for olive oil spread dosis 0,05% og denne transformation er anvendt på alle dosis. På tilsvarende måde som i afsnit 3.4.3 er der beregnet en række statistikker på de enkelte niveauer i Gauss pyramider og Laplace pyramider af billederne af den første MNF komponent. De 14 plots af statistikkerne regnet på disse pyramider ses i bilag D på figur D.7 og figur D.8. Ingen af niveauerne kan adskille dosis i rækkefølge.

Mahalanobis afstanden fra gennemsnittet af pixelværdierne i MNF komponent 1-5 af olive oil spread dosis 0,05% og til alle pixelværdier af MNF komponent 1-5 af samtlige dosis er beregnet. Figur 3.20 viser histogrammet af disse afstande.



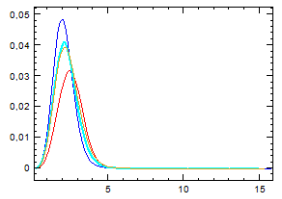
Figur 3.20: *Histogram over Mahalanobis afstanden fra gennemsnittet af pixelværdierne i MNF komponent 1-5 af olive oil spread dosis 0,05% og til alle pixelværdier af MNF komponent 1-5 af samtlige dosis. Det ses at Mahalanobis afstanden benyttet på denne måde ikke kan adskille de enkelte dosis, da histogrammerne overlapper kraftigt.*

Hvis fedtpletterne i de enkelte dosis klippes ud af de respektive billeder, og der beregnes en MNF transformation kun på fedtpletten i dosis 0,05% og de første fem komponenter benyttes repræsenteres 61,16% af variationen i fedtpletten for dosis 0,05%. For udklippet af fedtpletten gælder for alle dosis, at de første fem komponenter repræsenterer cirka 61% af variationen. De resterende komponenter ligner salt og peber støj. Da fedtpletterne er klippet ud, kan kontrasten mellem fedtplet og det rene stof ikke benyttes til at vurdere et rimeligt antal komponenter.
MNF transformationen er beregnet kun på fedtpletten fra olive oil spread dosis 0,05% og de første fem komponenter er anvendt på de alle dosis. På tilsvarende måde som i afsnit 3.4.3 er der beregnet en række statistikker på de enkelte niveauer i Gauss pyramider og Laplace pyramider af billederne af den første MNF komponent. De 14 plots af statistikkerne regnet på disse pyramider ses i bilag D

[9]Fedtpletten er *ikke* klippet ud af billedet, så både det rene stof og fedtpletten er synlig.

på figur D.9 og figur D.10. Ingen af niveauerne kan adskille dosis i rækkefølge.

Mahalanobis afstanden fra gennemsnittet af pixelværdierne i MNF komponent 1-5 af olive oil spread dosis 0,05% og til alle pixelværdier af MNF komponent 1-5 af samtlige dosis er beregnet. Figur 3.21 viser histogrammet af disse afstande.



Figur 3.21: *Histogram over Mahalanobis afstanden fra gennemsnittet af pixelværdierne i MNF komponent 1-5 af olive oil spread dosis 0,05% og til alle pixelværdier af MNF komponent 1-5 af samtlige dosis. Det ses at Mahalanobis afstanden benyttet på denne måde ikke kan adskille de enkelte dosis, da histogrammerne overlapper kraftigt.*

3.4.5 Delkonklusion

For olive oil spread fedtpletterne har hverken Mahalanobis afstanden eller teksturen i Gauss og Laplace pyramider af PCA og MNF transformationer af hele billedet og udklip af fedtpletterne, eller nCDA været i stand til unikt at adskille de enkelte dosis i rækkefølge.
Derfor er ovenstående teknikker ikke forsøgt på hamburger grease fedtpletterne.

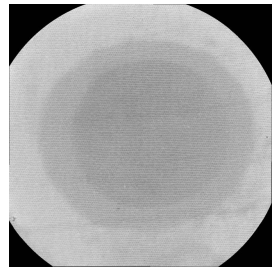
Alle hidtidige forsøg på at adskille dosis i rækkefølge efter den forventede renhed for et dosis respons forsøg er fejlet. I undren herover er udviklet en teori om fedtdiffusion. Denne teori gennemgås i næste afsnit.

3.5 Teori om fedtdiffusion

Som forklaret på side 4 foretager Novozymes målinger af renheden af vaskede stoflapper senest 24 timer efter endt vask. Som forklaret på side 43 i introduktionen til dette kapitel stammer de 24 stoflapper, der udgør det første af to stoflap-datasæt, fra et forsøg udført i august 2012. Der gik således seks måneder mellem at stoflapperne blev vasket og jeg fik dem udleveret og tog multispektrale billeder heraf. I disse seks måneder har stoflapperne været opbevaret i kølerum ved 5°C. Denne opbevaring på køl stemmer overens med Warwick Equests (producentens) anbefaling og stoflapperne oversteg ikke Warwick Equests fastsatte holdbarhedsdato.

Som mulig forklaring på, at alle ovenstående forsøg på at adskille de enkelte dosis i rækkefølge, er fejlet, udviklede jeg i samarbejde med Novozymes og mine vejledere en teori omkring fedtdiffusion i stoflapperne.

Teori: *Fedtpletten er en dynamisk størrelse, som diffunderer i porerne i stoffet. Diffusionen er dynamisk og detekterbarheden afhænger af tiden fra vask til måling. Diffusionen er både lateral og i dybderetningen. At diffusionen er lateral betyder, at fedtstoffet diffunderer væk fra centrum i samme plan som stoffet. Dette ses tydeligt ved nogle af hamburger grease pletterne. Se f.eks. figur 3.22. Når stoflapper med fedtpletter vaskes bliver overfladefedtet vasket af, mens noget af fedtpletten måske forbliver i dybden af tekstilet. Efter vask vil det resterende fedtstof diffunderer i dybderetningen, altså ud fra dybden af stoffet og tilbage til overfalden af stoffet.*
Det er netop denne dybderetningensdiffusion der forventes at være skyld i at alle hidtidige forsøg på separation af de forskellige dosis i korrekt rækkefølge har fejlet. Diffusionen har som en dynamisk proces med tiden øget fedtmængden nogle steder i stoflappen og mindsket fedtmængen andre steder.



Figur 3.22: *Stoflap med fedtplet af typen hamburger grease. Der ses tydeligt tegn på lateral diffusion af fedtstof. Den inderste mørke cirkel tolkes som fedtplettens oprindelige udbredelse, og cirkelen udenom tolkes som fedtstoffets nuværende udbredelse ved diffusion.*
stoflappen er vasket ved 0,25% Lipex og vist ved bånd 12, 850nm.

Hvis teorien er korrekt, kan det betyde, at de enkelte stoflapper ganske simpelt ikke kan rangordnes i dosis rækkefølge ud fra multispektrale billeder taget seks måneder efter vask. Dette fordi fedtfordelingen i stoflapperne har ændret sig i løbet af de seks måneder og stoflapperne derfor ikke længere har de renheder der forventes efter et dosis respons forsøg. En sådan diffusion vil forklare at alle hidtidige forsøg på rangordning i dosis rækkefølge er fejlet, da der i givet fald er ledt efter et system i data som ikke længere var til stede.
Min eksterne vejleder i Novozymes kender ikke til studier foretaget af Novozymes eller 3. part af fedts dynamiske egenskaber i stof, men finder ovenstående teori plausibel.
For at undersøge om ovenstående teori kan være korrekt, har jeg:

- Gentaget Novozymes oprindelige intensitetsmålinger af stoflapperne. Der er gået cirka otte måneder mellem de oprindelige målinger og gentagelsen.
- Vasket 40 nye stoflapper med lipidpletter i et dosis respons forsøg. I tidsrummet 23-24 timer efter endt vask er der taget multispektrale billeder af de nyvaskede stoflapper. Kapitel 4 beskriver arbejdet med disse nyvaskede stoflapper.

Gentagelsen af Novozymes oprindelige intensitetsmålinger er vist på figur 3.23 på side 66.

3.5.1 Gentagelse af intensitetsmålinger

For at undersøge om ovenstående teori om fedtdiffusion kan være korrekt er Novozymes oprindelige intensitetsmålinger gentaget. De oprindelige målinger blev foretaget i midt august 2012. Gentagelsen af målingerne er foretaget midt april 2013. I den mellemliggende periode på otte måneder har stoflapperne været opbevaret på køl ved 5°C. Dette stemmer overens med producentens anvisninger for langvarig opbevaring. Det organiske materiale i fedtpletterne burde således ikke have taget skade af opbevaringen.

På figur 3.23 på næste side ses til venstre de oprindelige intensitetsmålinger, som også vist på figur 3.3 på side 46, samt til højre gentagelsen af de samme intensitetsmålinger efter otte måneders opbevaring ved 5°C.

Hvis resultatet af de enkelte intensitetsmålinger betragtes som uafhængige identisk fordelte stokastiske variable kan intensitetsmålingerne illustreret i figur 3.23a og 3.23b meningsfyldt trækkes fra hinanden og usikkerheden i differensen betragtes.
For to uafhængige identisk fordelte stokastiske variable, X og Y, gælder at standardfejlen af differensen mellem gennemsnittet $\bar{x}$ og gennemsnittet $\bar{y}$ er [JFM11]:

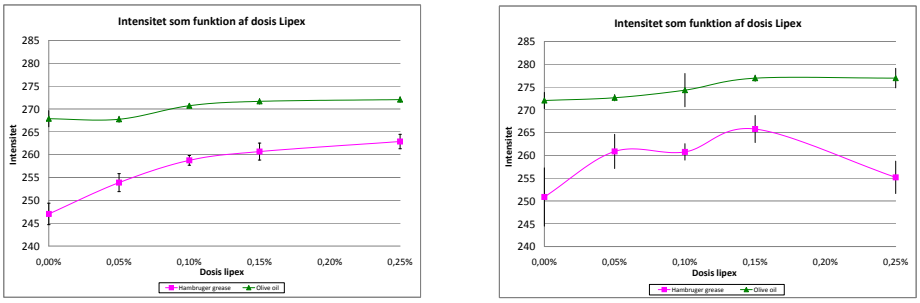
$$SE_{(\bar{x}-\bar{y})} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \tag{3.1}$$

hvor s_X^2 er variansen af n_X observationer af X med gennemsnittet $\bar{x}$. Tilsvarende er s_Y^2 variansen af n_Y observationer af Y med gennemsnittet $\bar{y}$.

Hvis X betragtes som de oprindelige intensitetsmålinger og Y betragtes som gentagelsen af de oprindelige intensitetsmålinger, svarer (3.1) til standardfejlen af differensen mellem de oprindelige intensitetsmålinger og gentagelsen af disse. Denne standardfejl er benyttet på figur 3.24 som mål for usikkerheden ved differensen.

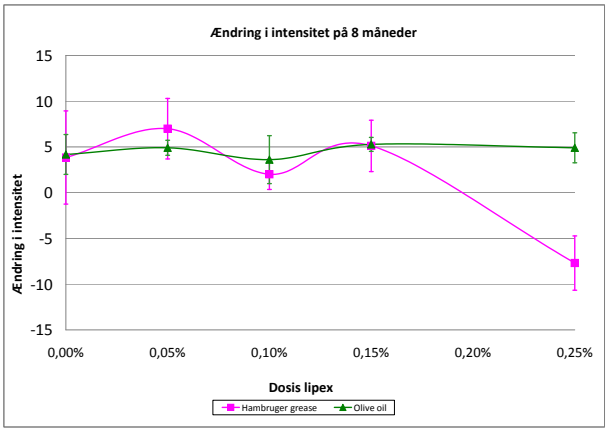
3.5.2 Delkonklusion

Figur 3.24 på side 67 viser differensen mellem de oprindelige målinger og gentagelsen. Der ses en signifikant ændring i farveintensiteten efter de otte måneders opbevaring. Teorien om fedtdiffusion kan derfor ikke afvises. Den påviste ændring forklarer, at alle hidtidige forsøg på rangordning i dosis rækkefølge er fejlet, da der er ledt efter et system i data, som ikke længere var til stede.



(a) **(b)**

Figur 3.23: *Til venstre (a) ses Novozymes oprindelige intensitetsmålinger. Til højre (b) ses en gentagelse af målingerne foretaget otte måneder senere.*
Der er angivet en standardfejl over og under hvert punkt. I det oprindelige forsøg er benyttet en firedobbelt bestemmelse og gentagelsen er baseret på en dobbeltbestemmelse.
Det bemærkes, at overlappende standardfejl betyder at pågældende stoflapper statistisk ikke kan adskilles fra hinanden. Det bemærkes desuden, at specielt stoflapperne med hamburger grease pletter har ændret renhed i løbet af de otte måneders opbevaring. Dette stemmer overens med den forventede laterale diffusion af fedtstoffet vist på figur 3.22.
Det bemærkes desuden, at standardfejlen for de enkelte dosis er steget markant over de otte måneders opbevaring, hvilket stemmer overens med teorien om fedtdiffusion.



Figur 3.24: *Differensen mellem Novozymes oprindelige intensitetsmålinger og gentagelsen af de samme intensitetsmålinger otte måneder senere. Det fremgår tydeligt, at renheden defineret ved farveintensiteten har ændret sig over de otte måneders opbevaring. Kun standardfejlen for hamburger grease dosis 0,00% overlapper nul, hvorfor ændringen er signifikant for de øvrige dosis og for alle dosis for olive oil spread. Farveintensiteten for olive oil spread har jævnt ændret sig 4-5 intensitetsenheder op, mens farveintensiteten for hamburger grease har ændret sig arbitrært skiftevis op og ned i farveintensitet. Disse ændringer skyldes formentlig dynamisk fedtdiffusion i stoffet som har afhængt af den tilbageværende fedtkoncentration i stoffet.*

Kapitel 4

Modellering af enzymatiske effekter

Efter gentagelsen af de otte måneder gamle intensitetsmålinger af stoflapperne med olive oil spread og hamburger grease fedtpletterne, kunne den i afsnit 3.5 beskrevne teori om fedtdiffusion ikke afvises. Derfor er vasket 40 nye stoflapper i et dosis respons forsøg[1]. I tidsrummet 23-24 timer efter endt vask er der taget multispektrale billeder af de nyvaskede stoflapper. Dette kapitel omhandler arbejdet med disse 40 nyvaskede stoflapper. På baggrund af multispektrale billeder af disse 40 nyvaskede stoflapper er udviklet en metode, der unikt kan adskille dosis benyttet i et dosis respons forsøg og derved renheden af stofstykkerne. I afsnit 4.8 på side 94 vises at denne metode ikke virker på de otte måneder gamle stoflapper, hvilket yderligere sandsynliggør teorien om fedtdiffusion.

De 40 nyvaskede stoflapper er blå stoflapper påført cirkulære lipidpletter som beskrevet i afsnit 1.1 på side 4. Lipidpletten (herfter kaldet fedtpletten) er af typen cooked beef fat[2]. Dosis respons forsøget er udført med enzymkoncentrationerne 0,00%, 0,10%, 0,20%, 0,30% og 0,40% Lipex, som er et af Novozymes lipaseprodukter[3]. Procentangivelserne er i vægtprocent af detergentet. Der er foretaget en ottedobbelt bestemmelse, hvilket vil sige at der er vasket otte ens stoflapper ved hver af de fem enzymkoncentrationer.

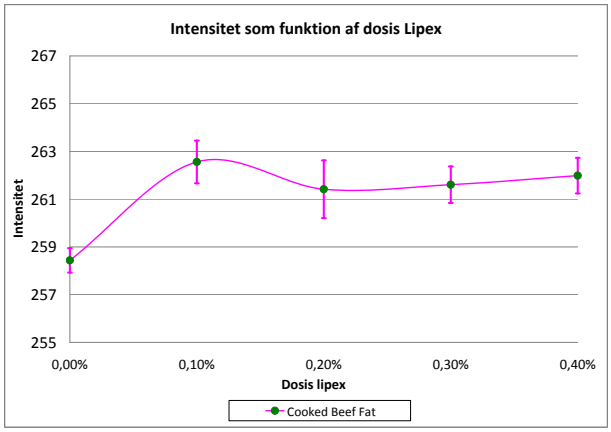
[1] Novozymes task no. H100-13.

[2] Identifikationsnummer hos Warwick Equest: 131BKC Cooked Beef Fat.

[3] Lipase er betegnelsen for vandopløselige enzymer, der nedbryder lipid. Lipid er en samlet betegnelse for fedtstof og fedtlignende kemiske forbindelser, der er hydrofobe, men opløselige i ikke-polære opløsningsmidler så som f.eks. benzen, æter og kloroform. Bl.a. triglycerider (populært kaldet fedtstoffer) og visse voksarter er lipider.

4.1 Intensitetsmålinger

Novozymes benytter farveintensiteten som et udtryk for renheden af blå stoflapper tilsvarende de 40 nyvaskede stoflapper med cooked beef fat. Jo højere farveintensitet desto renere anses stoffet. Jeg har foretaget intensitetsmålingerne cirka 21 timer efter endt vask. Intensitetsmålingerne er altså foretaget inden for Novozymes grænse på 24 timer efter endt vask. På figur 4.1 ses intensitetsmålingerne af de nyvaskede lapper. Gennemsnittet af intensitetsmålingerne for de otte bestemmelser ved hver dosis er vist ved en grøn prik, og standardfejlen ved de enkelte gennemsnit er illustreret ved lodrette usikkerhedsbarer.
For de vaske, der har indeholdt enzym, overlapper standardfejlene, hvorfor der statistisk ikke er forskel på middelværdierne. Hvis intensitetsmålingerne anses som et udtryk for renlighed, kan der altså ikke påvises en statistisk forskel i renligheden af stoflapperne vasket ved de forskellige enzymkoncentrationer.
Dette bekræfter at der er behov for en ny målemetode.
Novozymes mener dog, at det er rimeligt at antage, at der ér forskel i renheden af stoflapperne og at jo større dosis Lipex der er brugt jo mere af fedtpletten er vasket af[4]. Der arbejdes derfor videre under denne antagelse.



Figur 4.1: *Intensitetsmålinger af cooked beef fat pletterne. Der er angivet en standardfejl over og under gennemsnittet af de otte bestemmelser ved hver dosis. Det bemærkes, at for de vaske der har indeholdt enzym overlapper standardfejlene, hvorfor der statistisk ikke er forskel på middelværdierne og dermed renheden af stoflapperne.*

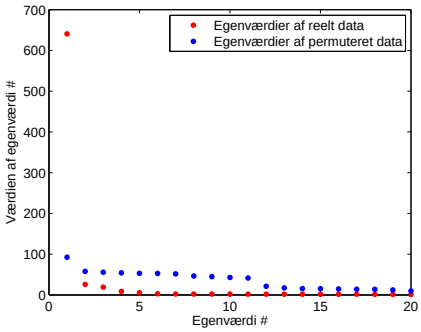
[4]Dog kun op til en naturlig grænse, hvor enzymets plateau niveau opnås. Når enzymkoncentrationen overstiger enzymets plateau niveau ses ikke længere en stigning i den enzymatiske effekt når der tilføjes mere enzym.

4.2 PCA: Variation versus relevant variation

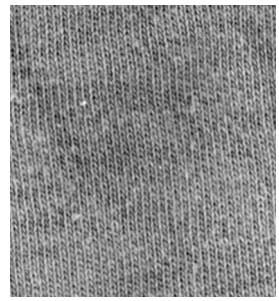
Principal component analysis[5] (PCA) er en ortogonal transformation, hvor den første principal komponent redegør for så meget af variationen i data som muligt, og de næste næste principal komponenter redegør for så meget af den tilbageværende variation i data som muligt under betingelsen at de næste principal komponenter skal være ortogonale på alle tidligere principal komponenter. På denne måde kan dimensionaliteten af data reduceres. I de fleste tilfælde kan PCA benyttes til at fjerne uønsket støj i data, ved at transformere data til et lavere dimensionelt rum udspændt af de første N principal komponenter. Antallet N af principal komponenter, der benyttes, kan udvælges f.eks. ud fra hvor mange procent af variationen i hele datasættet som de første N principal komponenter redegør for. Alternativt kan benyttes et plot tilsvarende figur 4.2 på den følgende side. Plottet er baseret på en stoflap vasket med dosis 0,10%. De røde prikker viser egenværdierne for principal komponenterne beregnet ud fra de oprindelige pixelværdier. De blå prikker viser egenværdierne for principal komponenterne beregnet hvor pixelværdierne i hvert bånd, forud for PCA, er vilkårligt permuteret. Hvis den N'te principal komponent af det oprindelige data ikke beskriver mere variation end den N'te principal komponent af det vilkårligt permuterede data, kan denne principal komponent antages blot at beskrive støj i data [Cle13]. Støj i denne sammenhæng er støj i PCA sammenhæng, dvs. data med lav variation.
Ved en stoflap tilført fedtplet, vil strukturen i det rene stof formentlig variere mere (grundet mønstret i vævningen) end en forholdsvis homogen tilført fedtplet, hvorfor PCA i denne situation vil behandle fedtpletten som støj og det rene stof som signal. På figur 4.3 på næste side ses den første principal komponent af en stoflap vasket med dosis 0,10%. Det bemærkes, at den første principal komponent netop redegør for variationen i stoffet og dermed ikke adskiller rent stof og fedtplet.

Det er altså vigtigt at være opmærksom på at variation i data ikke nødvendigvis er relevant variation. Af netop denne årsag benyttes PCA ikke yderligere i dette projekt.

[5] Teorien bag PCA kan bl.a. findes i [TSK06] og [EC12].



Figur 4.2: *Egenværdierne ved PCA af henholdsvis det oprindelige data og det vilkårligt permuterede data.*



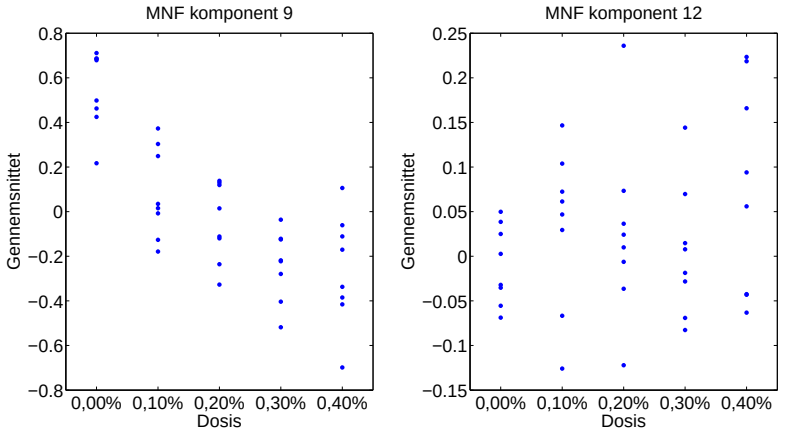
Figur 4.3: *Den første principale komponent af det oprindelige data. Det ses at denne principal komponent redegør for variationen i stoffet.*

4.3 Stabiliteten af MNF og CDA

Som en del af udviklingen af en metode til måling af den enzymatiske effekt er stabiliteten af minimum noise fraction (MNF) og CDA transformationer af de 40 fedtpletter undersøgt.

4.3.1 Stabiliteten af MNF

Ved at beregne gennemsnittet af hver MNF[Nie99] komponent af et af udklip af fedtpletterne er stabiliteten af MNF transformationen anvendt på stoflapperne undersøgt. Det beskrives i [Nie99] at den første MNF komponent har det højeste signal/støj forhold, og at signal/støj forholdet falder for de næste MNF komponenter. Derfor er det forventeligt, at de første MNF komponenter er mest stabile. I denne sammenhæng skal "mest stabile" tolkes som at gennemsnittet af MNF komponenterne varierer mindst for de første komponenter.
Stabiliteten er undersøgt ved at plotte gennemsnittet af de 20 MNF komponenter for alle 40 stoflapper. Ved at betragte disse 20 plots konstateres det dog, at stabiliteten af MNF komponenterne svinger arbitrært. For nogle MNF komponenter (f.eks. nr. 9) ses en tendens i gennemsnittet og for andre MNF komponenter (f.eks. nr. 12) ses blot tilsyneladende arbitrære værdier for gennemsnittet. Figur 4.4 viser plots af gennemsnittet af MNF komponent 9 og 12 for alle 40 stoflapper.
MNF transformationen er beregnet på et udklip af fedtpletten fra en stoflap vasket ved dosis 0,10%, og denne transformation er herefter anvendt på udklip af alle fedtpletterne.



Figur 4.4: *Plot af gennemsnittet af MNF komponent 9 henholdsvis 12 anvendt på udklip af fedtpletten for alle 40 stoflapper. For MNF komponent 9 ses en nedadgående tendens i gennemsnittet. For MNF komponent 12 ses blot tilsyneladende arbitrære værdier for gennemsnittet.*

4.3.2 Stabiliteten af CDA

Stabiliteten af forskellige CDA[6] transformationer af fedtpletterne er undersøgt ved at plotte gennemsnittet af scorebilledet for alle 40 fedtpletter. Dette er gjort for en række forskellige CDA transformationer, som er listet herunder.

1. CDA beregnet mellem en fedtplet vasket ved dosis 0,10% henholdsvis 0,20%.
2. CDA beregnet mellem en fedtplet vasket ved dosis 0,40% og det rene stof i samme stoflap.
3. CDA beregnet mellem dosisyderpunkterne 0,00% og 0,40%.
4. Den første CDF fra CDA beregnet mellem en fedtplet vasket ved dosis 0,00%, 0,10%, 0,20%, 0,30% og 0,40%.
5. Den anden CDF fra CDA beregnet mellem en fedtplet vasket ved dosis 0,00%, 0,10%, 0,20%, 0,30% og 0,40%.
6. Den tredje CDF fra CDA beregnet mellem en fedtplet vasket ved dosis 0,00%, 0,10%, 0,20%, 0,30% og 0,40%.
7. Den fjerde CDF fra CDA beregnet mellem en fedtplet vasket ved dosis 0,00%, 0,10%, 0,20%, 0,30% og 0,40%.

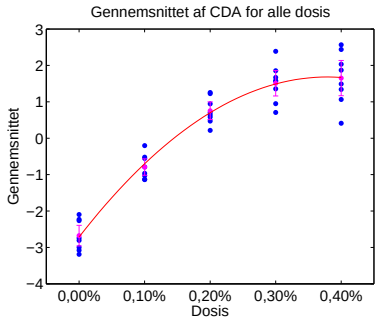
Renheden af de enkelte stoflapper forventes at stige asymptotisk imod enzymets plateau niveau, hvorefter tilføjelsen af mere enzym ikke længere vil øge den enzymatiske effekt. Denne stigning er illustreret på figur 4.5 på næste side ved et andengradspolynomium fittet til gennemsnittet for hver dosis. Med magenta er 95% konfidensintervallet for det sande gennemsnit illustreret som udtryk for måleusikkerheden. Hældningskoefficienten af andengradspolynomiumet (den røde linje) kan tolkes som "value for money", da hældningskoefficienten er svaret på spørgsmålet: Hvor meget renere bliver stoffet, hvis der tilføjes yderligere enzym?. Det bemærkes, at andengradspolynomiumets hældning mellem dosis 0,30% og 0,40% er cirka nul, hvilket indikerer at enzymets plateau niveau er nået.

Figur 4.5 og 4.6 illustrerer stabiliteten henholdsvis ustabiliteten af CDA transformation 1 og 5 fra ovenstående liste. De øvrige CDA transformationer er illustreret på tilsvarende måde i bilag E på side 127. Det bemærkes at de 40 gennemsnit af scorebillederne for CDA transformation 1 har en stigende tendens og

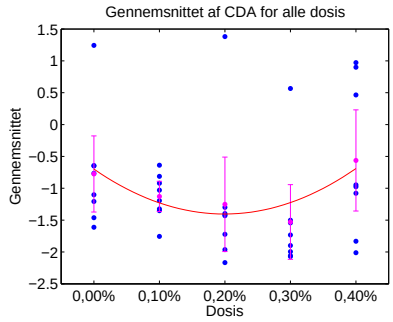
[6] Teorien bag CDA er gennemgået i afsnit 2.1 på side 8.

vokser asymptotisk som det biokemisk må forventes. CDA transformation 1 anses derfor for stabil. CDA transformation 5 anses derimod ikke som stabil, da de 40 gennemsnit af scorebillederne ikke (langt fra) viser det forventede mønster.

Af ovennævnte CDA transformationer anses 1, 3 og 4 for stabile.
I det videre arbejde med opstilling af en model for den enzymatiske effekt er benyttet CDA transformation 1.



Figur 4.5: *Stabiliteten af CDA transformation 1.*



Figur 4.6: *Stabiliteten af CDA transformation 5.*

4.3.3 Delkonklusion

Tilsyneladende svinger det arbitrært om gennemsnittet af MNF komponenterne afspejler det biokemisk forventede mønster i dosis (asymptotisk opadgående eller nedadgående) eller blot svinger arbitrært. Da CDA transformation 1, 3 og 4 er stabile og afspejler det biokemisk forventede mønster i dosis, konkluderes det, at det videre arbejde bedst udføres med en CDA transformation
Af disse CDA transformationer er nummer 1 valgt, da gennemsnitsbredden af 95% konfidensintervallerne er lavest for denne transformation, som derfor opfattes som den mest stabile.

4.4 Optimal maling ved CDA

For at beregne en CDA transformation skal først males, eller på anden måde udvælges, et repræsentativt udsnit af grupperne, der ønskes diskrimineret. Så længe de malede områder er repræsentative ændres CDFen ikke ved at male større områder. Det er derfor[7] interessant at undersøge, om et givent malet område er repræsentativt udvalgt. Hotelling's T^2 test bruges til at undersøge om prøver fra to normalfordelinger med samme varians-kovarians struktur kan antages at have samme middelværdi. Hvilket præcis er situationen ved maling ved CDA. De to prøver Hotelling's T^2 test skal benyttes på, vil så være det malede område og de omkringværende pixels. En antagelse ved Hotelling's T^2 test er, at de to prøver, der testes, er normalfordelte. Ved brug af Matlabfunktionen HZmvntest.m[8] er udført Henze-Zirkler's Multivariate Normality Test for at teste, om antagelsen om normalfordelt data er opfyldt. Herunder ses funktionens udskrift. Funktionen er kørt på 10.000 tilfældigt udvalgte pixels fra en fedtplet vasket ved dosis 0,10%.

```
Henze-Zirkler's Multivariate Normality Test
-------------------------------------------------------------------
Number of variables: 20
Sample size: 10000
-------------------------------------------------------------------
Henze-Zirkler lognormal mean: -0.0000970
Henze-Zirkler lognormal variance: 0.0001472
Henze-Zirkler statistic: 1.0014557
P-value associated to the Henze-Zirkler statistic: 0.0000000
With a given significance = 0.050
-------------------------------------------------------------------
Data analyzed do not have a normal distribution.
-------------------------------------------------------------------
```

Som det fremgår af funktionens udskrift, kan data ikke antages at være normalfordelt. Hotelling's T^2 test kan derfor ikke benyttes til at undersøge, om de malede områder ved CDA transformation er repræsentative.
Hvis Hotelling's T^2 test alligevel forsøges anvendt på udklip af fedtpletterne, vil testresultatet være, at selv højre og venstre billedhalvdel af f.eks. en fedtplet vasket ved dosis 0,10% ikke er repræsentativt udvalgt[9].
De malede områder males således alene ved øjemål sådan at øjet ser dem som repræsentative.

[7] Desuden afhænger beregningstiden for CDFen af størrelsen af de malede områder. Typisk går det dog så hurtigt, at beregningstiden er af mindre betydning.

[8] Downloaded fra MATLAB Central File Exchange, [TOHWBRCM].

[9] Hotelling's T^2 test er udført med Matlabfunktionen HotellingT2, downloaded fra MATLAB Central File Exchange, [TOHW13].

4.5 Modeller for den enzymatiske effekt

Der arbejdes videre med CDA transformation 1 fra afsnit 4.3.2 på side 74, da denne CDA transformation er den mest stabile. På figur 4.5 på side 75 ses gennemsnittet af scorebilledet for alle 40 fedtpletter efter anvendelsen af denne CDA transformation. På figuren er også vist 95% konfidensintervaller for de sande gennemsnit for de enkelte dosis. Hvis de enkelte stoflapper klassificeres som tilhørende den dosis, de er nærmest, klassificeres 29/40 stoflapper korrekt. Situationen er illustreret på figur 4.7 på side 80. Det bemærkes, at det primært er dosis 0,30% og 0,40% der klassificeres forkert, hvilket giver mening, da gennemsnittet for disse to dosis ligger meget tæt. Gennemsnitsbredden af de viste konfidensintervaller er 0,32.

For at mindske gennemsnitsbredden af konfidensintervallerne og øge antallet af stoflapper der klassificeres korrekt, er ved krydsvalidering benyttet sekventiel udvælgelse af attributter på baggrund af basisekspansioner, som herefter benyttes til at beregne en multiple lineær regression model (MLR model). Næsten al teorien gennemgået i kapitel 2 benyttes i dette afsnit.

Der er benyttet denne fremgangsmåde:

1. Fedtpletterne er klippet ud af billederne.[10]
2. CDA transformation 1 fra afsnit 4.3.2 er anvendt på alle 40 fedtpletter.
3. Gennemsnittet og standardafvigelsen for hver af de 40 scorebilleder er beregnet.
4. Standardafvigelsen, skewness, kurtosis og korrelationen[11] er beregnet for hvert niveau i en Gauss pyramide for niveauerne 0-5. Der er ikke benyttet normalisering.
5. Punkt 3 og 4 giver tilsammen en designmatrice X med 26 attributter for hver af de 40 fedtpletter. Der er beregnet basisekspansionerne $X = [X \quad X^2 \quad \sqrt{|X|} \quad log(|X|) \quad e^X]$ for at tage højde for, at nogle af attributterne måske har ulineær sammenhæng med renheden af stoffet. Ved at benytte førnævnte basisekspansioner forsøges at linearisere disse eventuelle ulineære tendenser, således at de kan beskrives ved en multiple lineær regression model. Designmatricen X indeholder nu 130 attributter for hver

[10] Dette er gjort manuelt, men kan f.eks. automatiseres ved brug af en CDA transformation tilsvarende den, der er vist nederst i afsnit 2.1.1 på figur 2.7 på side 13.

[11] Korrelationen af en gray-level co-occurrence matrix (GLCM). Se f.eks. [Car02] side 227 formel 4.

af de 40 fedtpletter. Den afhængige responsvariabel y er sat til de arbitrære værdier 0, 1, 2, 3 og 4 for de respektive dosis.

6. Der er foretaget to lag krydsvalidering som illustreret på figur 4.8 på side 80. To lags krydsvalidering gennemgås bl.a. i [TSK06].

   (a) Det ydre lag krydsvalidering er en 40-fold, svarende til leave-one-out, og sikrer, at den endeligt valgte model ikke er forudindtaget (eng: biased) af de pågældende 40 stoflapper. Da producenten af stoflapperne Warwick Equest ikke producerer stoflapperne med det formål at renheden skal aflæses maskinelt[Novozymes], kan der derfor forventes en høj batch til batch variation i både stoflapperne og fedtpletterne. Det ydre lag krydsvalidering har således til formål at sikre, at de endeligt udvalgte features er så robuste som muligt over for denne variation.

   (b) Det indre lag krydsvalidering er en 20-fold og benyttes til at foretage fremadgående[12] sekventiel udvælgelse af attributter. Det indre lag krydsvalidering sikrer, at det enkelte valg af attributter ikke er forudindtaget (eng: biased) for det testsæt, de valgte attributter er evalueret på.
   Summen af de kvadrerede fejl er benyttet som objektfunktion.

7. Der er nu 40 gange udvalgt en række attributter, der bedst beskriver data. På figur 4.9 på side 80 ses hvor mange gange hver attribut er valgt. Som det må forventes ses, at nogle attributter er vigtigere end andre. Alle attributter, der er udvalgt mere end 15 gange, benyttes i det videre arbejde. På denne måde sikres bedst imod batch til batch variation af stoflapperne og fedtpletterne. En oversigt over de udvalgte attributter kan ses i tabel 4.1 på modstående side.

8. De attributter der er udvalgt mere end 15 gange, kaldet γ_{15}, er ikke nødvendigvis udvalgt i samme fold af den ydre krydsvalidering og er således måske ikke alle sammen signifikante, når de benyttes samlet i én MLR model.
   For at teste signifikansen af γ_{15} attributterne beregnes en MLR model af disse, og ved t-test afgøres signifikansen. Der er benyttet signifikansniveauet $\alpha = 0,05$. Hvis en γ_{15} attribut ikke er signifikant udelades attributten af γ_{15}.

[12] Der benyttes *fremadgående* sekventiel udvælgelse af attributter, da bagudgående ikke er muligt, da designmatricen X i første iteration da vil indeholde alle 130 attributter for hver af de 40 fedtpletter. Herved overparameteriseres MLR modellen, da der forsøges løst 40 ligninger med 130 ubekendte. Uden regularisering kan dette ikke lade sig gøre.
Såfremt der ved fremadgående sekventiel udvælgelse udvælges mere end 40 attributter opstår samme problem. Dette er dog ikke sket og som det fremgår af det videre arbejde (se tabel 4.2 på side 84), bliver denne problemstilling formentlig heller ikke aktuel på fremtidige datasæt af stoflapper.

9. MLR modellen benyttet inkluderer et konstantled, hvis signifikans ligeledes tjekkes med t-test.

10. Der beregnes en MLR model af de signifikante γ_{15} attributter samt et konstantled hvis konstantleddet i forrige punkt viser sig signifikant.

11. MLR modellen testes ved leave-one-out krydsvalidering.

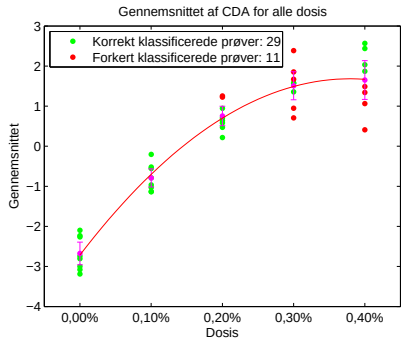
12. Hvis hver af de 40 fedtpletter klassificeres som hørende til den dosis, de er nærmest, klassificeres 35/40 fedtpletter korrekt. Se figur 4.10 på side 81. Der beregnes et 95% konfidensinterval for det faktiske sande gennemsnit af hver dosis. Det bemærkes, at ingen af disse konfidensintervaller overlapper. Dog er konfidensintervallet for dosis 0,30% og 0,40% meget lidt adskilt.

13. Det undersøges om residualerne opfører sig tilfældigt eller indeholder en eller flere tendenser. Dette gøres ved at teste om residualerne kan antages at have tilfældige fortegn og antages ikke at være korrelerede. Teorien bag disse to tests er gennemgået i afsnit 2.10.1 på side 36. Da $z = 0,34 < 1.96$ og $\frac{|\rho|}{T_\rho} = 0,74 < 1$ kan det ikke afvises, at residualerne har tilfældige fortegn og er ukorrelerede. Udover disse test betragtes et plot og histogram af residualerne. Se figur 4.11 på side 81.

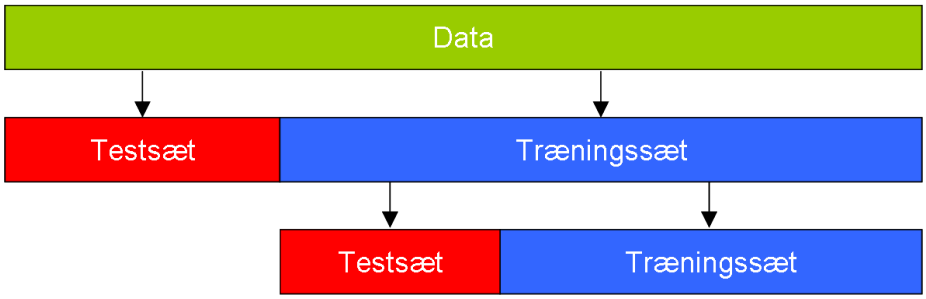
Ovenstående fremgangsmåde er benyttet i fem forskellige varianter. De fem varianter er gennemgået i afsnit 4.5.1 på side 82.

Tabel 4.1: *Oversigt over valgte attributter.*

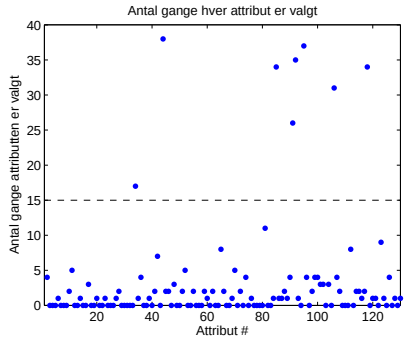
Attribut #	Antal gange valgt	Attribut navn
1	40	CDA Gennemsnit
34	17	(CDA Std.Dev. 5)^2
44	38	(CDA Kurtosis 3)^2
85	34	log(abs(CDA Std.Dev. 4))
91	26	log(abs(CDA Skewness 4))
92	35	log(abs(CDA Skewness 5))
95	37	log(abs(CDA Kurtosis 2))
106	31	exp(CDA Std.Dev.)



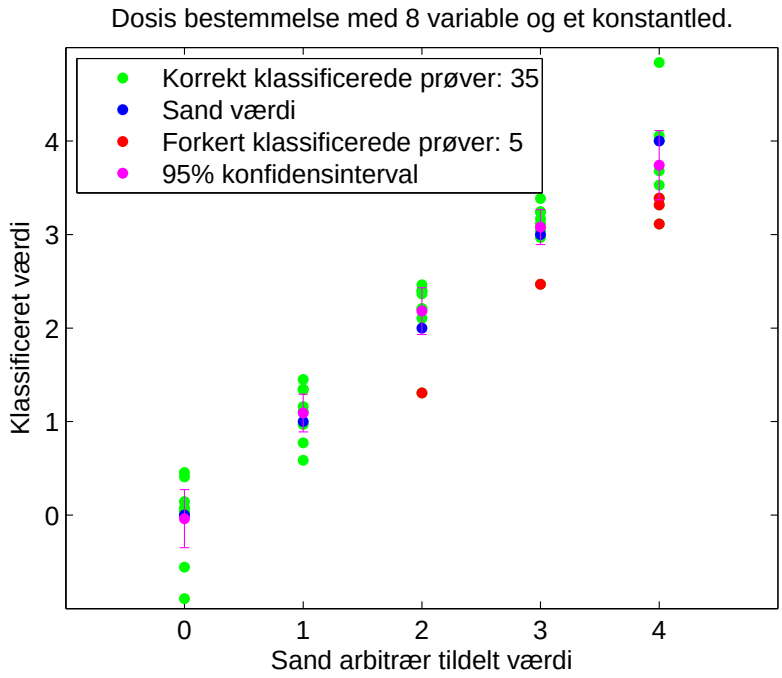
Figur 4.7: *De enkelte stoflapper er klassificeret som tilhørende den dosis de er nærmest.*



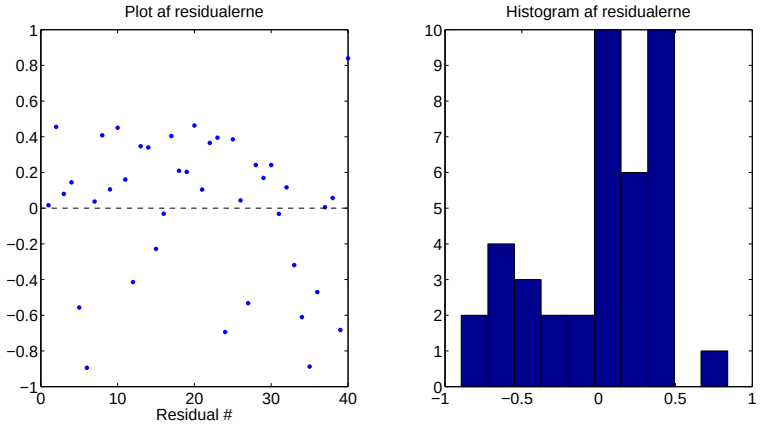
Figur 4.8: *Principskitse for to lags krydsvalidering.*



Figur 4.9: *Antal gange hvor enkelt attribut er udvalgt.*



Figur 4.10: *Hver fedtplet er klassificeret som tilhørende den nærmeste dosis. Der er vist 95% konfidensintervaller for det faktisk sande gennemsnit for hver dosis. Det bemærkes, at ingen af disse konfidensintervaller overlapper.*



Figur 4.11: *Residualerne for MLR modellen anvendt i figur 4.10.*

4.5.1 Forskellige varianter

Ovenstående metode er afprøvet i fem forskellige variationer som er beskrevet herunder. Tabel 4.2 på side 84 opsummerer resultaterne. Plots tilsvarende figurerne 4.10 og 4.11 på forrige side for variation 2 til 5 kan ses i bilag F på side 129.

Variation 1
Variation 1 er blot fremgangsmåden beskrevet på side 77.

Variation 2
Det fremgår af figur 4.7 på side 80, at hældningen af andengradspolynomiumet er cirka nul mellem dosis 0,30% og 0,40%, hvilket indikerer, at enzymets plateau niveau er nået. Hvis enzymets plateau niveau er nået, er der således ikke forskel i renheden af stoflapperne vasket ved dosis 0,30% og 0,40%. Dosis 0,30% og 0,40% er derfor behandlet som samme dosis. I ovenstående fremgangsmåde er derfor gjort følgende mellem punkt fire og fem.

- Hver anden stoflap fra dosis 0,30% og 0,40% er valgt og der ses bort fra de resterende, hvilket giver en repræsentativ delmængde.
- Årsagen hertil er, at MLR modellen ikke må være forudindtaget (eng: biased) imod den kombinerede 0,30%/0,40% dosis, hvorfor der skal være lige mange observationer til hver dosis. På denne måde undgås at der opstår et class imbalance problem[TSK06], hvor MLR modellen favoriserer den nye kombinerede dosis, fordi der er flere observationer ved denne kombinerede dosis.

Variation 3
Som det fremgår af figur 4.5 på side 75 stiger renheden[13] af en fedtplet ikke lineært med dosis. Dette stemmer også overens med biokemisk teori[Novozymes]. Derfor er forsøgt at benytte ovenstående fremgangsmetode, blot hvor den afhængige responsvariabel y i step 5 er sat til gennemsnittet af CDA scorebillederne. Dvs. $y = [-2,78 \ -0,80 \ 0,75 \ 1,51 \ 1,65]$ for de respektive dosis.

Variation 4
I denne variation er undersøgt betydningen af båndvis at benytte et 3×3 medianfilter på de 40 multispektrale billeder af stoflapperne inden fremgangsmåden beskrevet på side 77 benyttes.
I denne variation er den afhængige responsvariabel y ligeledes sat til gennemsnittet af CDA scorebillederne.

[13] Hvor renheden af en stoflap her defineres som gennemsnittet af scorebilledet af fedtpletten efter anvendelsen af CDA transformation 1 fra afsnit 4.3.2 på side 74.

At medianfiltrere før der foretages CDA mindsker within groups variationen og derfor stiger Rayleigh koefficienten, hvilket betyder, at CDFen separerer fedtpletten og det rene stof bedre. Resultatet af medianfiltreringen opsummeres i tabel 4.2 på næste side.
Der er ingen teori, der siger, at medianfiltreringen skulle gøre data normalfordelt. Ved matlabfunktionen HZmvntest.m[14] er udført Henze-Zirkler's multivariate normality test for at teste, om det kan antages, at pixelværdierne i et medianfiltreret billede er normalfordelte. Funktionen er kørt på 10.000 tilfældigt udvalgte pixels fra en fedtplet vasket ved dosis 0,10%. Funktionens udskrift ses i bilag F. For dette medianfiltrerede billede kan pixelværdierne ikke antages at være normalfordelte. Medianfiltrering kan således ikke benyttes til at muliggøre brugen af Hotelling's T^2 test til bestemmelse af optimalt malede områder ved dannelsen af en CDA transformationen.

Variation 5

Denne variation er en kombination af variation 2 og 3. Den afhængige responsvariabel y er sat til $y = [-2,78\ -0,80\ 0,75\ 1,58]$ for de respektive dosis. Hvor $1,58 = \frac{1}{2}(1,51+1,65)$ og den kombinerede 0,30%/0,40% dosis indeholder otte repræsentative punkter fra dosis 0,30% og 0,40%. De 8 repræsentative punkter er udvalgt således:

- Gennemsnittet af de 8 scorebilleder for dosis 0,30% og 0,40% er sorteret i numerisk rækkefølge.
- Hver anden stoflap er valgt og der ses bort fra de resterende, hvilket giver en repræsentativ delmængde.

I tabel 4.2 på den følgende side er opsummeret resultaterne for de fem variationer af metoden beskrevet i afsnit 4.5 på side 77. Der er benyttet leave-one-out krydsvalidering ved evaluering af modellerne.
Teststørrelsen for tilfældige fortegn er givet ved formel (2.15) fra side 37. Så længe $z < 1,96$ accepteres fortegnsrækkefølgen som tilfældig med et 5% signifikansniveau.
Som beskrevet i afsnit 2.10.1.3 på side 37 er der sandsynligvis trends til stede i residualerne, hvis absolutværdien af autokorrelationen overstiger grænsen for autokorrelation, dvs. hvis $|\varrho| > T_\varrho$.
Gennemsnittet af 95% konfidensintervallerne ønskes mindst muligt, da mindre konfidensintervaller betyder, at modellen er mere stabil.
Som det fremgår af figur 4.7 på side 80 er gennemsnittet af CDA scorebillederne tættest for dosis 0,30% og 0,40%. Derfor er afstanden mellem konfidensintervallerne for dosis 0,30% og 0,40% beregnet og vist i tabel 4.2 på den følgende side. I tabel 4.2 er den bedste værdi i hver række markeret med grøn og den værste

[14] Downloaded fra MATLAB central file exchange, [TOHWBRCM].

værdi med rød. Et "k" i rækken med antal attributter i modellen betyder, at modellen indeholder et konstantled.
Det bemærkes, at ingen af variationerne benytter mere end 40 attributter, hvorfor problemstillingen med en overparameteriseret model ved fremadgående sekventiel udvælgelse af attributter ikke opstår[15].

Tabel 4.2: *Opsummering af resultaterne for de fem variationer af metoden.*

Variation	1	2	3	4	5
Teststørrelsen for tilfældige fortegn	0,34	0,51	0,19	0,01	0,23
Teststørrelsen for autokorrelation	0,74	1,83	0,79	0,77	1,73
Gns. af 95% konfidensintervallerne	0,26	0,26	0,25	0,22	0,27
Afstanden mellem 0,30% og 0,40%	0,11	-	-0,44	-0,25	-
Antal korrekt klassificerede	35/40	29/32	29/40	30/40	30/32
Antal attributter i modellen	8 + k	3 + k	3 + k	4 + k	3

4.5.2 Delkonklusion

I det forrige afsnit er gennemgået fem variationer af en model for den enzymatiske effekt. Resultaterne er opsummeret i tabel 4.2. De fem variationer kan opdeles i to kategorier, alt efter om der er antaget, at enzymets plateau niveau er opnået, og dosis 0,30% og 0,40% derfor er opfattet som én kombineret dosis.

Variationer uden antagelse om ens dosis 0,30% og 0,40%

Variation 1
Af variation 1, 3 og 4 er variation 1 den eneste hvor 95% konfidensintervallerne for dosis 0,30% og 0,40% ikke overlapper. Det er også den variation, der klassificerer bedst.

Variation 3
Variation 3 klassificerer næst dårligst og klassificerer kun 29/40 stoflapper korrekt. Desuden overlapper 95% konfidensintervallerne kraftigt for dosis 0,30% og 0,40%. Denne variation benyttes derfor ikke yderligere.

Variation 4
Variation 4, som benytter medianfiltrering, er den variation, der klassificerer dårligst. Dette er til trods for at medianfiltreringen leder til

[15]Da det maksimalt udvalgte antal attributter er 8 plus et konstantled, er det usandsynligt, at fremtidige målinger på stoflapper vil føre til en situation med en overparameteriseret model.

en højere Rayleigh koefficient. Dette illustrerer, at Rayleigh koefficienten ikke er altafgørende, hvis MLR modellen benyttet på scorebillederne ikke er tilsvarende effektiv. Da variation 4 kun klassificerer 30/40 stoflapper korrekt, benyttes variationen ikke yderligere.

Variationer med antagelse om ens dosis 0,30% og 0,40%

Variation 2 & 5

Variation 2 & 5 klassificerer næsten lige godt, men variation 5 gør det med en simplere model (uden konstantled) og klassificerer én stoflap mere rigtigt ift. variation 2. Variation 2 er desuden den variation, der blandt alle variationer har de højeste værdier for teststørrelserne for tilfældige fortegn henholdsvis autokorrelerede residualer. Variation 2 benyttes derfor ikke yderligere.

Afhængig af om der antages eller ej, at enzymets plateau niveau er opnået, og dosis 0,30% og 0,40% derfor opfattets som en fælles dosis, er det enten variation 1 eller 5, der giver den bedste model for den enzymatiske effekt. Det må biokemisk afgøres, om en antagelse om fælles 0,30%/0,40% dosis er en rimelig antagelse, men set fra et statistisk synspunkt er variation 5 den bedste.
Dette skyldes, at variation 5 klassificerer $\frac{30}{32} \approx 94\%$ rigtige imod variation 1s $\frac{35}{40} \approx 88\%$, og dette endda ved brug af kun 3 attributter og intet konstantled ift. variation 1s 8 attributter og konstantled. MLR modellen i variation 5 er således både simplere og klassificerer bedre. Dette kunne tyde på, at MLR modellen i variation 1 er overtilpasset (eng: overfitted) i et forsøg på at modellere en ekstra dosis i data, som ikke er til stede.
En oversigt over de udvalgte attributter for variation 1 kan ses i tabel 4.1 på side 79. En oversigt over de udvalgte attributter for variation 5 kan ses i tabel 4.3[16].

Tabel 4.3: *Oversigt over valgte attributter ved variation 5.*

Attribut #	Antal gange valgt	Attribut navn
1	32	CDA Gennemsnit
18	17	CDA Kurtosis 3
105	32	exp(CDA Gennemsnit)

[16] "CDA Kurtosis 3"er kurtosis af 3. niveau i Gauss pyramiden.

4.6 Validering af modellen for enzymatisk effekt

For at validere de opnåede modeller for den enzymatiske effekt er genereret[17] et uafhængigt valideringssæt bestående af 40 nyvaskede stoflapper. Det uafhængige valideringssæt stammer fra en ny batch[18] af stoflapper fra Warwick Equest. Valideringen validerer således også de opnåede MLR modellers robusthed over for batch til batch variation. Stoflapperne i valideringssættet er vasket tilsvarende de 40 stoflapper i træningssættet. Der er således benyttet samme vandtemperatur, enzymkoncentration, enzymbatch, detergentkoncentration, detergentbatch, og hårdhedsgrad af vandet mm.

Som konkluderet i delkonklusion 4.5.2 herover, er det enten model variation 1 eller 5, der giver det bedste resultat når modellen ved krydsvalidering benyttes på træningsdataet. Herunder er model variation 1 og 5 testet på det uafhængige valideringssæt.

Testet på det uafhængige valideringssæt er foretaget således:

- MLR modellen er beregnet ved at følge punkt 1 til 10 af fremgangsmåden beskrevet i afsnit 4.5 på side 77. MLR modellen beregnes således på de 40 stoflapper i *træningssættet*.
- Attributterne i MLR modellen er beregnet for de 40 fedtpletter, der udgør det uafhængige *valideringssæt*.
- MLR modellen er evalueret på de beregnede attributter for de 40 fedtpletter, der udgør det uafhængige *valideringssæt*.
- Hver af de 40 fedtpletter klassificeres som hørende til den dosis de er nærmest.
- Der er beregnet et 95% konfidensinterval for det faktiske sande gennemsnit af hver dosis.

Figur 4.12 og 4.13 på side 88 viser valideringen af model variation 1.
Figur 4.14 og 4.15 på side 89 viser valideringen af model variation 5.

Resultaterne er opsummeret i tabel 4.4 på modstående side.
De to MLR modeller er opskrevet i formel (4.1) og (4.2).

[17] Novozymes task no. H247-13.
[18] De 40 stoflapper, der udgør testsættet, er fra batch nr. 85 og de 40 stoflapper, der udgør valideringssættet, er fra batch nr. 92.

Tabel 4.4: *Opsummering af resultaterne ved validering af variation 1 og 5 af metoden.*

Variation	1	5
Teststørrelsen for tilfældige fortegn	2,93	0,71
Teststørrelsen for autokorrelation	5,23	1,37
Gns. af 95% konfidensintervallerne	0,46	0,27
Afstanden mellem 0,30% og 0,40%	-0,71	-
Antal korrekt klassificerede	11/40	35/40
Antal attributter i modellen	8 + k	3

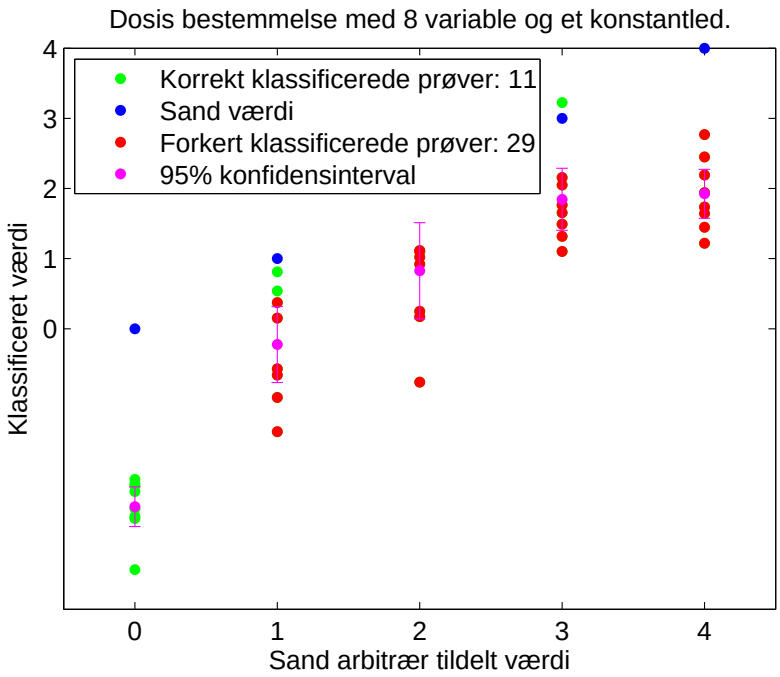
MLR modellen for variation 1 er givet ved:

$$\begin{aligned} f_1(x) = {} & 9,46 + 0,81 \cdot (\text{CDA gennemsnit}) - 1,66 \cdot (\text{CDA Std.Dev. 5})^2 \\ & - 1,14 \cdot (\text{CDA Kurtosis 3})^2 + 3,03 \cdot \log(\text{abs}(\text{CDA Std.Dev. 4})) \\ & - 0,29 \cdot \log(\text{abs}(\text{CDA Skewness 4})) + 0,28 \cdot \log(\text{abs}(\text{CDA Skewness 5})) \\ & + 0,27 \cdot \log(\text{abs}(\text{CDA Kurtosis 2})) - 1,98 \cdot \exp(\text{CDA StdDev}) \end{aligned} \tag{4.1}$$

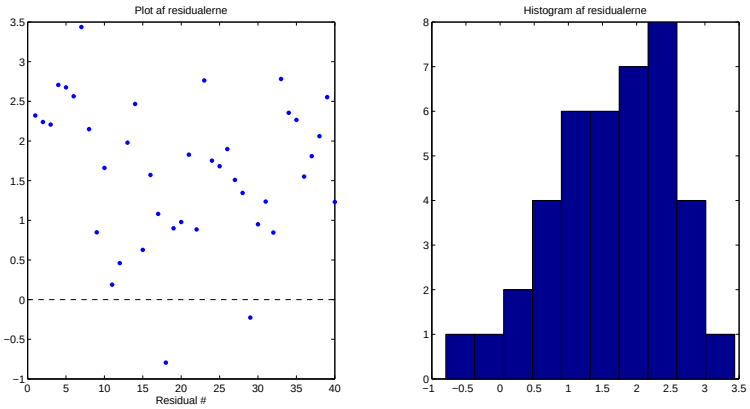
MLR modellen for variation 5 er givet ved:

$$\begin{aligned} f_5(x) = {} & 1,00 \cdot (\text{CDA gennemsnit}) - 0,47 \cdot (\text{CDA Kurtosis 3}) \\ & - 0,04 \cdot \exp(\text{CDA gennemsnit}) \end{aligned} \tag{4.2}$$

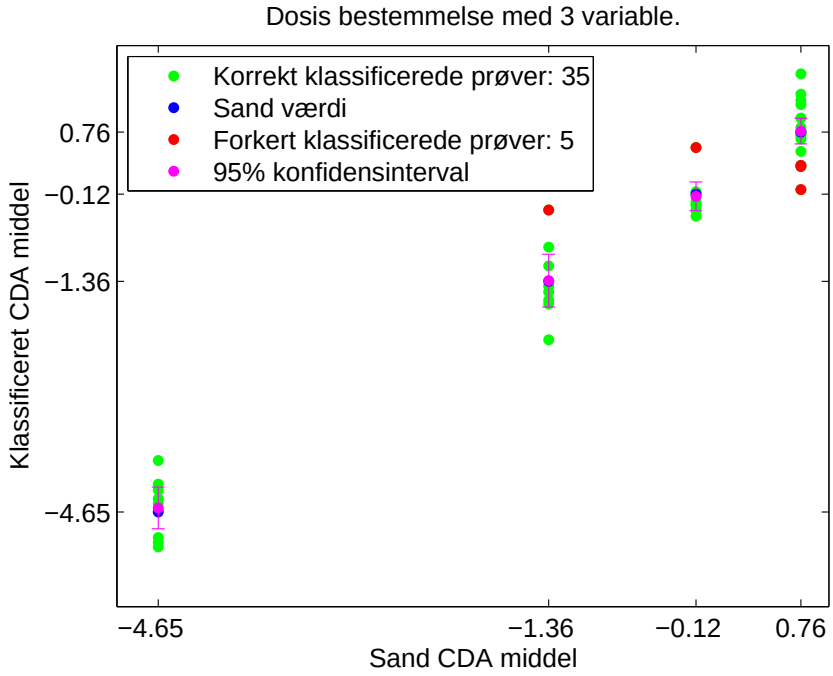
Begge modeller er afrundet til 2 decimaler.
Det bemærkes, at MLR model $f_5(x)$ er markant simplere end MLR model $f_1(x)$.



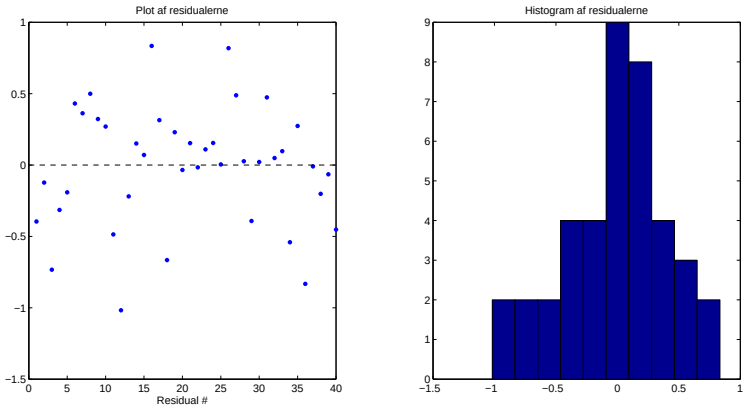
Figur 4.12: *Valideringsresultatet for model variation 1.*



Figur 4.13: *Residualerne for variation 1 af MLR modellen.*
Det bemærkes, at det visuelt ***ikke*** *er plausibelt, at residualerne er normalfordelte med gennemsnit i nul. MLR modellen giver systematisk for lave værdier og den marginale fordeling af residualerne har negativ skewness.*



Figur 4.14: *Valideringsresultatet for model variation 5.*



Figur 4.15: *Residualerne for variation 5 af MLR modellen.*
Det bemærkes, at det visuelt er plausibelt, at residualerne er normalfordelte med gennemsnit i nul.

4.6.1 Delkonklusion

Som det fremgår at figurerne på side 88 og 89 og tabel 4.4 på side 87 er det kun MLR modellen for variation 5, $f_5(x)$, der giver brugbare resultater på det uafhængige valideringssæt.
MLR model $f_5(x)$ er således vist robust overfor den høje batch til batch variation i stoflapperne og fedtpletterne.
MLR modellen for variation 1, $f_1(x)$, er derfor ikke brugbar i praksis.
Dette kan skyldes, at modellen $f_1(x)$ er for følsom over for batch til batch variation i stoflapperne. Der må forventes en høj batch til batch variation både i stoflapperne og fedtpletterne, da producenten af stoflapperne ikke producerer stoflapperne med det formål at renheden skal aflæses maskinelt[Novozymes].
Modellen burde ikke være overtilpasset (eng: overfitted) til træningsdataet, da der er benyttet krydsvalidering under træningen.

Såfremt enzymets plateau niveau rent faktisk ér nået, og der derfor ikke er forskel i renheden af dosis 0,30% og 0,40%, er dette en mulig forklaring på den dårlige præstation af MLR model $f_1(x)$ på valideringssættet. I givet fald er det blot tilfældig støj i træningsdataet, der udgør forskellen mellem dosis 0,30% og 0,40%. En modellering af dette tilfældige støj vil naturligt føre til en dårligere præstation på valideringssættet. Det høje antal attributter i MLR model $f_1(x)$ indikerer, at dette ér tilfældet. Fra et statistisk synspunkt vurderes det derfor, at enzymets plateau niveau rent faktisk ér nået, og der derfor ikke er forskel i renheden af dosis 0,30% og 0,40%.

En opsummering af præsentationen af MLR model $f_5(x)$, givet ved formel (4.2), på træningssættet/testsættet henholdsvis valideringssættet fremgår af tabel 4.5.

MLR model $f_5(x)$ er projektets endelige model for den enzymatiske effekt ved blå stoflapper med fedtpletter af typen cooked beef fat.

Tabel 4.5: *Sammenligning af MLR model $f_5(x)$ anvendt på træningssættet/testsættet og på valideringssættet.*

	Træningssæt/testsæt	Valideringssæt
Teststørrelsen for tilfældige fortegn	0,23	0,71
Teststørrelsen for autokorrelation	1,73	1,37
Gns. af 95% konfidensintervallerne	0,27	0,27
Antal korrekt klassificerede	30/32	35/40

4.7 Korrelation med vægtdata

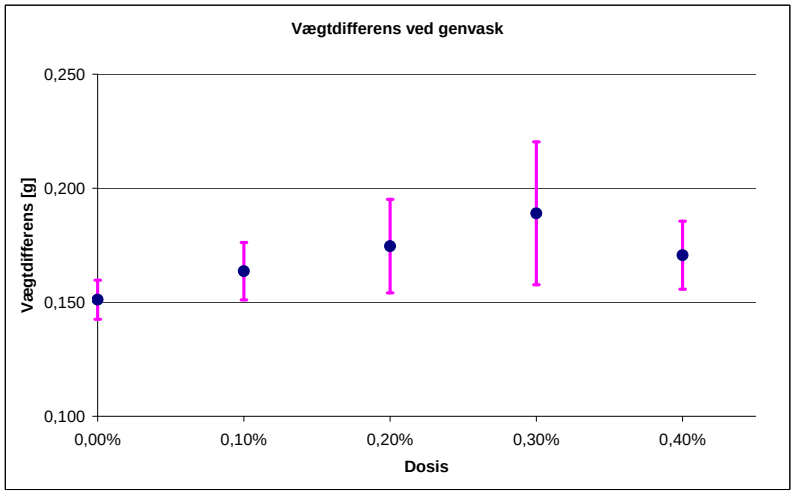
Det er undersøgt om resultaterne fra MLR modellen $f_5(x)$ (4.2) kan korreleres med vægtdata. Ideen er, at *hvis* det kan vejes, hvor meget fedt der er vasket ud af stoflapperne, så kan det undersøges, om disse vægtdata korrelerer med resultaterne fra MLR modellen $f_5(x)$.
Den optimale forsøgsudførelse ville have været at veje stoflapperne før og efter vask. Differensen i vægt må så være den mængde fedtstof, der er vasket ud af stoflapperne.
Imidlertid blev stoflapperne ikke vejet før vask og Novozymes foreslog[19] derfor at stoflapperne blev:

1. Tørret i varmeskab ved 65°C i en time[20].
   Dette for at opnå et kontrolleret (dog ukendt) niveau for fugtigheden i stoflapperne, uafhængig af luftfugtigheden den pågældende dag.
2. Vejet.
3. Vasket så rene, som det overhovedet kan lade sig gøre, ved at placere dem i en spand med 5L 30°C varmt vand, 50g detergent og 5ml Lipex.
   Ved disse koncentrationer sikres, at enzym- og detergentkoncentrationen ikke er begrænsende faktorer i renheden af stoflapperne.
   Stoflapperne er vasket i et døgn for at sikre, at vasketiden heller ikke er en begrænsende faktor.
4. Skyllet grundigt i rindende vand i 15 minutter for at sikre, at alle enzym- og detergentrester er skyllet ud af stoffet.
5. Lufttørret i et døgn.
6. Tørret i varmeskab ved 65°C i en time[21].
7. Vejet.

Stoflapperne er nu vasket så rene, som det overhovedet kan lade sig gøre[Novozymes]. Vægtdifferensen mellem før og efter denne genvaskning er et udtryk for, hvor meget fedtstof der var tilbage i stoflapperne efter første vask.

[19] Novozymes task no. H247-13.
[20] I praksis blev temperaturen de første 35 minutter holdt på 68°C, hvorefter varmelegemet blev slukket og temperaturen gradvist over de resterende 25 minutter faldt til 52°C.
[21] Starttemperaturen i varmeskabet er målt til 65°C og sluttemperaturen er målt til 61°C.

Gennemsnittet af vægtdifferenserne for de otte stoflapper fra hver dosis er beregnet. På figur 4.16 er plottet disse gennemsnit samt standardfejlen for hver gennemsnit. Rådata fremgår af tabel G.1 på side 136. Det bemærkes, at alle standardfejlene overlapper, hvorfor der statistisk ikke er forskel i gennemsnittene af vægtdifferenserne. Dvs. at ifølge vægtdifferenserne har der efter første vask ikke været forskel i renheden af stoflapperne vasket ved de forskellige dosis. Mængden af residualfedt i stoflapperne kan altså ikke afgøres ved vægtdataet[22]. Dette betyder, at vægtdataet ikke kan korreleres med MLR modellen, som derfor ikke kan valideres ud fra de givne vægtdata.



Figur 4.16: *Vægtdifferens ved genvask. Det bemærkes, at alle standardfejlene overlapper, hvorfor der statistisk ikke er forskel i gennemsnittene af vægtdifferenserne.*

I løbet af det døgn stoflapperne lå til tørre efter vask, blev der af Novozymes sået tvivl om, hvorvidt 65°C i en time var nok til at kontrollere fugtniveauet i stoflapperne. Efter at stoflapperne var vejet efter en time ved 65°C, blev de derfor placeret tilbage i varmeskabet og blev opbevaret der ved 65°C i yderligere halvanden time. Herefter blev stoflapperne vejet igen. Alle vejninger er foretaget med den samme vægt og måleresultaterne fremgår af tabel G.1 på side 136.
Differensen er beregnet mellem de 40 vægtmålinger foretaget efter en 1 time i varmeskab og de 40 vægtmålinger foretaget efter yderligere halvanden time i varmeskab. Gennemsnittet af disse 40 differenser er $0,015\text{g} \pm 0,080\text{g}$. Bemærk, at usikkerheden ($\pm 0,080\text{g}$) overstiger værdien af gennemsnittet ($0,015\text{g}$).

[22]Den benyttede vægt har en måleusikkerhed på ± 1mg.

Derfor kan det ikke afvises, at den sande værdi af differensen er nul, og at der derved ingen forskel er i vægten af stoflapperne før og efter de yderligere 1,5 time i varmeskab. Dette betyder, at det statistisk ikke kan afvises, at 1 time ved 65°C har været nok til at opnå et kontrolleret fugtniveau i stoflapperne.

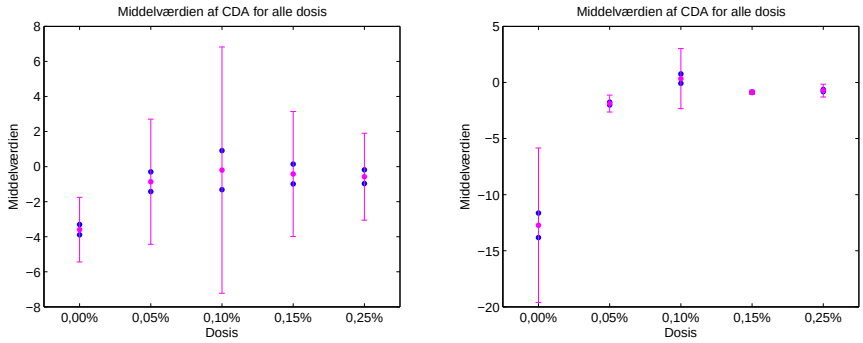
4.7.1 Delkonklusion

På basis af ovenstående konkluderes følgende:

- Statistisk kan det ikke afvises, at 1 time i varmeskab ved 65°C var nok til at opnå et kontrolleret fugtniveau i stoflapperne.
- Ud fra vægtdataet er det ikke muligt at konkludere på, om enzymets plateau niveau er nået, og derfor om stoflapperne vasket ved dosis 0,30% og 0,40% reelt er lige rene.
- Vejning med anvendelse af en vægt med en usikkerhed på ±1mg er ikke en mulig konkurrent til brugen af en MLR model.
- Med den anvendte vægt er det ikke muligt at korrelere vægtdata med MLR modellen.

4.8 Modeltest på 8 måneder gamle fedtpletter

For at teste om den udviklede metode til måling af den enzymatiske effekt også virker på otte måneder gamle stoflapper, er metoden forsøgt anvendt på de 10 vaskede olive oil spread fedtpletter og de vaskede 10 hamburger grease fedtpletter omtalt igennem kapitel 3.
Grundlaget for alle variationer af metoden er stabiliteten af CDA transformationen, som MLR modellen bygger på. I afsnit 4.3.3 på side 75 er konkluderet, at CDA transformationen mellem dosis 0,20% og 0,30% er den mest stabile til at adskille forskellige dosis. De tilsvarende dosis for olive oil spread og hamburger grease fedtpletterne er dosis 0,10% og 0,15%. Derfor er beregnet en CDA transformation herimellem og gennemsnittet af scorebillederne for hver fedtplet ses på figur 4.17. Det bemærkes, at begge CDA transformationer er ustabile, da gennemsnittene af scorebillederne ikke, som biokemisk forventet, udgør en asymptotisk monoton funktion. At CDA transformationerne er ustabile skyldes formentlig, at fedtpletterne er otte måneder gamle, og fedtpletterne derfor er diffunderet i stoffet, som det er beskrevet i afsnit 3.5.
At begge CDA transformationer er ustabile betyder, at MLR modellen ikke kan anvendes meningsfyldt på de otte måneder gamle fedtpletter.
At 95% konfidensintervallerne er så brede skyldes, at de er beregnet på baggrund af kun to målinger. Bredden kan gøres mindre ved at benytte et større antal målinger (stoflapper). Hvis de yderligere målinger ikke ændrer nævneværdigt ved de nuværende gennemsnit, vil et øget antal målinger dog ikke gøre CDA transformationerne stabile og derved muliggøre brugen af MLR model $f_5(x)$ på otte måneder gamle stoflapper. Da Novozymes altid måler deres fedtlapper efter senest 24 timer, er det dog af mindre betydning, om metoden virker på otte måneder gamle fedtpletter. Denne situation vil formentlig aldrig blive aktuel i praksis.



Figur 4.17: *Stabiliteten af CDA anvendt på de otte måneder gamle fedtpletter. Til venstre ses olive oil spread og til højre hamburger grease. Det bemærkes, at begge CDA transformationer er ustabile.*

4.9 Modeltest på glastallerkner

I forbindelse med forsøg med opvaskemidler benytter Novozymes glastallerkner som manuelt påføres besmudsning og herefter vaskes i en opvaskemaskine. Det er undersøgt, om en MLR model tilsvarende $f_5(x)$ kan benyttes til at prædiktere renheden af disse glastallerkner.
Novozymes vejer hver glastallerken tre gange efter følgende fremgangsmåde:

- Den rene tallerken vejes.
- Hver glastallerken påføres manuelt cirka 4,5g besmudsning og bages i varmeovn i to timer ved 180°C. Da besmudsningen tilføres manuelt, varierer den påførte mængde lidt.
- Glastallerknerne afkøles og vejes.
- Glastallerknerne vaskes under kendte vaskebetingelser[23].
- Glastallerknerne vejes.
- Der beregnes hvor mange % af den påførte besmudsning, der er fjernet.

Da det er en tidskrævende proces at veje hver glastallerken tre gange, er det undersøgt, om tilsvarende resultater kan opnås ved én måling med VideometerLab (et multispektralt billede), som herefter evalueres ved en MLR model.

I forsøget[24] hvor glastallerknerne, der er taget multispektrale billeder af, er genereret, er benyttet en femdobbelt bestemmelse. Dvs. der er vasket fem ens glastallerkner ved hvert sæt af forskellige vaskebetingelser. Det totale glastallerkendatamateriale består således af 25 multispektrale billeder. Besmudsningen fjernet ved hvert sæt af vaskebetingelser fremgår af tabel 4.6.

Tabel 4.6: *Besmudsning fjernet ved forskellige vaskebetingelser.*

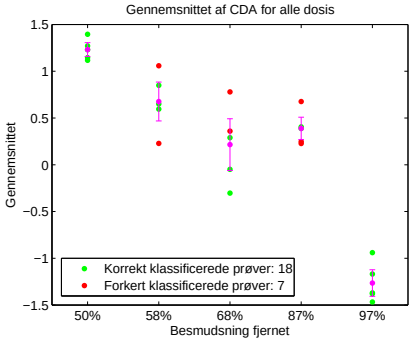
Sæt af vaskebetingelser	Besmudsning fjernet
# 1	48,78%
# 2	58,74%
# 3	68,49%
# 4	87,91%
# 5	97,86%

[23] F.eks. enzymkoncentration, vandets pH, vandets temperatur, detergenter mm.
[24] Novozymes task no. H183-13. Wash 5, 9, 13, 18 og 21.

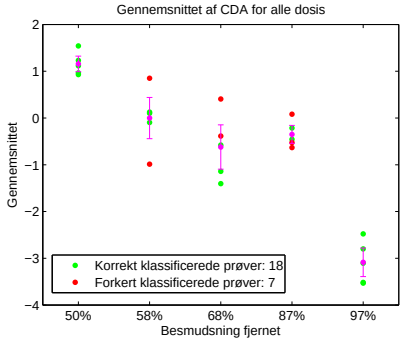
Stabiliteten af tre forskellige CDA transformationer er undersøgt efter tilsvarende princip som gennemgået i afsnit 4.3.2 på side 74 ved at plotte gennemsnittet af scorebilledet for alle 25 glastallerkner.
Nedenstående tre CDA transformationer er undersøgt:

1. CDA beregnet mellem en tallerken vasket med # 1 vaskebetingelser og en tallerken vasket med # 5 vaskebetingelser. Bemærk, at dette svarer til yderpunkterne i besmudsning fjernet.
2. CDA beregnet mellem en tallerken vasket med # 1 vaskebetingelser og en tallerken vasket med # 2 vaskebetingelser.
3. CDA beregnet mellem en tallerken vasket med # 1 vaskebetingelser og en tallerken vasket med # 4 vaskebetingelser.

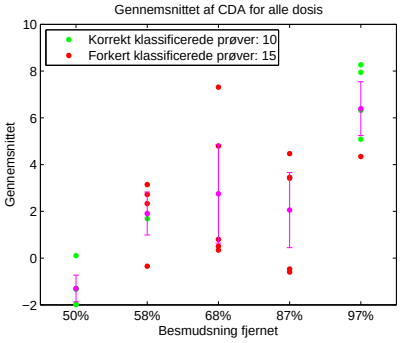
Figur 4.18, 4.19 og 4.20 herunder viser stabiliteten af de respektive CDA transformationer. Bemærk, at figur 4.18 og 4.19 **ikke** er ens, blot meget enslignende.



Figur 4.18: *Stabiliteten af CDA transformation 1.*



Figur 4.19: *Stabiliteten af CDA transformation 2.*



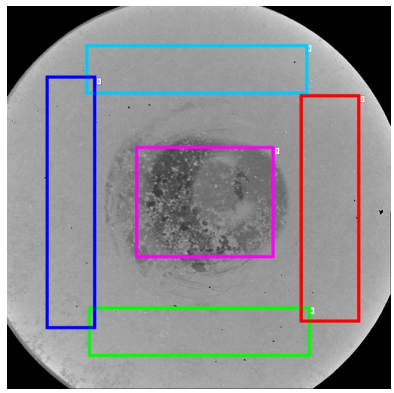
Figur 4.20: *Stabiliteten af CDA transformation 3.*

Af data illustreret på figur 4.18, 4.19 og 4.20 på forrige side konkluderes det, hvis der ses bort fra de fem glastallerkner vasket med vaskebetingelse # 4, så er CDA transformation 1 og 2 stabile og gennemsnittet af CDA scorebillederne afspejler data i tabel 4.6 på side 95. CDA transformation 3 ses ikke at være stabil.
Det bemærkes, at de fem glastallerkner vasket med vaskebetingelse # 4 skiller sig ud ved alle tre CDA transformationer. For CDA transformation 1 og 2 er 95% konfidensintervallet[25] desuden meget smalt, hvilket indikerer, at der begået en systematisk målefejl ved vejningen af disse fem glastallerkner.

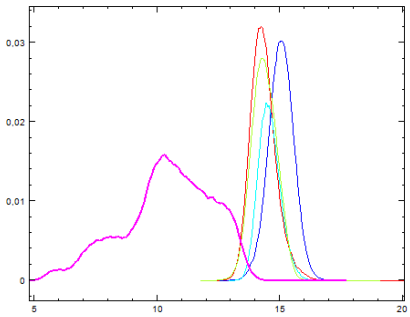
Baseret på CDA transformation 1 er der beregnet og testet to MLR modeller efter tilsvarende fremgangsmåde som beskrevet i afsnit 4.5 på side 77. Den ene MLR model $f_M(x)$ er trænet og testet på alle 25 glastallerkner. I trænings- og testsættet for den anden MLR model $f_U(x)$ er udeladt de fem glastallerkner vasket med vaskebetingelse # 4. På figur 4.25 på side 99 ses den første MLR models prædiktioner og på figur 4.26 på side 99 ses prædiktionerne for den anden MLR model. De to MLR modeller ses af formel (4.3) og (4.4) på side 99.

Bunden af nogle af glastallerknerne er en smule konkav og på nogle af de multispektrale billeder betyder dette, at et cirkulært område i midten af tallerknen ikke er et repræsentativt udsnit for hele tallerknen. Måske fordi lyset reflekteres mellem glastallerknen og bundpladen i VideometerLab og på denne måde optisk lokalt forstyrrer billedet. Båndvis er besmudsningen på glastallerknerne forholdsvis homogen.
MLR modellerne er derfor kun baseret på pixelværdierne i de yderste markerede regioner på figur 4.21 på næste side. Regionerne er valgt således, at de ikke overlapper hverken med området i centrum af glastallerknen eller yderkanten af billedet. Yderkanten af billedet er undgået, da yderkanten ikke er repræsentativ, da der er reflekteret ude fra kommende lys ind i ulbrichtkuglen[26] gennem selve glastallerknen.
Figur 4.24 på den følgende side viser et spektrum af gennemsnittet og standardafvigelsen af regionerne. Figur 4.22 og 4.23 på næste side viser histogrammer over regionerne ved henholdsvis bånd 3 (450nm, blåt lys) og bånd 8 (630nm, rødt lys).
Alt i alt bemærkes det af figur 4.21 til 4.24, at området i midten af glastallerknen skiller sig kraftigt ud fra resten af glastallerknen, hvilket er årsagen til at området i midten ikke er medtaget i beregningen af MLR modellen.

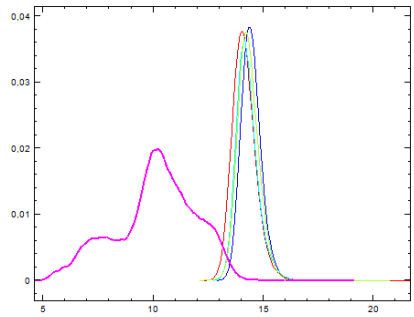
[25] For det sande gennemsnit af CDA scorebillederne for de fem glastallerkner vasket med vaskebetingelse # 4.
[26] Se afsnittet om VideometerLab på side 5.



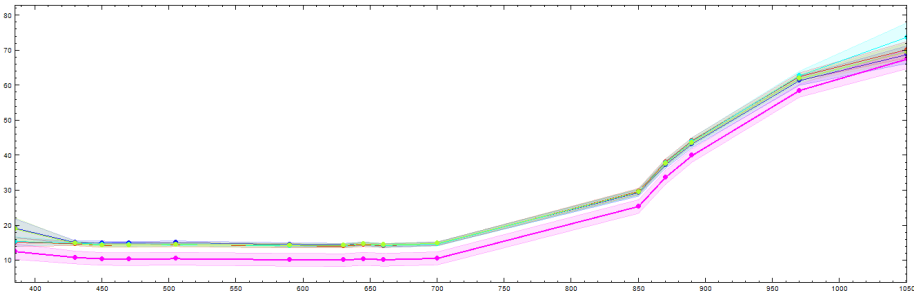
Figur 4.21: *Opdeling i regioner. De yderste fire regioner er medtaget ved beregningen af MLR modellerne. Regionen i midten er* ***ikke*** *medtaget i beregningen af MLR modellen, men er benyttet i figur 4.22, 4.23 og 4.24.*



Figur 4.22: *Histogrammer af regionerne figur 4.21 ved bånd 3 (450nm).*

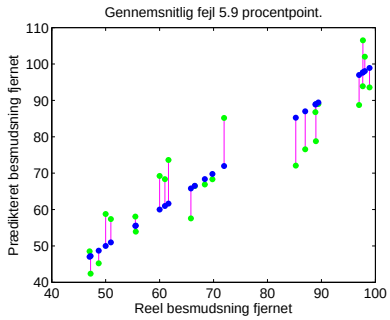


Figur 4.23: *Histogrammer af regionerne i figur 4.21 ved bånd 8 (630nm).*

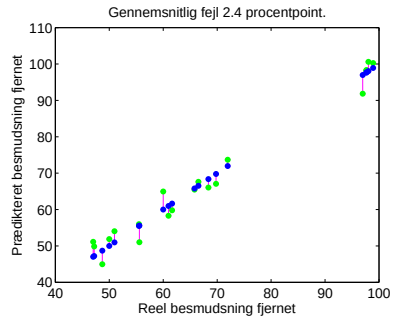


Figur 4.24: *Spektrum af gennemsnittet og standardafvigelsen af regionerne i figur 4.21. Bemærk, at den lilla region skiller sig kraftigt ud ved alle bølgelængder.*

På figur 4.25 og 4.26 ses prædiktionerne for MLR modellen trænet henholdsvis med og uden de fem glastallerkner vasket ved vaskebetingelser # 4. Det bemærkes, at den gennemsnitlige fejl (angivet i procentpoint) er $\approx 60\%$ mindre for MLR modellen, hvor vaskebetingelse # 4 er udeladt. Tabel 4.7 opsummerer de to modellers præstationer. Det bemærkes, at MLR modellen uden vaskebetingelse # 4 giver det bedste resultat i alle parametre. Dette indikerer yderligere, at der måske er foretaget en systematisk fejl ved vejningen af de fem glastallerkner.
Plots og histogrammer af residualerne for de to MLR modeller kan ses i bilag H på side 137.



Figur 4.25: *Prædiktioner for MLR modellen $f_M(x)$ baseret på alle 25 glastallerkner.*



Figur 4.26: *Prædiktioner for MLR modellen $f_U(x)$ hvor de fem glastallerkner vasket ved vaskebetingelser # 4 er udeladt.*

Tabel 4.7: *Opsummering af resultaterne for MLR modellen med og uden de 5 glastallerkner vasket ved vaskebetingelser # 4.*

	Med	Uden
Teststørrelsen for tilfældige fortegn	0,36	0,09
Teststørrelsen for autokorrelation	0,67	0,58
Gennemsnitlig fejl (procent point)	5,95	2,42
Attributter i modellen	3 + k	4+k

$$f_M(x) = 56,59 + 25,48 \cdot (\text{CDA StdDev})^2 + 14,31 \cdot \sqrt{|\text{CDA gennemsnittet}|} - 12,08 \cdot \exp(\text{CDA Mean}). \quad (4.3)$$

$$f_U(x) = 78,07 - 17,69 \cdot \text{CDA gennemsnittet} + 5,21 \cdot (\text{CDA Kurtosis 1})^2 + 19,85 \cdot \exp(\text{CDA Std.Dev. 4}) - 23,45 \cdot \exp(\text{CDA Std.Dev. 5}). \quad (4.4)$$

4.9.1 Delkonklusion

Det er undersøgt, om der kan benyttes en MLR model baseret på en CDA transformation til at prædiktere renheden af glastallerkner. Dette kan ikke endeligt konkluderes ud fra resultaterne.
Det er indikeret, at en MLR model baseret på en CDA transformation kan benyttes til at prædiktere renheden af glastallerkner, men yderligere undersøgelser er nødvendige før spørgsmålet endeligt kan besvares.
Det vil formentlig være muligt at konkludere endeligt på en MLR models brugbarhed, hvis der genereres et større datasæt. Datasættet skal bestå af multispektrale billeder af f.eks.[27] 40 glastallerkner, som er vasket og vejet efter normal procedure. Ved efterfølgende at undersøge om vægtdata korrelerer med MLR modellens prædiktioner, kan der formentlig konkluderes på brugbarheden af en MLR model ifm. glastallerkner.

[27] Desto flere glastallerkner der benyttes, desto større sikkerhed opnås ved valideringen.

Kapitel 5

Konklusion

Der er udviklet en metode til måling og dokumentation af enzymatiske effekter med spektral billedanalyse. Metoden er primært baseret på normalized canonical discriminant analysis (nCDA), sekventiel fremadgående udvælgelse af attributter (SUA) samt multiple lineær regression (MLR). Den udviklede metode giver relative værdier for renheden af de målte emner og det er foreslået, hvordan metoden måske kan udvides til at give absolutte værdier.
Metoden er udviklet på blå stoflapper med lipidpletter vasket i et dosis respons forsøg. Metoden er valideret på et tilsvarende vasket uafhængigt valideringssæt fra en ny batch og derved vist robust overfor batch til batch variation. Det er antaget, at enzymets plateau niveau blev opnået ved begge vaske og at stoflapperne vasket ved de to højeste enzymkoncentrationer derfor er vasket lige rene. Ved vejning har det ikke været muligt at afvise gyldigheden af denne antagelse. Så længe det ikke kan måles, om enzymets plateau niveau er nået, kan det ikke endegyldigt konkluderes, om metoden virker.
Blå stoflapper med lipidpletter er noget af det sværeste at måle ved remission og farveintensitet. Hvide stoflapper kan derimod nemt måles med remission. Da den udviklede metode er valideret på blå stoflapper med lipidpletter forventes det derfor, at den udviklede metode også fungerer for andre plettyper og stoftyper. Dette kan verificeres ved at korrelere metoden med f.eks. remissionsmålinger af hvide stoflapper.
Der er endvidere opnået lovende resultater ved at anvende modellen på glastallerkner. Et større datasæt er dog nødvendig før modellens brugbarhed på glastallerkner endeligt kan konkluderes.
Hvis metoden ikke tidligere har været anvendt på pågældende kombination af plettype og stoffarve skal metoden initialiseres, hvilket forventes at tage cirka et minuts brugerinteraktion og under et minuts beregningstid.
En renhedsmåling med modellen tager 5-10 sekunder og kræver blot ét multispektralt billede af hvert emne.
Måling af enzymatiske effekter med spektral billedanalyse er således muligt.

5.1 Protokol for måling på stoflapper

Herunder er beskrevet en protokol for brugen af den udviklede metode til måling af enzymatiske effekter med spektral billedanalyse. I protokollen er det antaget, at der benyttes VideometerLab til at tage multispektrale billeder af stoflapperne og at databehandlingen foregår i VideometerLab softwaren. Den reelle protokol vil afhænge af den konkrete implementering i VideometerLab softwaren, hvilket ikke er en del af dette projekt. Der kan derfor forekomme mindre afvigelser mellem protokollen beskrevet herunder og den faktiske protokol, såfremt denne implementeres i praksis.

Bemærk, at selvom metoden er udviklet ved brug af CDA, benyttes der **n**CDA i den praktiske udførelse. Dette skyldes, at det er nødvendigt at kontrollere orienteringen af CDFen, således at renhedsskalaen ikke arbitrært skifter fortegn[1].

Måling af renheden af stoflapper foregår efter fremgangsmåden:

1. Stoflapperne vaskes enten i en LOM[2] eller en fuldskala vaskemaskine. Der benyttes den til enhver tid gældende SOP[3].
2. Det trin i gældende SOP der omhandler remissionsmåling/intensitetsmåling kan springes over.
   Alternativt kan remissionsmålingen/intensitetsmålingen udføres og derefter sammenlignes med resultatet af MLR modellen.
3. Senest 24 timer efter endt vask tages der multispektrale billeder af stoflapperne. Der tages ét multispektralt billede af den éne side af hver stoflap. Det er ikke nødvendigt at holde styr på, om der benyttes "forsiden" eller "bagsiden" af de enkelte stoflapper.

[1] Se evt. teorien beskrevet i afsnit 2.1.1 på side 11.

[2] Launder-O-Meter, benyttes til at simulere en typisk europæisk automatisk vaskemaskine i lille skala.

[3] Standard operating procedure. Stoflapperne omtalt i kapitel 4 er vasket i LOM efter SOP No. CS-SM-2000.uk version 4.0.

4. Næste trin afhænger af, om det er første gang den pågældende kombination af plettype[4] og stoffarve[5] benyttes. For hver kombination gøres følgende.

   Ny kombination

   (a) Der skal genereres en nCDA[6] transformation og en MLR[7] model.

   (b) I VideometerLab softwaren udvælges et repræsentativt område af én af de pletter, der forventes at være gået *mindst* af i vask. Der kan med fordel vaskes en stoflap uden enzym til det dette formål.

   (c) I VideometerLab softwaren udvælges et repræsentativt område af én af de pletter, der forventes at være gået *mest* af i vask. Benyt f.eks. pletten på en af de stoflapper der er vasket ved højest enzymkoncentration.

   (d) VideometerLab softwaren beregner nu den nødvendige nCDA transformation og derefter MLR modellen[8]. MLR modellen gemmes i en database, så den kan anvendes igen ved fremtidige målinger på samme kombination af plettype og stoffarve.

   Tidligere benyttet kombination

   Fra en database i VideometerLab softwaren vælges MLR modellen hørende til pågældende kombination af plettype og stoffarve.

5. VideometerLab softwaren beregner nu de relative renheder af stoflapperne, for hver kombination af plettype og stoffarve.
   For at sikre at beregningstiden er mindst mulig, benyttes metoden til effektiv leave-one-out krydsvalidering som beskrevet i afsnit 2.5 på side 19.

På baggrund af de relative renheder kan det konkluderes, hvilke vaskebetingelser der fører til de reneste stoflapper.
Hvis metoden tidligere har været anvendt på pågældende kombination af plettype og stoffarve forventes det at renheden af en stoflap kan måles på 5-10 sekunder[9]. Hvis metoden ikke tidligere har været benyttet på pågældende kombination af plettype og stoffarve forventes initialiseringen at tage cirka et minuts brugerinteraktion og under et minuts beregningstid.

[4] F.eks. Warwick Equests stain type 131BKC - Cooked beef fat.
[5] F.eks. Warwick Equests hvide og blå stoflapper.
[6] Normalized canonical discriminant analysis, evt. se teorien i afsnit 2.1.1 på side 11.
[7] Multiple lineær regression, evt. se teorien i afsnit 2.3 på side 15.
[8] Pletten isoleres automatisk i billedet ved brug af en nCDA transformation beregnet mellem det rene stof og pletten. Den samme nCDA transformation benyttes på alle stoflapperne. Gauss pyramiderne, der benyttes i træningen af MLR modellen, beregnes kun af de isolerede pletter.
[9] Det tager 5-10 sekunder at tage ét multispektralt billede med VideometerLab. www.videometer.com.

5.2 Absolutte renheder

Hvis det ønskes at sammenligne renheden af stoflapper med to forskellige plettyper, er det nødvendigt at kende den absolutte renhed af stoflapperne.
Efter nedenstående fremgangsmåde er det formentlig også muligt at opnå absolutte renheder.

1. Fastsæt en vaskeprocedure der kan reproduceres for alle plettyper.
2. Vask to stoflapper fra samme batch ved den fastsatte vaskeprocedure, dog hvor den ene af stoflapperne vaskes uden enzym.
   (Hvis der ønskes øget robusthed[10] over for den høje batch til batch variation i pletterne kan alternativt vaskes f.eks. 10 stoflapper uden brug af enzym samt f.eks. 10 stoflapper ved den fastsatte vaskeprocedure. Alle stoflapperne vaskes ved den fastsatte vaskeprocedure. Disse 20 stoflapper skal parvist være fra 10 forskellige batches.)
3. Tag multispektrale billeder af disse 2 (20) stoflapper. Dette gøres tilsvarende punkt 3 herover.
4. Ved beregning af between groups matricen (2.1) og within groups matricen (2.2) benyttes et repræsentativt udsnit af de 2 (20) pletter i X.
   (Hvis der blev benyttet 20 stoflapper, er der opnået en nCDA transformation, der er mere robust over for batch til batch variationen i både stoflapperne og pletterne. Se evt. teorien om CDA i afsnit 2.1 på side 8.)
5. Baser MLR modellen på denne nCDA transformation.

Ved at følge ovenstående fremgangsmåde for forskellige plettyper kan der formentlig opnås målinger af den absolutte renhed, som således kan sammenlignes på tværs af både plettyper og stoffarver. Dette fordi de to grupper i nCDA transformationen er af fast renhed og er robuste over for batch til batch variation.

[10] Metoden er allerede vist robust i afsnit 4.6.1.

5.3 Videre arbejde

Det vil være interessant at undersøge om:

1. Metoden også virker på andre kombinationer end blå stoflapper med lipidpletter af typen cooked beef fat. Dette forventes, men det er ikke definitivt afklaret. Dette er nemt og hurtigt at undersøge, hvis der vaskes f.eks. 40 stoflapper af f.eks. hvidt stof med en plettype, hvor renheden kan måles med remission. Modellens generalitet kan testes ved brug af multispektrale billeder af disse stoflapper.
2. Det er muligt at benytte den samme nCDA transformation og MLR model på flere forskellige kombinationer plettyper og stoffarver. Hvis det er muligt, vil det gøre metoden simplere ved at mindske behovet for brugerinteraktion.
3. Det er muligt at benytte ideen[11] til at modificere nCDA således, at der opnås målinger af absolutte renheder i stedet for relative renheder.

Efter undersøgelsen af punkt 1 kan der træffes endelig beslutning om metodens praktiske anvendelighed i Novozymes. Det er min vurdering, at metoden kan benyttes i Novozymes, hvis blot punkt 1 kan opnås. Såfremt andet og/eller tredje punkt også kan opnås vil det yderligere øge anvendeligheden af metoden.

[11] Beskrevet i afsnit 5.2 på modstående side.

Bilag A

Implementering af LOOKV

Dette bilag indeholder Matlab implementeringen af effektiv leave-one-out krydsvalidering som gennemgået i afsnit 2.5 på side 19.

```
function [ SKR_K ] = EffektivLOOKV(X, y, dummy1, dummy2)
%EffektivLOOKV Effektiv leave-one-out krydsvalidering.
%   Denne funktion foretager effektiv leave-one-out krydsvalidering som
%   gennemgået i afsnit 2.5. Herved er det muligt momentant at beregne
%   summen af de kvadrerede leave-one-out krydsvaliderede residualer
%   ud fra et datasæt (og dermed hatmatricen), uden at skulle beregne
%   modellen om og om igen, n gange.
%
%   EffektivLOOKV kaldes med sequentialfs således:
%   [F, H] = sequentialfs(@EffektivLOOKV, X_train, y_train,'cv','none');
%
%   Input variablerne dummy1 og dummy2 er nødvendige, da sequentialfs
%   forventer, at objektfunktionen har fire inputvariable.

% Tilføjer en søjle med et-taller til designmatricen,
% hvorved der medtages et konstant led i MLR modellen.
X=[ones(size(X,1),1) X];

% Beregner hatmatricen
H=X/(X'*X)*X'; % H=X*inv(X'*X)*X';

% N = Antal observationer.
% M = Antal attributter.
[N M]=size(X);

% Beregner summen af kvadrerede leave-one-out krydsvaliderede residualer
SKR_K=(y'*(eye(N)-H))*((diag(eye(N)-H).^(-2)).*((eye(N)-H)*y));

end
```

Bilag B

Eksempeldata

Dette bilag indeholder eksempeldataet benyttet i eksemplet af en hypotesetest for en multiple lineær regressionsmodel. Eksemplet står på side 32.

$$X = \begin{pmatrix} 1 & 36,1 & 19,1 \\ 1 & 45,5 & 19,1 \\ 1 & 38,9 & 19,3 \\ 1 & 44,8 & 19,7 \\ 1 & 41,6 & 19,9 \\ 1 & 46,1 & 19,7 \\ 1 & 48,5 & 20,0 \\ 1 & 53,6 & 20,0 \\ 1 & 57,2 & 20,3 \\ 1 & 57,4 & 20,6 \\ 1 & 57,4 & 20,4 \\ 1 & 63,0 & 20,8 \\ 1 & 65,0 & 21,1 \\ 1 & 65,0 & 21,1 \\ 1 & 68,5 & 21,7 \\ 1 & 70,0 & 21,2 \\ 1 & 67,9 & 21,3 \\ 1 & 75,2 & 21,2 \\ 1 & 73,0 & 22,1 \\ 1 & 75,2 & 22,3 \end{pmatrix} \qquad y = \begin{pmatrix} 243,6 \\ 243,6 \\ 264,9 \\ 256,5 \\ 271,2 \\ 276,2 \\ 292,4 \\ 286,1 \\ 289,1 \\ 298,1 \\ 303,3 \\ 307,7 \\ 309,5 \\ 311,6 \\ 324,5 \\ 321,4 \\ 328,2 \\ 345,2 \\ 341,7 \\ 348,3 \end{pmatrix}$$

Tabel B.1: *Samtidige observationer af udbyttet y_i og de to variable x_{i1} og x_{i2}.*

Observationsnummer	Variabel 1 (x_{i1})	Variabel 2 (x_{i2})	Udbyttet (y_i)
1	36,1	19,1	243,6
2	45,5	19,1	243,6
3	38,9	19,3	264,9
4	44,8	19,7	256,5
5	41,6	19,9	271,2
6	46,1	19,7	276,2
7	48,5	20,0	292,4
8	53,6	20,0	286,1
9	57,2	20,3	289,1
10	57,4	20,6	298,1
11	57,4	20,4	303,3
12	63,0	20,8	307,7
13	65,0	21,1	309,5
14	65,0	21,1	311,6
15	68,5	21,7	324,5
16	70,0	21,2	321,4
17	67,9	21,3	328,2
18	75,2	21,2	345,2
19	73,0	22,1	341,7
20	75,2	22,3	348,3

Bilag C

Tekstur sammenligning, øvrige statistikker

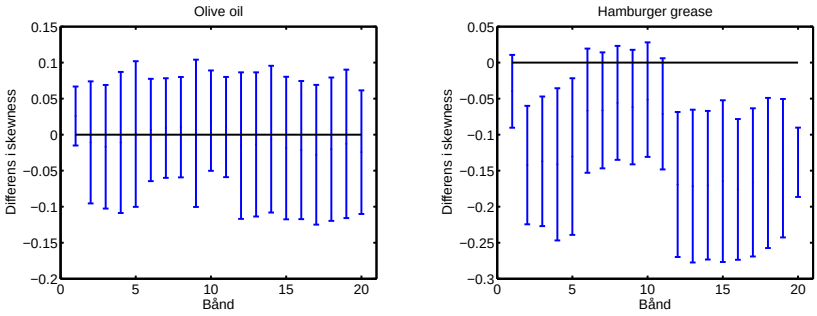
Dette bilag indeholder de øvrige statistikker, skewness, kurtosis, energi og entropi, fra tekstur sammenligningen af stoflapper med olive oil spread og hamburger grease fedtpletter fra afsnit 3.3 på side 50. Beregningerne og argumentationen er tilsvarende afsnit 3.3.

C.1 Skewness

For hvert bånd er skewness beregnet for hver af de fire inddelinger af olive oil spread og hamburger grease pletterne ved henholdsvis dosis 0,05% og 0,25%. Differensen mellem de to dosis er beregnet og plottet på figur C.1. Standardafvigelsen af skewness for de fire dele af pletterne er betragtet som mål for metodens usikkerhed og plottet som usikkerhedsbarer på figuren. Hvis metodens usikkerhed overstiger differensen mellem de to dosis forkastes metoden for pågældende bånd. Metoden forkastes altså for et givent bånd, hvis usikkerhedsbaren for givende bånd overlapper nul.

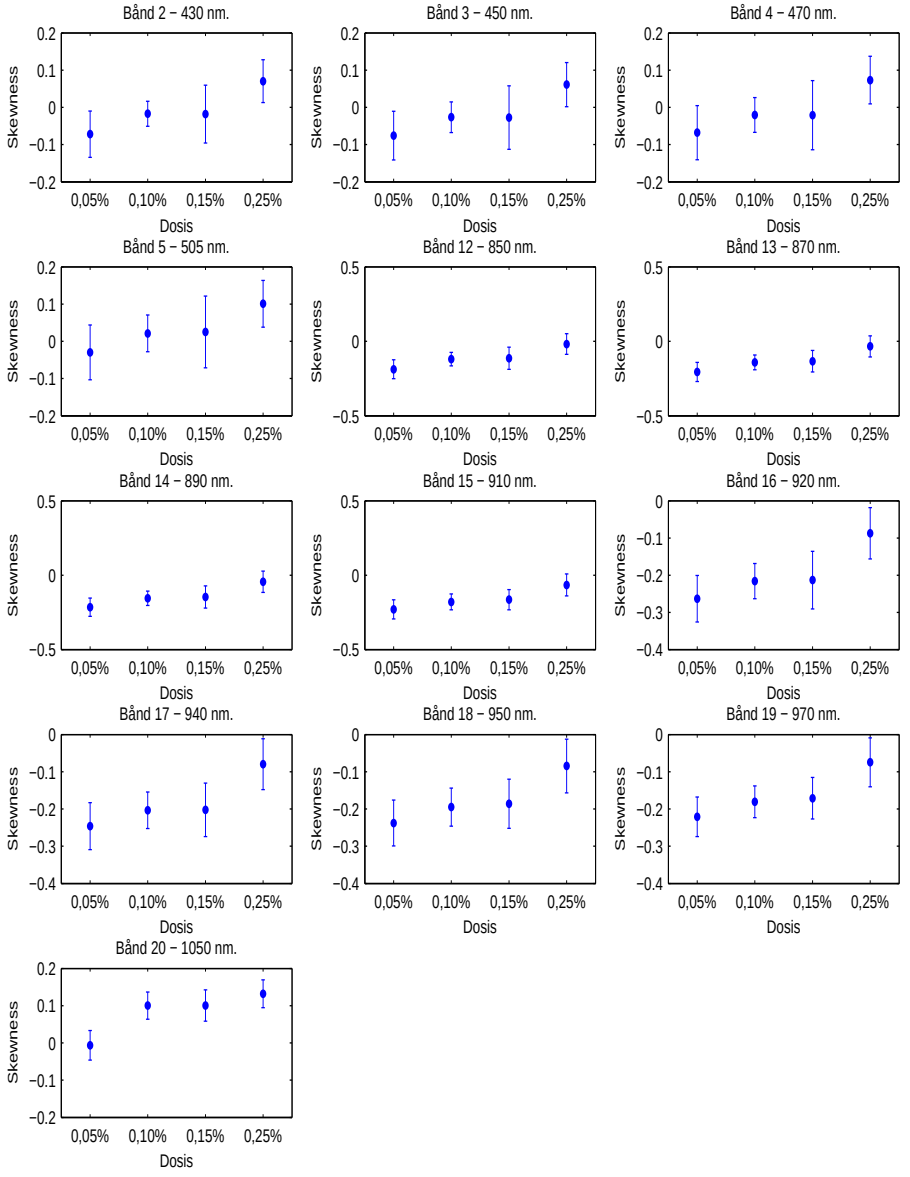
De bånd hvor usikkerheden på skewness ikke overstiger differensen mellem dosis 0,05% og 0,25% undersøges nærmere. Hvis skewness ikke kan adskille yderpunkterne i dosis er ingen grund til at undersøge om skewness kan adskille øvrige dosis.
Af figur C.1 fremgår det at skewness ikke kan adskille yderpunkterne i dosis for et eneste bånd for olive oil spread pletterne. Derfor forkastes metoden for plettypen olive oil spread.

Det er således kun bånd fra hamburger grease pletterne der undersøges nærmere. Resultatet fra denne undersøgelse fremgår af figur C.2.
Af figur C.1 og C.2 kan det samlet konkluderes at skewness ikke kan adskille de enkelte dosis i rækkefølge for nogen bånd, hverken for plettypen olive oil spread eller hamburger grease. Metoden forkastes derfor.



(a) **(b)**

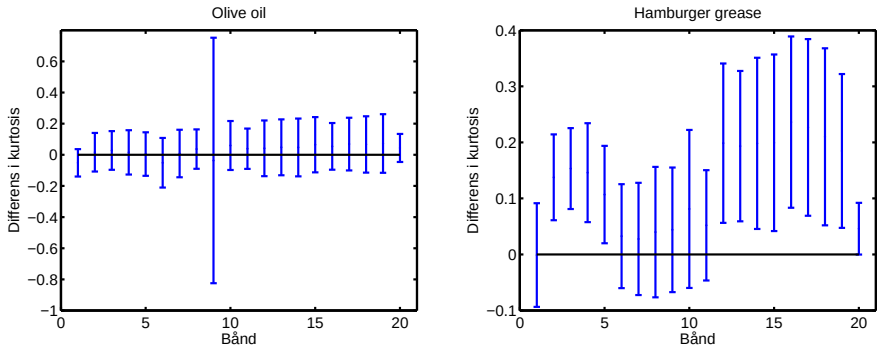
Figur C.1: *Differens i skewness som funktion af bånd for olive oil spread henholdsvis hamburger grease.*
For olive oil spread forkastes metoden for alle bånd.
For hamburger grease forkastes metoden for bånd 1 og bånd 6 til 11.



Figur C.2: *Skewness for hver dosis for de bånd der ikke blev forkastet af figur C.1. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9 på side 50.*
Det ses at skewness ikke unikt kan adskille de enkelte dosis i rækkefølge for nogen af ovenstående bånd.

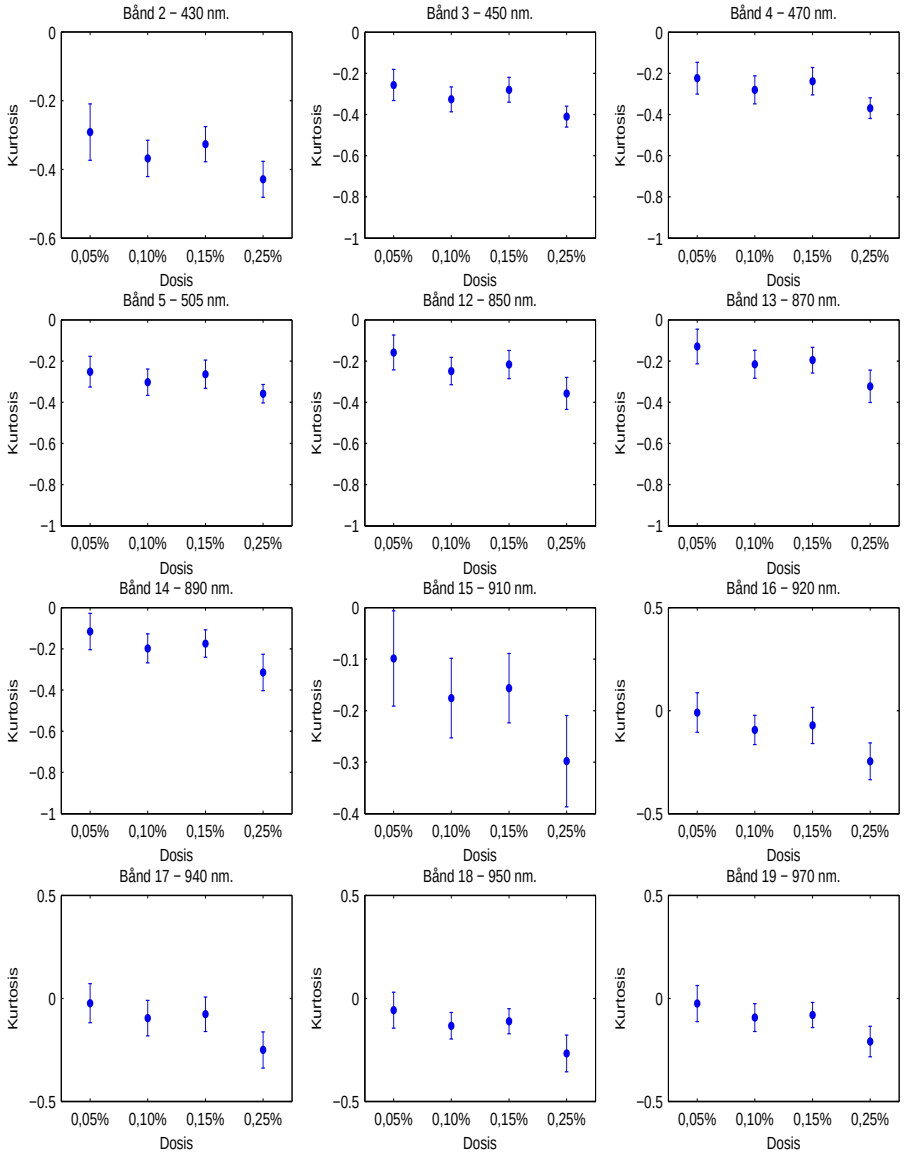
C.2 Kurtosis

For hvert bånd er kurtosis beregnet for hver af de fire inddelinger af olive oil spread og hamburger grease pletterne ved henholdsvis dosis 0,05% og 0,25%. Differensen mellem de to dosis er beregnet og plottet på figur C.3. Standardafvigelsen af kurtosis for de fire dele af pletterne er betragtet som mål for metodens usikkerhed og plottet som usikkerhedsbarer på figuren. Hvis metodens usikkerhed overstiger differensen mellem de to dosis forkastes metoden for pågældende bånd. Metoden forkastes altså for et givent bånd, hvis usikkerhedsbaren for det givende bånd overlapper nul.



Figur C.3: *Differens i kurtosis som funktion af bånd for olive oil spread henholdsvis hamburger grease.*
For olive oil spread forkastes metoden for alle bånd.
For hamburger grease forkastes metoden for bånd 1, bånd 6 til 11 og bånd 20.

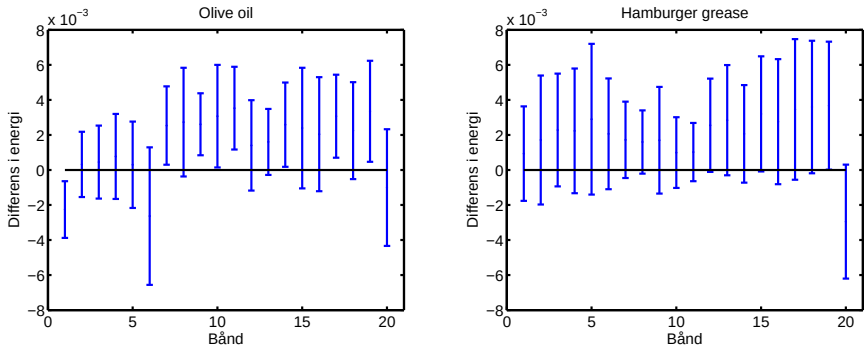
De bånd hvor usikkerheden på kurtosis ikke overstiger differensen mellem dosis 0,05% og 0,25% undersøges nærmere. Hvis kurtosis ikke kan adskille yderpunkterne i dosis er der ingen grund til at undersøge om kurtosis kan adskille øvrige dosis.
Af figur C.3 fremgår det at kurtosis ikke kan adskille yderpunkterne i dosis for et eneste bånd for olive oil spread pletterne. Derfor forkastes metoden for plettypen olive oil spread.
Det er således kun bånd fra hamburger grease pletterne der undersøges nærmere. Resultatet fra denne undersøgelse fremgår af figur C.4.
Af figur C.3 og C.4 kan det samlet konkluderes at kurtosis ikke kan adskille de enkelte dosis i rækkefølge for nogen bånd, hverken for plettypen olive oil spread eller hamburger grease. Metoden forkastes derfor.



Figur C.4: *Kurtosis for hver dosis for de bånd der ikke blev forkastet af figur C.3. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9 på side 50.*
Det ses at kurtosis ikke unikt kan adskille de enkelte dosis i rækkefølge for nogen af ovenstående bånd.

C.3 Energi

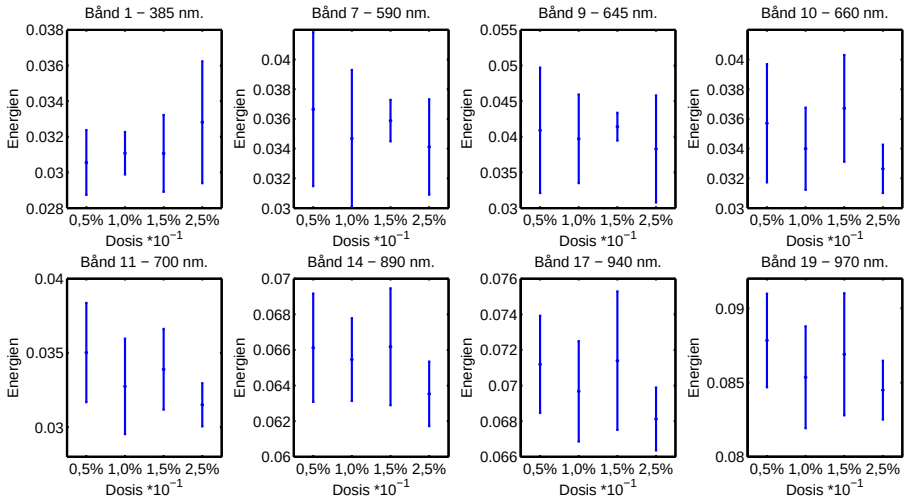
For hvert bånd er energien beregnet for hver af de fire inddelinger af olive oil spread og hamburger grease pletterne ved henholdsvis dosis 0,05% og 0,25%. Differensen mellem de to dosis er beregnet og plottet på figur C.5. Standardafvigelsen af energien for de fire dele af pletterne er betragtet som mål for metodens usikkerhed og plottet som usikkerhedsbarer på figuren. Hvis metodens usikkerhed overstiger differensen mellem de to dosis forkastes metoden for pågældende bånd. Metoden forkastes altså for et givent bånd, hvis usikkerhedsbaren for det givende bånd overlapper nul.



Figur C.5: *Differens i energi som funktion af bånd for olive oil spread henholdsvis hamburger grease.*
For olive oil spread forkastes metoden for bånd 2 til 6, bånd 8, 12, 13, 15, 16, 18 og 20.
For hamburger grease forkastes metoden for alle bånd på nær bånd 19.

De bånd hvor usikkerheden på energien ikke overstiger differensen mellem dosis 0,05% og 0,25% undersøges nærmere. Hvis energien ikke kan adskille yderpunkterne i dosis er der ingen grund til at undersøge om energien kan adskille øvrige dosis. Af figur C.5 fremgår det at både bånd fra olive oil spread og hamburger grease pletterne skal undersøges nærmere.

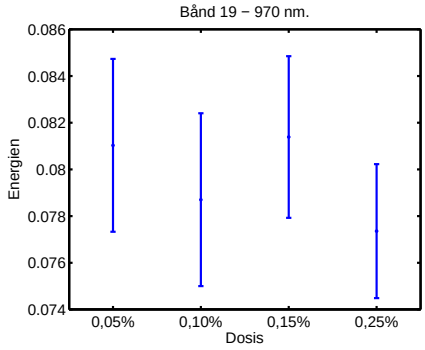
Nærmere undersøgelse af olive oil spread
På figur C.6 ses energien for hver dosis i dosis response forsøget. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9.
Som det fremgår af figuren kan de enkele dosis ikke adskilles i rækkefølge af energien ved nogen af de pågældende bånd. For olive oil spread må metoden derfor forkastes.



Figur C.6: *Energien beregnet for hver dosis for de otte bånd aflæst af figur C.5. Det ses at energien ikke unikt kan adskille de enkelte dosis i rækkefølge for nogen af de pågældende bånd.*

Nærmere undersøgelse af hamburger grease

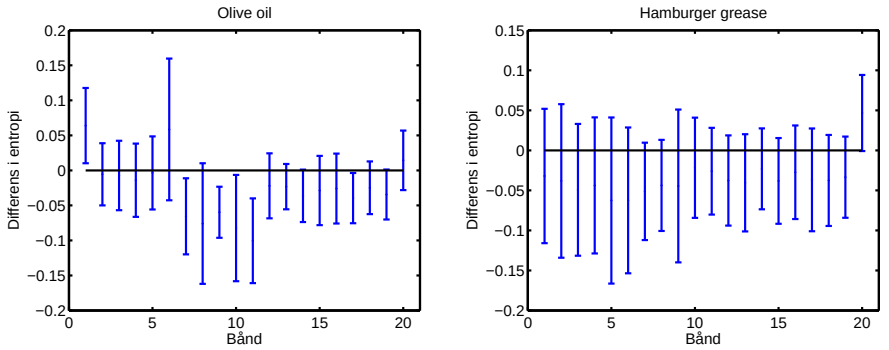
På figur C.5 ses at for hamburger grease er det kun bånd 19 hvor usikkerheden ikke overstiger differensen mellem dosisyderpunkterne 0,05% og 0,25%. Bånd 19 er derfor undersøgt nærmere. På figur C.7 ses energien for bånd 19 for alle dosis i dosis response forsøget. Det ses at usikkerheden er for stor til at metoden i rækkefølge kan adskille de enkelte dosis. Metoden forkastes derfor for hamburger grease.



Figur C.7: *Energien af bånd 19 beregnet for hver dosis. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9. De enkelte dosis kan ikke unikt adskilles i rækkefølge.*

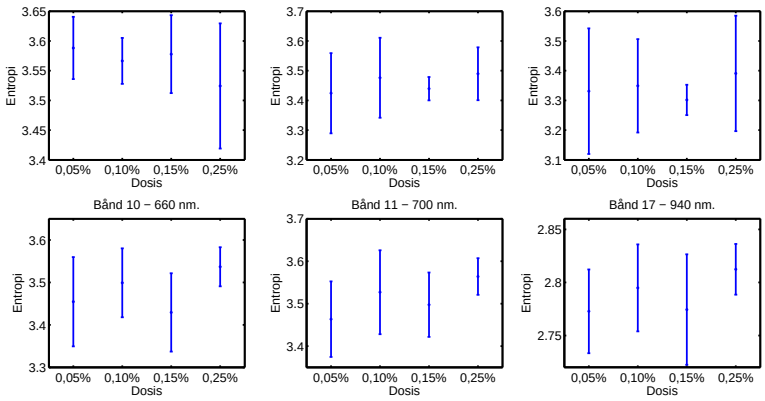
C.4 Entropi

For hvert bånd er entropien beregnet for hver af de fire inddelinger af olive oil spread og hamburger grease pletterne ved henholdsvis dosis 0,05% og 0,25%. Differensen mellem de to dosis er beregnet og plottet på figur C.8. Standardafvigelsen af entropien for de fire dele af pletterne er betragtet som mål for metodens usikkerhed og plottet som usikkerhedsbarer på figuren. Hvis metodens usikkerhed overstiger differensen mellem de to dosis forkastes metoden for pågældende bånd. Metoden forkastes altså for et givent bånd, hvis usikkerhedsbaren for det givende bånd overlapper nul.



Figur C.8: *Differens i entropi som funktion af bånd for olive oil spread henholdsvis hamburger grease.*
For olive oil spread forkastes metoden for bånd 2-6, bånd 8, bånd 12-16 og bånd 18-20.
For hamburger grease forkastes metoden for alle bånd.

De bånd hvor usikkerheden på entropien ikke overstiger differensen mellem dosis 0,05% og 0,25% undersøges nærmere. Hvis entropien ikke kan adskille yderpunkterne i dosis er der ingen grund til at undersøge om entropien kan adskille øvrige dosis.
Af figur C.8 fremgår det at entropien ikke kan adskille yderpunkterne i dosis for et eneste bånd for hamburger grease pletterne. Derfor forkastes metoden for plettypen hamburger grease.
Det er således kun bånd fra olive oil spread pletterne der undersøges nærmere. Resultatet fra denne undersøgelse fremgår af figur C.9.
Af figur C.8 og C.9 kan det samlet konkluderes at entropien ikke kan adskille de enkelte dosis i rækkefølge for nogen bånd, hverken for plettypen olive oil spread eller hamburger grease. Metoden forkastes derfor.

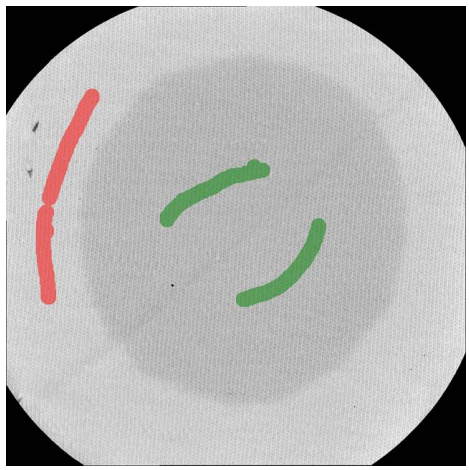


Figur C.9: *Entropien for hver dosis for de bånd der ikke blev forkastet af figur C.8. Usikkerhedsbaren for hver dosis er beregnet som standardafvigelsen mellem de fire dele af billedet af pågældende dosis. De fire dele er illustreret på figur 3.9 på side 50.*
Det ses at entropien ikke unikt kan adskille de enkelte dosis i rækkefølge for nogen af ovenstående bånd.

Bilag D

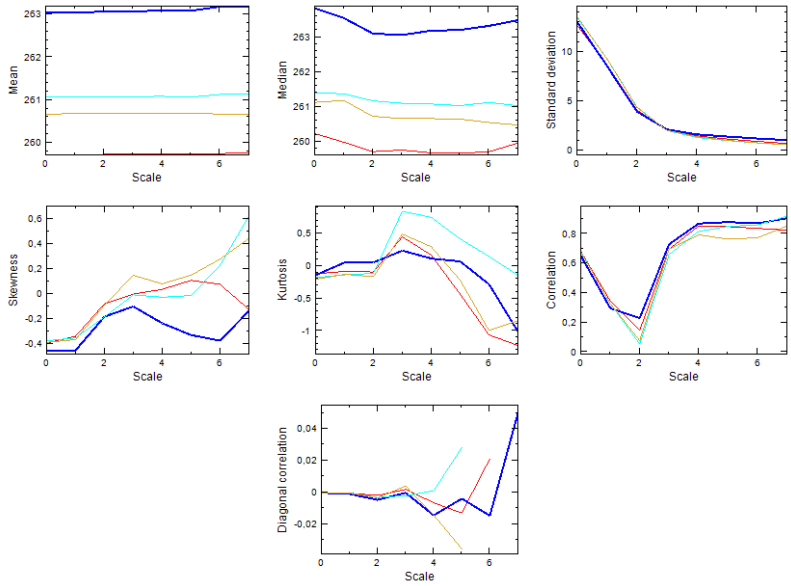
Gauss og Laplacian pyramider

Dette bilag indeholder de Gauss og Laplacian pyramider der er henvist til i afsnit 3.4.2 på side 57 til afsnit 3.4.4 på side 61.

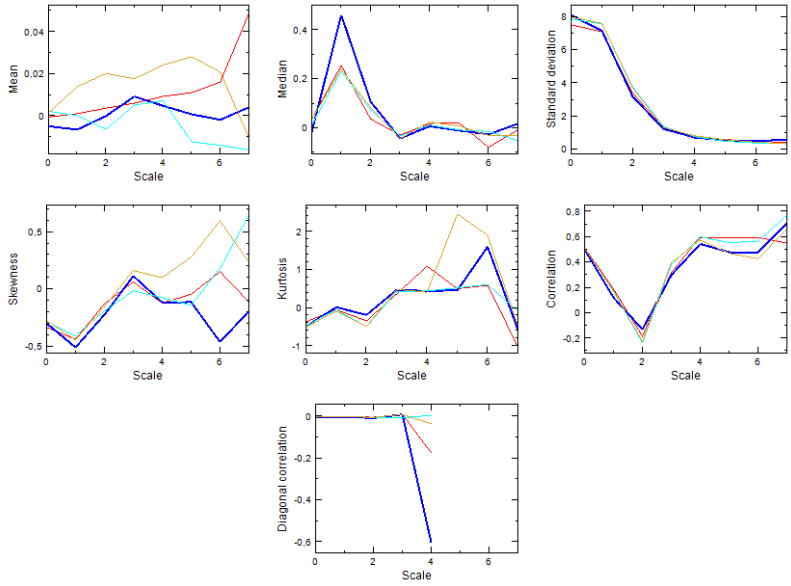


Figur D.1: *De malede områder for olive oil spread dosis 0,05%.*

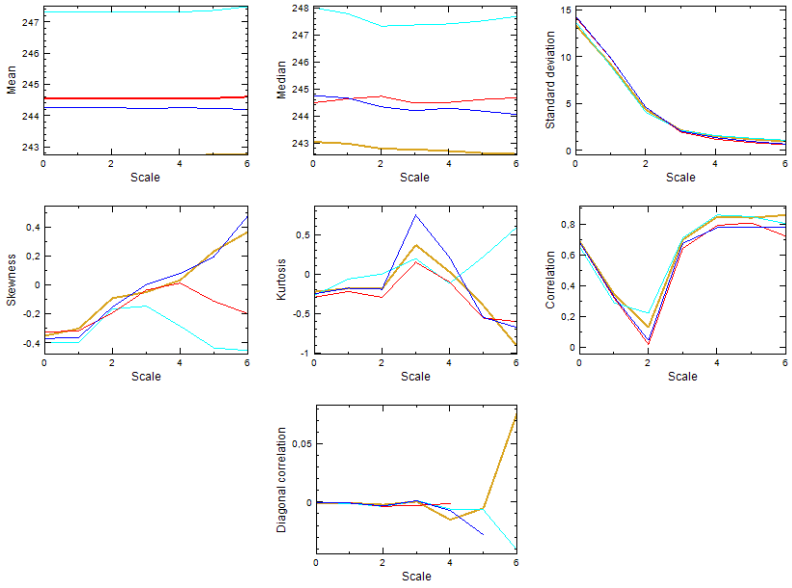
Figur D.2: *De malede områder for olive oil spread dosis 0,05% (rød) og 0,25% (grøn).*



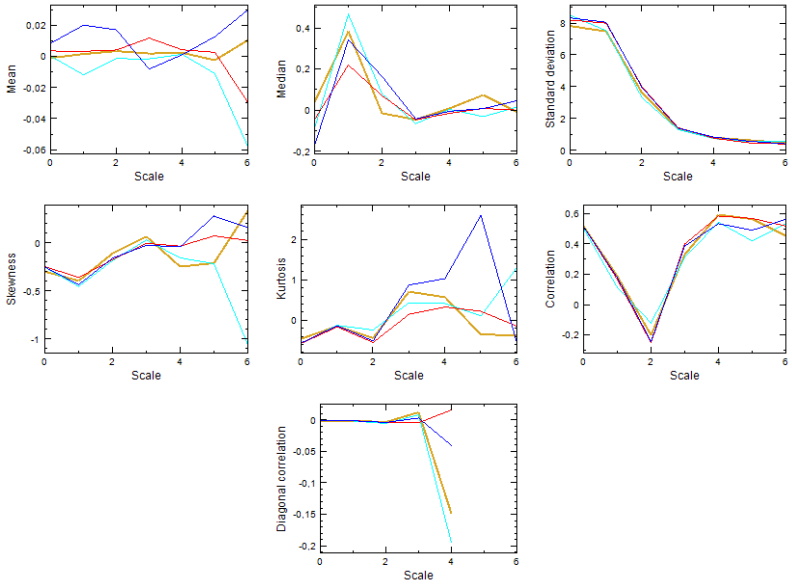
Figur D.3: *Statistikker beregnet i Gauss pyramider af den første principal komponent. Transformationen er beregnet på hele billedet af olive oil spread dosis 0,05% og anvendt på alle dosis.*
Brun: 0,05%, blå: 0,10%, rød: 0,15%, turkis: 0,25%.



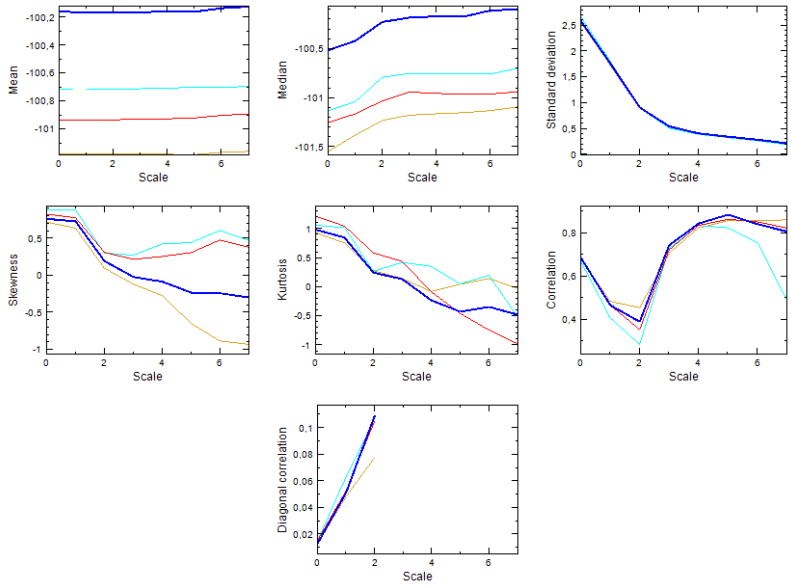
Figur D.4: *Statistikker beregnet i Laplacian pyramider af den første principal komponent. Transformationen er beregnet på hele billedet af olive oil spread dosis 0,05% og anvendt på alle dosis.*
Brun: 0,05%, blå: 0,10%, rød: 0,15%, turkis: 0,25%.



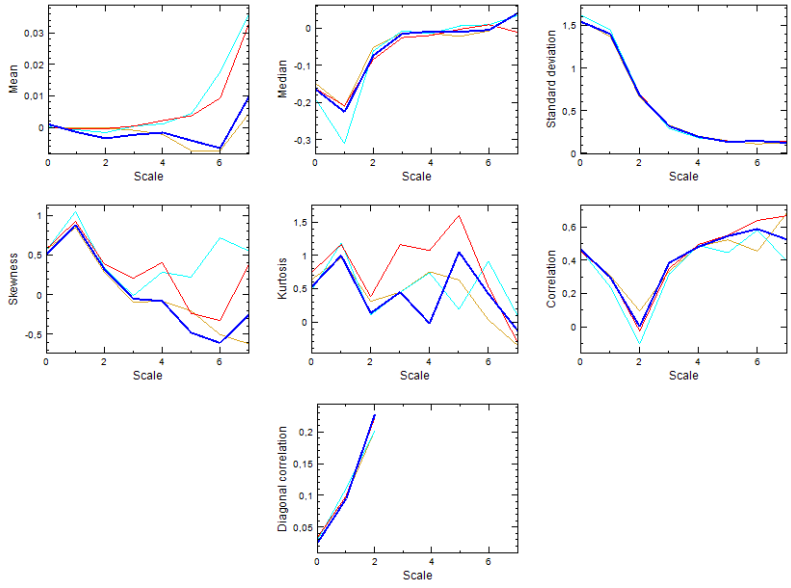
Figur D.5: *Statistikker beregnet i Gauss pyramider af den første principal komponent. Transformationen er beregnet på et udklip af fedtpletten af olive oil spread dosis 0,05% og anvendt på udklip af fedtpletterne for alle dosis. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, brun: 0,25%.*



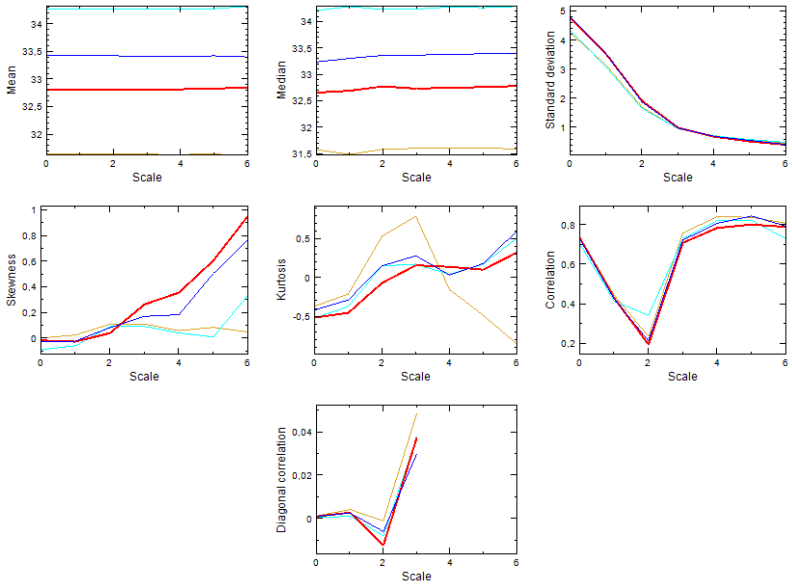
Figur D.6: *Statistikker beregnet i Laplacian pyramider af den første principal komponent. Transformationen er beregnet på et udklip af fedtpletten af olive oil spread dosis 0,05% og anvendt på udklip af fedtpletterne for alle dosis. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, brun: 0,25%.*



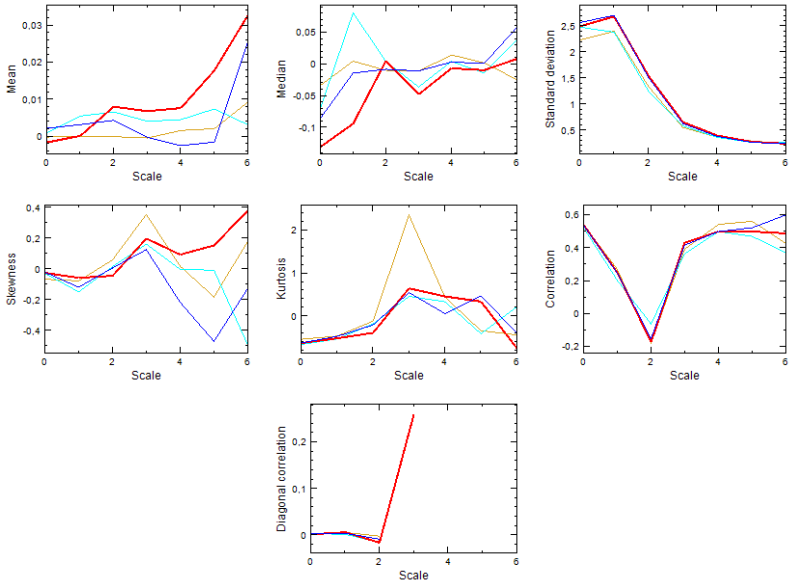
Figur D.7: *Statistikker beregnet i Gauss pyramider af den første MNF komponent. Transformationen er beregnet på hele billedet af olive oil spread dosis 0,05% og anvendt på alle dosis.*
Turkis: 0,05%, brun: 0,10%, rød: 0,15%, blå: 0,25%.



Figur D.8: *Statistikker beregnet i Laplacian pyramider af den første MNF komponent. Transformationen er beregnet på hele billedet af olive oil spread dosis 0,05% og anvendt på alle dosis.*
Turkis: 0,05%, brun: 0,10%, rød: 0,15%, blå: 0,25%.



Figur D.9: *Statistikker beregnet i Gauss pyramider af den første MNF komponent. Transformationen er beregnet på et udklip af fedtpletten af olive oil spread dosis 0,05% og anvendt på udklip af fedtpletterne for alle dosis. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, brun: 0,25%.*

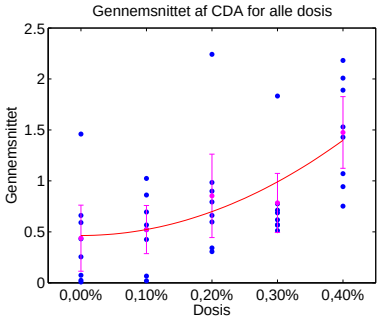


Figur D.10: *Statistikker beregnet i Gauss pyramider af den første MNF komponent. Transformationen er beregnet på et udklip af fedtpletten af olive oil spread dosis 0,05% og anvendt på udklip af fedtpletterne for alle dosis. Blå: 0,05%, rød: 0,10%, turkis: 0,15%, brun: 0,25%.*

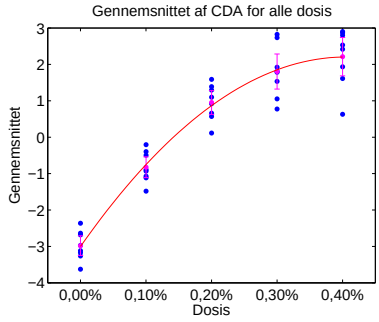
Bilag E

Stabiliteten af CDA

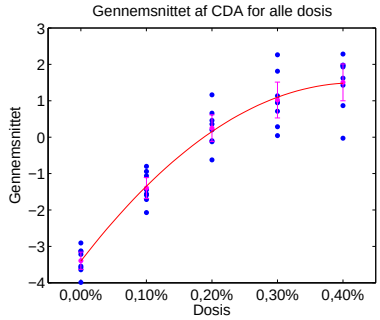
Dette bilag indeholder de resterende fem illustrationer af stabiliteten af CDA som omtalt i afsnit 4.3.2 på side 74.



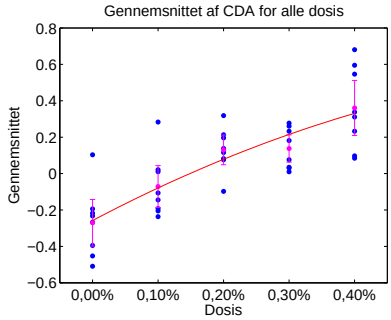
Figur E.1: *CDA transformation nr. 2.*



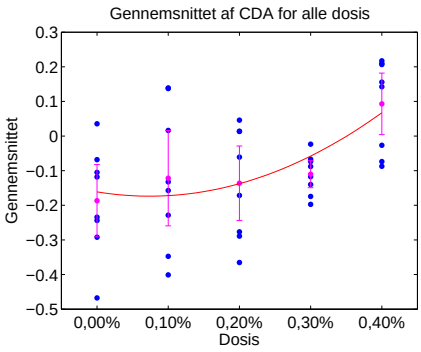
Figur E.2: *CDA transformation nr. 3.*



Figur E.3: *CDA transformation nr. 4.*



Figur E.4: *CDA transformation nr. 6.*



Figur E.5: *CDA transformation nr. 7.*

Bilag F

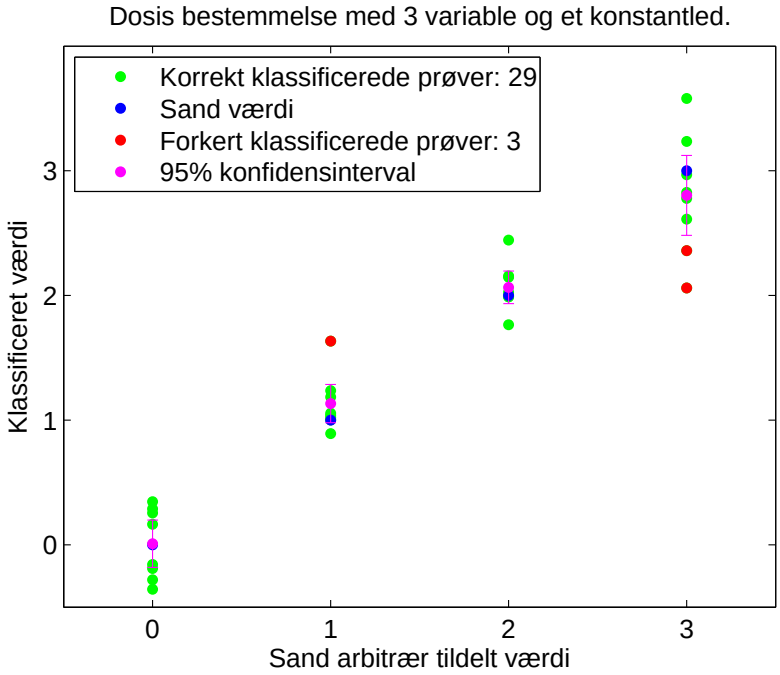
Variationer af modellen

Dette bilag indeholder plots for variation 2 til 5 af modellen beskrevet i afsnit 4.5.1 på side 82.

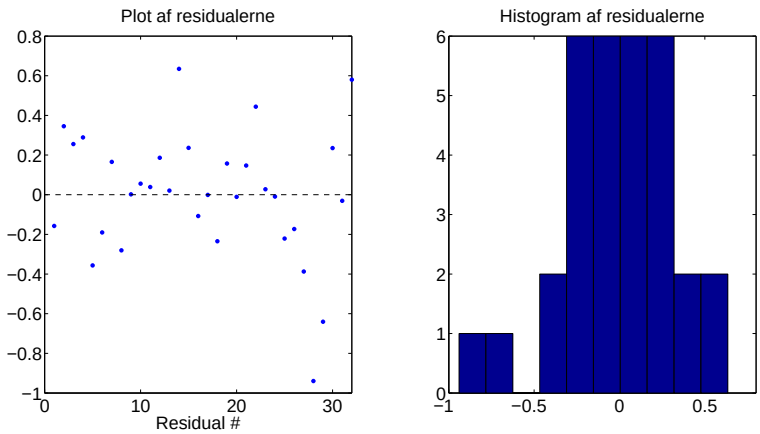
Plot af resultatet ved krydsvalidering for variation 1 ses af figur 4.10 på side 81.
Plot af residualerne for variation 1 ses af figur 4.11 på side 81.

Dette bilag indeholder desuden nedenstående programudskrift som refereret til i afsnit 4.5.1 på side 83.

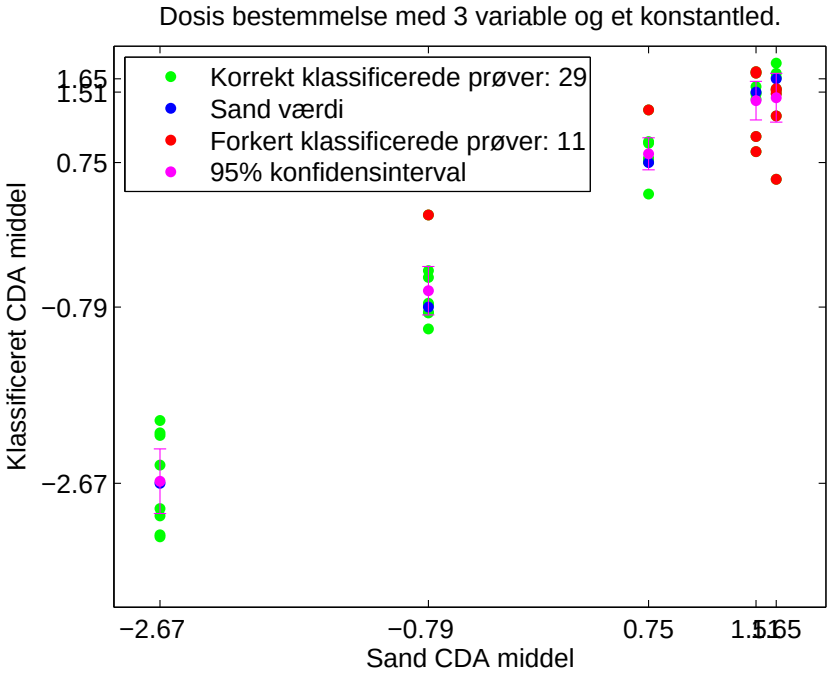
```
Henze-Zirkler's Multivariate Normality Test
---------------------------------------------------------------------
Number of variables: 20
Sample size: 10000
---------------------------------------------------------------------
Henze-Zirkler lognormal mean: -0.0000970
Henze-Zirkler lognormal variance: 0.0001472
Henze-Zirkler statistic: 1.0213302
P-value associated to the Henze-Zirkler statistic: 0.0000000
With a given significance = 0.050
---------------------------------------------------------------------
Data analyzed do not have a normal distribution.
---------------------------------------------------------------------
```



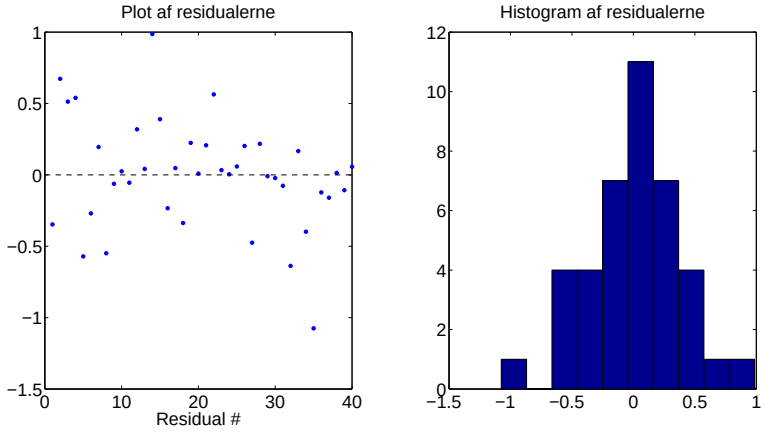
Figur F.1: *Resultatet ved krydsvalidering for variation 2.*



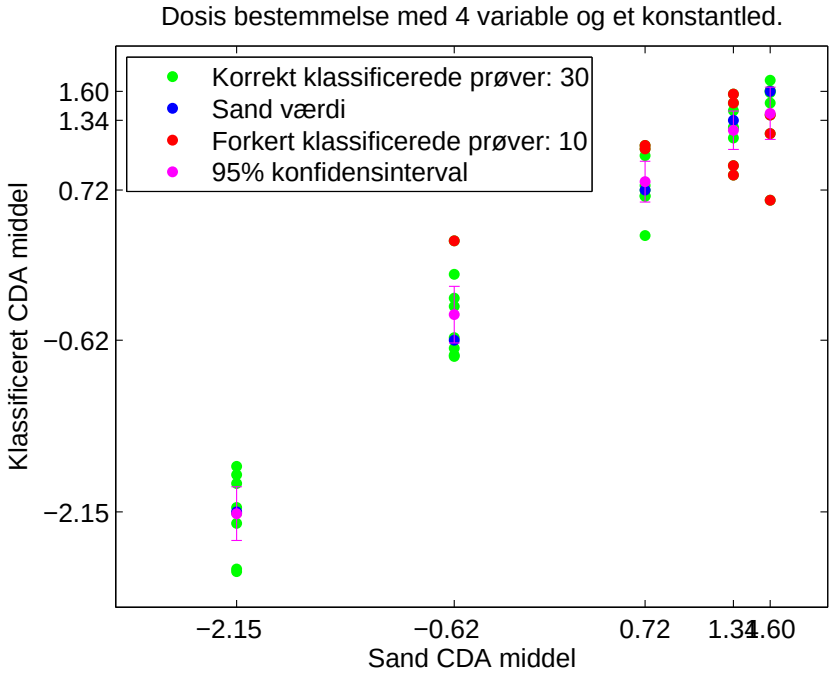
Figur F.2: *Residualer for variation 2.*



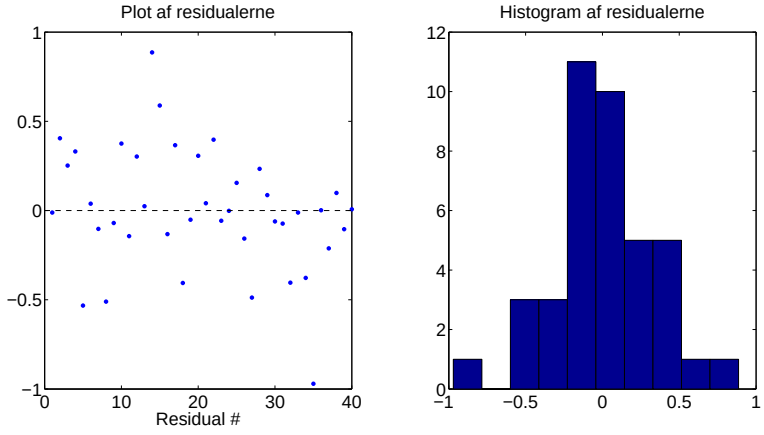
Figur F.3: *Resultatet ved krydsvalidering for variation 3.*



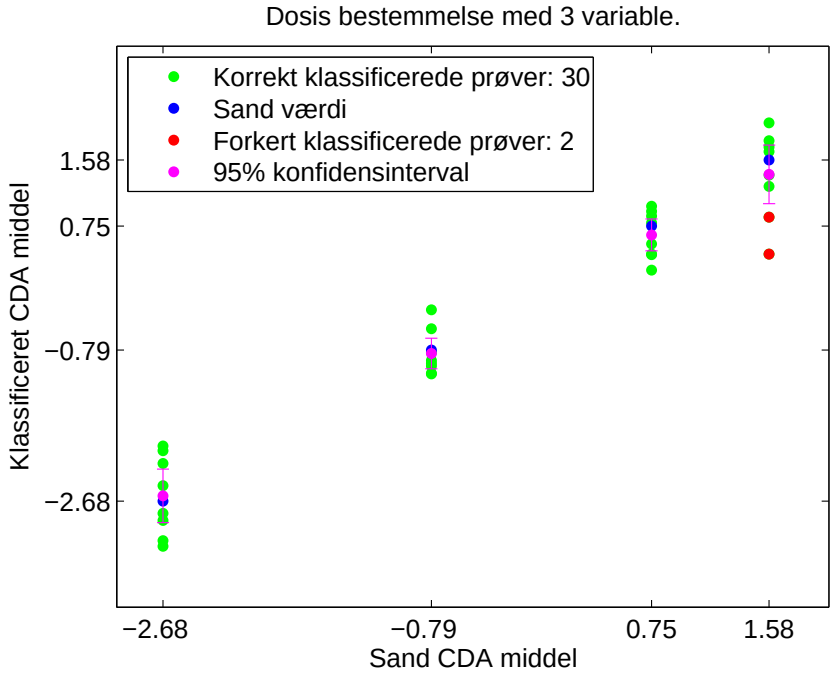
Figur F.4: *Residualer for variation 3.*



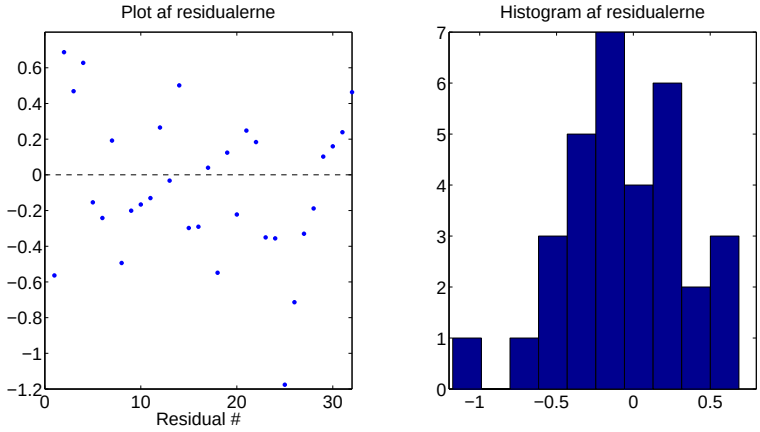
Figur F.5: *Resultatet ved krydsvalidering for variation 4.*



Figur F.6: *Residualer for variation 4.*



Figur F.7: *Resultatet ved krydsvalidering for variation 5.*



Figur F.8: *Residualer for variation 5.*

Bilag G

Vægtdata

Dette bilag indeholder vægtdataet omtalt i afsnit 4.7 på side 91.

Data fremgår af tabel G.1.
Første kolonne angiver vægten af stoflapperne *før* genvasken, men efter en times tørring i varmeskab.
Anden kolonne angiver vægten af stoflapperne *efter* genvasken, lufttørring og en times tørring i varmeskab.
Tredje kolonne angiver vægten af stoflapperne efter yderligere 1,5 time i varmeskab.
Fjerde kolonne angiver differensen mellem kolonne 2 og 1.
Femte kolonne angiver differensen mellem kolonne 3 og 2.

Stoflab 1A til 2D er vasket ved dosis 0,00%.
Stoflab 3A til 4D er vasket ved dosis 0,10%.
Stoflab 5A til 6D er vasket ved dosis 0,20%.
Stoflab 7A til 8D er vasket ved dosis 0,30%.
Stoflab 9A til 10D er vasket ved dosis 0,40%.

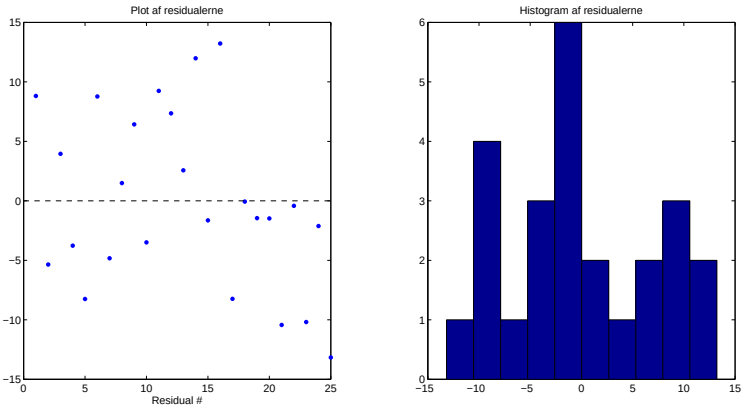
Tabel G.1: *Vægtdata. Alle målinger er i gram med en usikkerhed på ±1mg.*

Stoflap	Før	Efter	Efter yderligere 1,5 time	Differens 1	Differens 2
1A	3,475	3,349	3,337	-0,126	-0,012
1B	3,467	3,335	3,316	-0,132	-0,019
1C	3,575	3,459	3,466	-0,116	0,007
1D	3,664	3,516	3,515	-0,148	-0,001
2A	3,607	3,440	3,437	-0,167	-0,003
2B	3,650	3,475	3,446	-0,175	-0,029
2C	3,706	3,542	3,520	-0,164	-0,022
2D	3,648	3,467	3,445	-0,181	-0,022
3A	3,687	3,518	3,488	-0,169	-0,030
3B	3,629	3,459	3,461	-0,170	0,002
3C	3,663	3,468	3,461	-0,195	-0,007
3D	3,711	3,490	3,494	-0,221	0,004
4A	3,363	3,235	3,193	-0,128	-0,042
4B	3,493	3,383	3,378	-0,110	-0,005
4C	3,497	3,351	3,350	-0,146	-0,001
4D	3,580	3,410	3,361	-0,170	-0,049
5A	3,727	3,441	3,433	-0,286	-0,008
5B	3,667	3,428	3,435	-0,239	0,007
5C	3,501	3,367	3,290	-0,134	-0,077
5D	3,554	3,377	3,369	-0,177	-0,008
6A	3,546	3,417	3,404	-0,129	-0,013
6B	3,528	3,370	3,368	-0,158	-0,002
6C	3,452	3,308	3,295	-0,144	-0,013
6D	3,516	3,386	3,367	-0,130	-0,019
7A	3,770	3,625	3,636	-0,145	0,011
7B	3,411	3,301	3,290	-0,110	-0,011
7C	3,572	3,410	3,409	-0,162	-0,001
7D	3,690	3,501	3,471	-0,189	-0,030
8A	3,640	3,547	3,528	-0,093	-0,019
8B	3,694	3,460	3,455	-0,234	-0,005
8C	3,820	3,445	3,430	-0,375	-0,015
8D	3,663	3,459	3,437	-0,204	-0,022
9A	3,536	3,363	3,344	-0,173	-0,019
9B	3,662	3,402	3,392	-0,260	-0,010
9C	3,649	3,457	3,430	-0,192	-0,027
9D	3,627	3,446	3,398	-0,181	-0,048
10A	3,608	3,472	3,458	-0,136	-0,014
10B	3,615	3,478	3,466	-0,137	-0,012
10C	3,707	3,560	3,531	-0,147	-0,029
10D	3,428	3,289	3,310	-0,139	0,021

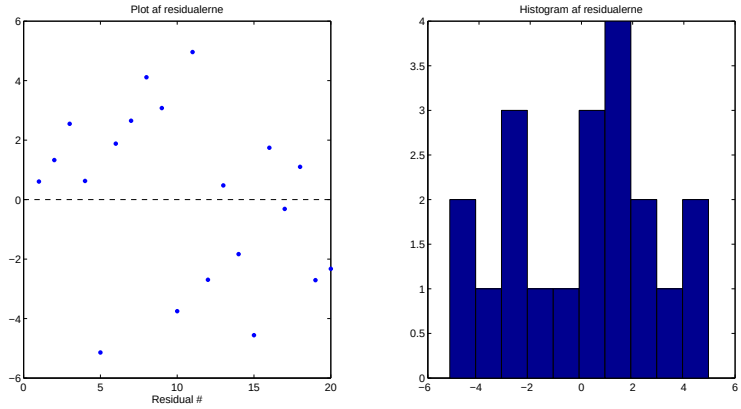
Bilag H

Residualer for glastallerkner

Dette bilag indeholder plots og histogrammer af residualerne for de to MLR modeller omtalt i afsnit 4.9 på side 95.



Figur H.1: *Plot og histogram af residualerne for MLR modellen baseret på alle 25 glastallerkner.*



Figur H.2: *Plot og histogram af residualerne for MLR modellen, hvor de 5 glastallerkner vasket ved vaskebetingelser #4 er udeladt.*

Litteratur

[Ans73] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[Arv13] Lars Arvastson. Bayesian data analysis. University lecture, 2013.

[BB09] Robert P. Burns and Richard Burns. *Business Research Methods and Statistics Using SPSS*. SAGE Publications Ltd, 2009.

[Car02] J. M. Carstensen. *Image Analysis, Vision and Computer Graphics*. Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2002.

[Cle13] Line Harder Clemmensen. Principal component analysis. University lecture, 2013.

[CS] Jens Michael Carstensen and Nette Schultz. Dermatological conditions assessed by spectral imaging and normalized canonical discriminant analysis. Paper intended for the International Conference on Image Analysis and Processing 2013.

[Dal12] Gerard E. Dallal. Why p=0.05?, December 2012.

[EC12] Bjarne Kjær Ersbøll and Knut Conradsen. *Multivariate Statistics - An Introduction*. DTU Informatics, Department of Informatics and Mathematical Modeling, 8. edition, 2012.

[Hag89] William W. Hage. Updating the inverse of a matrix. *Society for Industrial and Applied Mathematics*, 31(2):221–239, 1989.

[HPS13] Per Christian Hansen, Víctor Pereyra, and Godela Scherer. *Least Squares Data Fitting with Applications*. Johns Hopkins University Press, 2013.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2009.

[JFM11] Richard Johnson, John Freund, and Irwin Miller. *Probability and Statistics for Engineers*. Pearson International Education, 8. edition, 2011.

[Mij10] Rosa Mijer. Effcient approximate leave-one-out cross-validation for ridge and lasso. Master's thesis, Delft University of Technology, 2010.

[Nie99] A. A. Nielsen. *Orthogonal Transformations*. IMM, DTU, 1999.

[Rod] Germán Rodríguez. Princeton University. The Anscombe Datasets.

[Sch12] Mikkel Schmidt. Overfitting and performance evaluation. University lecture, 2012.

[She09] Simon J. Sheather. Multiple linear regression. In *A Modern Approach to Regression with R*, Springer Texts in Statistics, pages 125–149. Springer New York, 2009.

[SW91] David J. Saville and Graham R. Wood. *Statistical Methods: The Geometric Approach*. Springer Texts in Statistics, 1991.

[TOHW13] A. Trujillo-Ortiz and R. Hernandez-Walls. Hotelling T-squared test. MATLAB Central File Exchange, Retrieved May 25, 2013.

[TOHWBRCM] A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, and L. Cupul-Magana. Hzmvntest: Henze-zirkler's multivariate normality test. MATLAB Central File Exchange, Retrieved May 25, 2013.

[TSK06] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson International Education, 2006.