Forord

Dette eksamensprojekt er udført på Institut for Informatik og Matematisk Modellering, IMM, Danmarks Tekniske Universitet i perioden 5. august 2002 til 31. marts 2003.

Centre for Proteome Analysis, CPA, i Odense har leveret datamateriale til projektet, og en stor tak skal gå til Peter Mose Larsen (CPA) og Stephen J. Fey (CPA) for hjælp til en bedre forståelse af baggrunden for data samt besvarelse af spørgsmål.

Jeg vil derudover gerne takke min vejleder Bjarne Ersbøll for god vejledning og støtte gennem projektforløbet, for stort engagement og ikke mindst for altid at være behjælpelig med besvarelse af spørgsmål.

Lyngby, marts 2003

Mia Skettrup

Abstract

The aim of this thesis is to examine whether it is possible by multivariate data analysis to find single proteins or groups of proteins in brain tissue that distinguish between rats with epilepsy and healthy rats. The composition of proteins in brain tissue from rats is determined by 2d-electrophoresis gels. Samples from 19 rats, 12 with three different levels of epilepsy (under development, mild and severe) and a control group of 7 healthy rats, were analysed. Hence the data set consists of 19 observations and 1849 variables (different proteins).

Standard procedure today is to use univariate methods for statistical analysis of data from 2d-electrophoresis gels. As a reference for the multivariate analysis applied in this thesis, a non-parametric one-way analysis of variance and a t-test is performed. Multivariate statistical methods, contrary to the univariate methods, take into account correlations between the variables. The methods can be divided into two main groups, non-supervised and supervised. The non-supervised methods include principal component analysis, factor analysis, cluster analysis and canonical correlation analysis, while the supervised methods include discriminant analysis, logistic regression and classification trees.

Principal component analysis and canonical correlation analysis proved to be suitable for distinguishing the three ill groups of rats from each other and from the healthy group. With factor analysis and cluster analysis it is possible to divide the 1849 variables into smaller groups of which some seem to be interesting in relation to epilepsy.

Of the supervised methods only the discriminant analysis proved to be suitable for the data and within this, a couple of variables were found that appear to have an influence on whether the rat suffers from epilepsy or not and which level of epilepsy it could be.

Resumé

Dette projekts formål er at undersøge, om det med multivariat dataanalyse er muligt at finde enkelt-proteiner eller grupper af proteiner i hjernevæv, der adskiller epilepsiramte rotter fra en gruppe raske rotter. Proteinsammensætningen i hjernevæv fra rotter er bestemt ved hjælp af 2d-elektroforesegeler. Der blev analyseret prøver fra 19 rotter, dels fra 12 rotter med tre forskellige grader af epilepsi (under udvikling, mild og svær) og dels fra en kontrolgruppe af 7 raske rotter. Datasættet består således af 19 observationer og 1849 variable (forskellige proteiner).

Det er i dag standard at benytte univariate metoder til statistisk analyse af data fra 2d-elektroforesegeler, og som sammenligningsgrundlag for de, i dette projekt, benyttede multivariate analyser, er der udført en ikke-parametrisk ensidet variansanalyse og en t-test. Multivariate statistiske metoder tager, modsat de univariate metoder, højde for korrelationer mellem variablene. Metoderne kan deles op i to hovedgrupper, ikke-superviserede og superviserede. De ikke-superviserede metoder omfatter principal komponentanalyse, faktoranalyse, clusteranalyse og kanonisk korrelationsanalyse, mens de superviserede metoder omfatter diskriminantanalyse, logistisk regression og klassifikationstræer.

Principal komponentanalyse og kanonisk korrelationsanalyse viste sig rimelig velegnede til at adskille de tre syge grupper af rotter fra hinanden og fra de raske. Med faktoranalyse og clusteranalyse er det muligt at inddele de 1849 variable i mindre grupper, hvor nogle af disse ser ud til at være interessante i forbindelse med epilepsi.

Af de superviserede metoder var det kun diskriminantanalysen, der viste sig velegnet til data, og med denne blev fundet et par variable, som ser ud til at have en indflydelse på, om rotten har epilepsi eller ej, og hvilken grad af epilepsi der kunne være tale om.

Indholdsfortegnelse

1 Indledning

Ordet epilepsi stammer fra græsk og betyder angrebet eller anfald – i det gamle Grækenland troede man, at mennesket blev kastet til jorden af guderne ved krampeanfaldet. I dag ved man fortsat ikke præcist, hvad det er, der udløser et epileptisk anfald. I ca. 70 % af epilepsitilfældene kender man dog årsagerne. Af disse kan blandt andet nævnes medfødte misdannelser i hjernen, hjernesvulster, følger efter hjernebetændelse, en blodprop i hjernen eller en hjerneblødning og alkoholmisbrug. For de sidste 30 % af epilepsitilfældene gælder det, at epilepsien er nedarvet. Årsagen til denne type epilepsi kendes ikke, men det formodes, at den skyldes nogle ukendte kemiske forandringer i hjernen, som muligvis har noget at gøre med cellemembranerne og neurotransmitterne. Neurotransmitter er stoffer, der leder impulser fra en nervecelle til en anden.

Nogle af de behandlingsformer der findes mod epilepsi i dag, er forskellige former for medicin og operation. Operation er dog kun en mulighed for en meget lille procentdel af de epilepsiramte, det er kun for omkring 3-5 %, det er relevant. Det er altså hovedsageligt medicin, der bruges til behandling. Problemet med de nuværende former for medicin er, at der dels stadig er 25-30 % af patienterne, der er vanskelige at behandle, og som ikke bliver anfaldsfrie, og dels har ca. halvdelen af alle patienter bivirkninger af medicinen. En anden metode, der benyttes i dag, anvender samme teknik som en pacemaker og benyttes til patienter med svær epilepsi. Metoden er kendt som nervus vagus-stimulation. Ved denne metode gives der med bestemte intervaller korte serier af svage strømimpulser til den ene af de to store nerver, der går fra hjernestammen ned forbi strubehovedet og hjertet. Det er lykkedes at reducere anfaldshyppigheden hos patienterne ved hjælp af nervus vagus-stimulation, men meget få bliver dog helt anfaldsfrie.

Der er altså et behov for at finde enten andre behandlingsformer eller ny medicin mod epilepsi, men problemet med udviklingen af ny medicin er, at der fortsat er en meget mangelfuld forståelse af de mekanismer i hjernen, der ligger bag anfaldene.

Data til dette projekt stammer fra 'Centre for Proteome Analysis', CPA, i Odense, hvor prøver af hjernevævet for rotter med forskellige grader af epilepsi samt for raske rotter er blevet undersøgt ved hjælp af 2d-elektroforese, en teknik der kan bestemme prøvernes

proteinsammensætning. Man håber at kunne finde nogle proteiner, der adskiller sig fra en syg til en rask gruppe, det vil sige, finde nogle proteiner der kan have en indflydelse på epilepsiudvikling og dernæst undersøge disse proteiners funktion. Ved at finde ud af hvilke mekanismer i hjernen proteinerne er involveret i, kan man måske få et bedre indblik i, hvordan epilepsi udvikler sig.

Formålet med dette projekt er at undersøge, om det med multivariat dataanalyse er muligt at finde nogle enkelte proteiner eller eventuelt nogle grupper af proteiner, der adskiller sig fra den raske gruppe til de forskellige grupper af epilepsiramte, sagt på en anden måde om det er muligt at finde et eller flere proteiner, der har betydning for eller ligefrem er årsag til, om en rotte har epilepsi eller ej og eventuelt, hvor svær en grad af epilepsi der er tale om.

Hvis der findes sådanne proteiner, kan det måske på længere sigt blive muligt blot ved hjælp af en blodprøve, hvis proteinsammensætningen i blodet altså kan antages at være den samme som proteinsammensætningen i en prøve af hjernevævet, at afgøre, om en person har anlæg for at udvikle epilepsi, og det kan muligvis lade sig gøre at udvikle en ny behandlingsform eller ny medicin, der kan bremse eller endda behandle epilepsi, hvis man ved, hvilke proteiner (hvis der da er nogen) der spiller en rolle ved sygdommen.

Dette projekt bygger i nogen grad videre på Line Conradsens eksamensprojekt [8], hvor der også blev arbejdet med analyse af 2d-elektroforesegeler. I Line Conradsens eksamensprojekt var det kræft i livmoderen, der blev undersøgt, hvor undersøgelsen her handler om udviklingen af epilepsi. Line Conradsen benyttede i sit projekt principal komponentanalyse, clusteranalyse på observationerne og diskriminantanalyse til undersøgelsen, og det viste sig, at principal komponentanalyse og diskriminantanalyse gav nogle rimelige resultater, mens den benyttede form for clusteranalyse ikke var særlig velegnet til denne type data. De metoder, der er benyttet til dette projekt, er derfor dels de to, Line Conradsen fandt velegnede, dels clusteranalyse på variablene i stedet for på observationerne og dels en række andre multivariate metoder.

1.1 Rapportens opbygning

Efter indledningen i kapitel 1 følger i kapitel 2 en beskrivelse af de udleverede data samt en gennemgang af den indledende databehandling, her beskrives desuden 2d-

elektroforeseteknikken. De benyttede metoder kan opdeles i tre grupper, univariate, ikke-superviserede og superviserede, og disse vil blive gennemgået i kapitel 3, hvorefter gennemgang og diskussion af de vigtigste resultater af analyserne følger i kapitel 4. Her bliver først resultaterne af de univariate metoder præsenteret, dernæst følger resultaterne af de ikke-superviserede metoder og resultaterne af de superviserede metoder. Til sidst bliver resultaterne af de ikke-superviserede metoder og resultaterne af de superviserede metoder sammenlignet.

2 Data

Data kommer som nævnt fra 'Centre for Proteome Analysis', CPA, i Odense. De stammer fra en undersøgelse af udvikling af epilepsi i rotter. Rotterne blev i undersøgelsen stimuleret med elektrisk chok i en specifik del af hjernen og efter omkring en måned, udviklede de epilepsilignende anfald. Ifølge CPA er dette den bedste model for epilepsi hos mennesker, der eksisterer i øjeblikket. Hvis der derfor kunne findes nogle relevante proteiner for de undersøgte rotter og disse proteiners funktioner, og de mekanismer i hjernen de styrer, blev klarlagt, kunne man forestille sig, at det vil være de samme mekanismer, der ligger bag 'naturligt' udviklet epilepsi hos mennesker. Derved kunne man forhåbentlig komme en effektiv behandling af epilepsi et skridt nærmere.

De undersøgte rotter blev efter stimulationen med elektrisk chok overvåget og anfaldene målt med et video-EEG monitor system. Kortslutningen i hjernen og varigheden af anfaldene blev målt, og antallet af anfald samt hvor mange dage, der gik, til det første anfald blev noteret. Ud fra disse målinger blev rotterne inddelt i grupper, efter hvor svær grad af epilepsi de havde udvist.

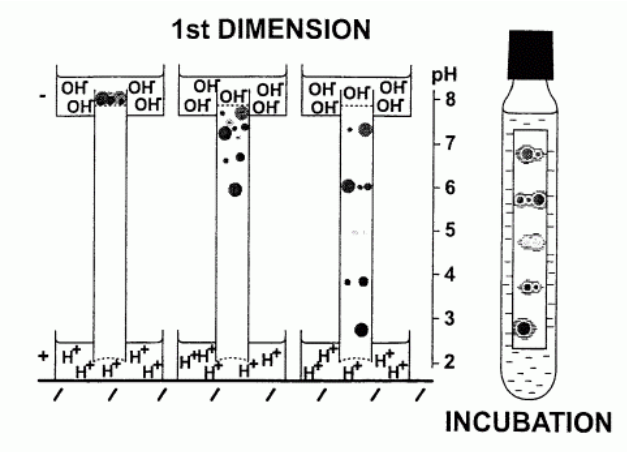
Ved undersøgelsen blev herefter udtaget prøver fra hippocampus i hjernen på de 19 involverede rotter og for hver af disse prøver blev lavet 2d-elektroforesegeler.

2.1 2d-elektroforese

For at adskille proteinerne i prøverne blev todimensionale elektroforesegeler benyttet. Første dimension adskiller proteinerne efter deres isoelektriske punkt (pI), mens anden dimension adskiller dem efter deres molekylvægt. Der er to forskellige slags geler, en som dækker pH-området fra 4-7, og en som dækker pH-området fra 6-9.

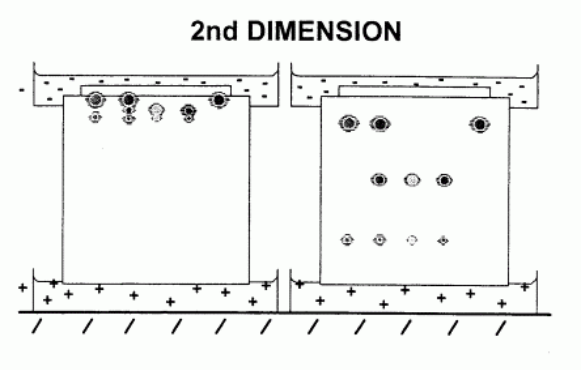
For at gøre det muligt at detektere proteinerne efter adskillelsen, blev de først 'mærket'. Dette blev gjort ved at nedfryse prøverne i flydende kvælstof, hvilket bevirker, at prøverne forbliver, som de var i organismen, og dernæst behandle proteinerne ved hjælp af et fluorescerende stof [9].

Efter 'mærkning' blev proteinerne tilført gelen, i hvilken en pH-gradient er indstøbt, og herefter blev en spænding sat over gelen. De ladede proteiner begyndte herved at bevæge sig ind i gelen, og da proteinerne har forskellig ladning, blev de skilt fra hinanden efter deres isoelektriske punkt. Et proteins pI angiver den pH-værdi, hvor proteinets samlede ladning er neutral i forhold til omgivelsernes pH-værdi. Proteinerne bevægede sig, indtil de nåede den pH-værdi, der svarer til deres isoelektriske punkt, det vil sige til den pH-værdi, hvor proteinet er neutralt. Der er her ingen elektrisk kraft, der kan påvirke proteinet i hverken den ene eller den anden retning. Figur 1 nedenfor illustrerer princippet ved første dimension:



Figur 1: Princippet ved første dimension af 2d-elektroforese [15]

Efter ligevægt var nået, det vil sige, når proteinerne havde bevæget sig til deres isoelektriske punkt (pH = pI), blev første dimensions gelen ækvilibreret, inden den blev placeret på den øverste kant af en polyakrylamid gel. I denne gel er der ingen pH-gradient, men derimod en størrelsesgradient. Porerne i gelen bliver gradvis mindre og mindre, så gelen fungerer ligesom et filter, som proteinerne kan bevæge sig gennem i forskellig hastighed alt efter deres molekylvægt. Proteiner med en lille molekylvægt vil bevæge sig hurtigere gennem gelen end store proteiner. En spænding blev sat over gelen, og proteinerne begyndte at bevæge sig ind i polyakrylamid gelen. Ved første dimension stoppede proteinernes bevægelse, når ligevægt var nået, men ved anden dimension er der ikke et ligevægtspunkt. Proteinerne bliver ved at bevæge sig, så længe det elektriske felt opretholdes. Derfor blev processen stoppet, når de mindste proteiner nåede bunden af gelen, og inden de bevægede sig helt ud af gelen. På figur 2 på næste side er princippet ved anden dimension vist:



Figur 2: Anden dimension af 2d-elektroforese [15]

Efter anden dimension var fuldført, blev gelen tørret, og proteinerne kunne herefter detekteres med et kamera, der er følsomt overfor fluorescensen.

2.2 Sammenligning af 2d-elektroforesegeler

Efter alle geler var blevet fremstillet, og alle proteiner var blevet detekteret, hvilket kan ses som pletter på gelen, blev pletterne på alle geler sammenlignet og matchet med en mastergel med hjælp fra billedbehandlingsprogrammet BioImage. Mastergelen blev valgt som den af de 19 geler, hvor pletterne var bedst afgrænsede. Pletterne på en 2d-elektroforesegel er ikke alle tydeligt adskilt fra hinanden, der kan være en flydende overgang fra en plet til en anden, og det kan derfor nogle gange være vanskeligt at skelne de enkelte pletter fra hinanden.

For hver gel blev intensiteten af hver enkelt plet ligeledes bestemt. I den forbindelse må det lige nævnes, at det også kan være svært at bestemme intensiteten af hver af pletterne, hvis der er en glidende overgang mellem pletterne. Intensiteten blev bestemt sådan, at alle pletter for alle rotter blev matchet. De pletter, der fandtes på mastergelen, blev altså tilsvarende fundet på alle de andre geler. På CPA foregår det sådan, at hvis en plet på mastergelen ikke kan findes på en af de andre geler, antages det, at denne plet alligevel eksisterer, men at intensiteten er så lille, at den ikke er synlig, og den manglende plet gives derfor værdien 0.

CPA bruger en anden fremgangsmåde end den, der blev benyttet til Line Conradsens data. Her blev en ikke-fundet plet betragtet som en manglende værdi i datasættet.

Problemet med manglende værdier i data er, at mange analyser ikke kan udføres, hvis datasættet ikke er fuldstændigt. Der er mange måder at gribe problemet med manglende værdier an på, hvor for eksempel kan nævnes:

- Udelad observationer med manglende værdier
- Erstat manglende værdier med middelværdien for de ikke-manglende værdier
- Erstat manglende værdier med medianen for de ikke-manglende værdier

CPA har dog valgt, at hvis en plet ikke kan findes, får den værdien 0, og mit datasæt er derfor fuldstændigt. Jeg vil af denne grund ikke komme nærmere ind på problemerne omkring manglende værdier i data, men alene gøre opmærksom på at CPA's fremgangsmåde er en anden.

Da der var store problemer med at detektere proteinerne på de basiske geler, er kun de 19 geler med pH-værdier fra 4-7 blevet undersøgt i dette projekt.

12 af de 19 undersøgte rotter er blevet stimuleret med elektrisk chok, mens de sidste syv udgør en kontrolgruppe. De 12 stimulerede rotter er af CPA blevet inddelt i to grupper, en med rotter der er ved at udvikle epilepsi, og en med rotter der har epilepsi. Epilepsi under udvikling er defineret ved, at ingen synlige anfald har fundet sted, men et udslag på EEG-monitoren er alligevel blevet målt. Gruppen af rotter med epilepsi er yderligere inddelt i to, rotter med mild grad af epilepsi og rotter med svær grad af epilepsi, ud fra de tidligere nævnte målinger af blandt andet varigheden af anfaldene og antallet af anfald. Blandt de 12 rotter er der desuden en enkelt rotte, som er blevet stimuleret, men som hverken har epilepsi eller epilepsi under udvikling. En mulig forklaring på den manglende reaktion kan være, at elektroderne ikke har siddet, som de skulle. En anden forklaring kan, ifølge Peter Mose Larsen, CPA, være, at rotten er resistent overfor behandlingen. Da denne rotte ikke har reageret på det elektriske chok (er en outlier), har jeg udeladt den i mine analyser.

I nedenstående tabel ses en oversigt over, hvilke undergrupper rotterne tilhører.

Ikke reageret	Under udvikling	Mild grad	Svær grad	Kontrolgruppe
R11	R3, R4, R10, R12	R1, R7, R8, R9	R2, R5, R6	R13 - R19

Tabel 1: Inddelingen af rotter i undergrupper.

Det udleverede datasæt indeholder 39 søjler og 1849 rækker, hvor hver række angiver en plet på gelen. Den første søjle angiver proteinpletternes numre, og de resterende 38 angiver intensiteten af proteinpletterne samt denne i procent af den totale intensitet for hver af de 19 rotter.

Til de videre analyser blev kun den procentvise intensitet benyttet. Det vil sige, som udgangspunkt for analyserne blev dannet et nyt datasæt, hvor kun den procentvise intensitet for hver rotte optræder. Det nye datasæt blev desuden transponeret, så proteinpletternes numre blev variable. Proteinpletternes oprindelige numre blev ligeledes erstattet af fortløbende numre, spot1-spot1849. Herefter haves et datasæt med kun 19 observationer (en for hver rotte) og 1850 variable, en for hver proteinplet samt en der angiver numre for rotterne, og fra dette datasæt blev fjernet den rotte, som er blevet stimuleret, men som ikke har reageret. Til sidst blev tilføjet en variabel, epilep, der har værdierne 0, 1, 2, 3, alt efter hvilken grad af epilepsi den pågældende rotte har. For kontrolgruppen har epilep værdien 0, og for rotter med svær grad af epilepsi er værdien 3.

2.3 Curse of dimensionality

Som tilfældet er ved denne undersøgelse, er man ved mange undersøgelser i dag i den situation, at det er muligt at indsamle oplysninger om utrolig mange forskellige parametre (i dette tilfælde proteiner), der på den ene eller den anden måde beskriver det, der ønskes undersøgt. Fordelen ved det er, at man kan få belyst mange aspekter af det ønskede, og man kan have et håb om, at der er taget højde for det meste, men det giver samtidig et problem at have mange variable. Blandt flere hundreder eller tusinder af variable kan det for eksempel være vanskeligt at finde de variable, der har størst betydning for undersøgelsen.

De vanskeligheder, der er ved at foretage analyser på data i multidimensionelle rum, er kendt som ’curse of dimensionality’. Nogle af problemerne er, at i situationer med mange variable vil observationerne dels befinde sig langt fra hinanden, så det ikke vil være muligt at foretage undersøgelser lokalt, og dels vil observationerne alle befinde sig i et grænseområde af det givne rum. Det er ikke muligt at løse disse problemer blot ved at forøge mængden af observationer, da antallet af observationer, der er nødvendige for at

kunne foretage en analyse med en given nøjagtighed, vokser eksponentielt med antallet af variable [11].

I dette projekt er der 1849 variable til at beskrive data og kun oplysninger fra 18 rotter. Det vil sige, der arbejdes i 1849 dimensioner, og de 18 observationer er fordelt i det 1849-dimensionelle rum.

For at give et billede af hvor spredt observationerne ligger, ses på en p-dimensional enhedshyperkubus, i hvilken observationerne er uniformt fordelt. Sidelængden af den hyperkubus, der er nødvendig for at dække r % af enhedskubusen, er nu, givet et target punkt, et udtryk for, hvor stort et område der skal medtages for at dække r % af de omkringliggende observationer. For det givne target punkt kan størrelsen af dette område i procent af det totale område beregnes som [13]:

$$e_p(r) = r^{1/p}$$

For at dække blot 1 % når p = 1849, skal der altså medtages hele 99.75 % af det område, hver enkelt variabel spænder over, og så er der jo ikke længere tale om en lokal undersøgelse.

Ses nu på N uniformt fordelte observationer i en p-dimensional enhedskugle kan afstanden fra centrum af kuglen til den tætteste observation bestemmes som [12]:

$$d(p,N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

Når p er 1849, og N er 18, bliver afstanden fra centrum til den tætteste observation 0.9982, hvilket betyder, at den tætteste observation altså befinder sig meget tæt ved kanten af enhedskuglen. Grunden til, at det ikke er hensigtsmæssigt, at observationerne befinder sig i grænseområdet, er, at prædiktion er meget vanskeligere nær kanterne af det givne rum.

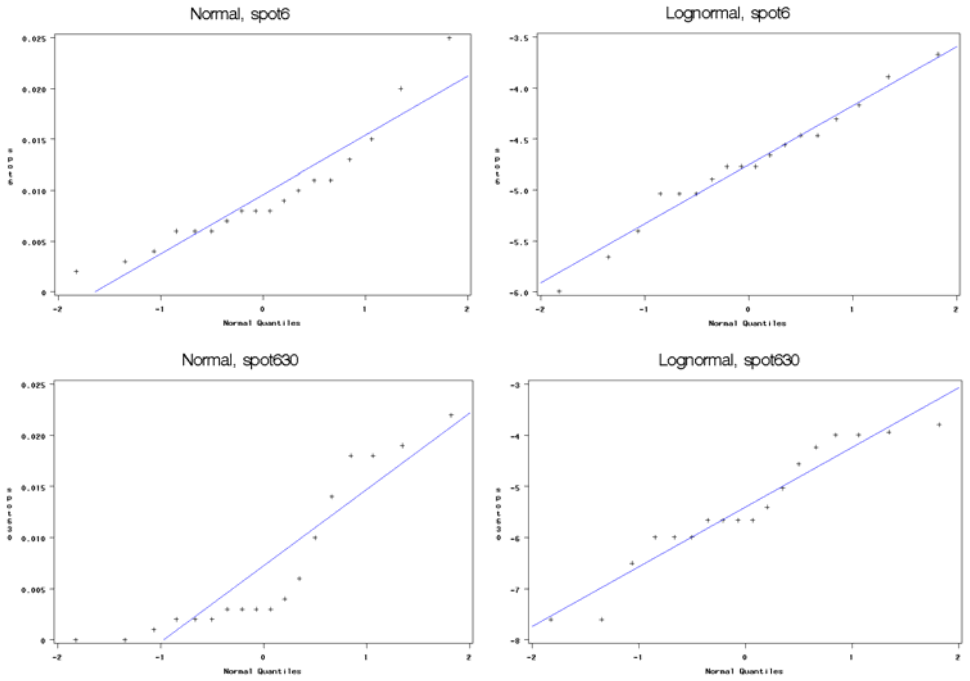
Ved begge ovenstående beregninger forudsættes det, at data er uniformt fordelt, hvilket nok ikke er tilfældet for mine data, så de fundne værdier er blot for at give et indtryk af, hvor spredt observationerne ligger, når der arbejdes med multidimensionelle data.

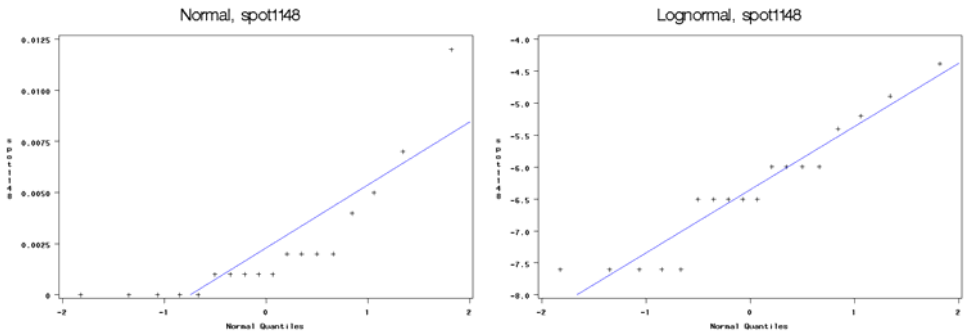
2.4 Transformation af data

En forudsætning for at udføre mange af de, i dette projekt, anvendte analyser er, at data er normalfordelt. Dette gælder dog hverken for ikke-parametrisk ensidet variansanalyse eller for ikke-parametriske klassifikationstræer og heller ikke logistisk regression forudsætter normalfordelte variable. Alle de øvrige analyser har dog denne forudsætning. For at undersøge om data eller eventuelt en logtransformation af data er normalfordelt, blev lavet fraktildiagrammer for det originale datasæt og for det logtransformerede.

Da der er adskillige observationer i det oprindelige datasæt, der har værdien 0, kunne data ikke logtransformeres direkte. Først måtte der vælges et tal, der skulle lægges til samtlige tal i datasættet. Som tommelfingerregel bliver dette tal valgt som halvdelen af den mindste talværdi i datasættet. I dette tilfælde er den mindste talværdi 0.001, og der blev derfor lagt 0.0005 til alle tal. Herefter blev data logtransformeret.

Da der er 1849 variable, skulle der i princippet laves 1849 fraktildiagrammer, men jeg har valgt blot at udvælge et par af disse for at illustrere pointen. I figur 3 nedenfor ses fraktildiagrammerne for tre variable, dels for de originale data og dels for de logtransformerede:





Figur 3: Fraktildiagrammer for tre forskellige pletter på gelen, dels for originaldata (til venstre) og dels for logtransformerede data (til højre).

Som det ses af ovenstående fraktildiagrammer, følger de originale data ikke en normalfordeling, hvilket de logtransformerede data derimod tilnærmet kan siges at gøre. Da fraktildiagrammerne antyder, at de logtransformerede data er tilnærmet normalfordelte, er det resultaterne af analyser af disse, jeg vil koncentrere mig mest om i denne rapport. Alle analyser er dog foretaget både på det logtransformerede datasæt samt på det originale.

Endvidere er alle analyser foretaget på rangen af data, hvorved det opnås, at analyserne bliver ikke-parametriske. For hver variabel i datasættet nummereres alle observationer efter voksende størrelse, og en observations rang defineres, som det nummer det har i opstillingen. Den observation, der har den laveste talværdi, får altså rangen et, observationen med den næstlaveste talværdi får rangen to osv. Hvis to eller flere observationer har samme talværdi, for eksempel hvis der er fire observationer, der alle har talværdien 1, og denne værdi er den laveste talværdi i datasættet, får disse fire observationer alle den samme rang, nemlig gennemsnittet af de rangværdier der skal tilordnes observationerne. Det vil sige, de får rangen $(1+2+3+4)/4 = 2.5$.

Ved at se på rangen af data beholdes altså for hver variabel kun information om, i hvilken rækkefølge observationerne kan ordnes størrelsesmæssigt. Al information om, hvor store værdierne for de forskellige observationer er i forhold til hinanden, forsvinder. En ulempe ved ikke-parametriske analyser er altså, at der ikke bliver taget højde for en stor del af informationen i data.

En væsentlig fordel ved ikke-parametriske analyser er derimod, at de ikke er baseret på en forudsætning om, at fordelingen af data kendes, modsat de parametriske analyser hvor

en af de vigtigste forudsætninger netop er en kendt fordeling. En ikke-parametrisk metode vil ofte udnytte data mere effektivt end en parametrisk, hvis forudsætningerne for denne ikke er helt opfyldt [5]. Man kunne derfor forestille sig, når der var tvivl om fordelingen af data, at ikke-parametriske analyser ville være velegnede. Dette har dog ikke vist sig at være tilfældet for data til dette projekt. Da de logtransformerede data, som nævnt ovenfor, med god tilnærmelse kan siges at følge en normalfordeling, er det da også rimeligt, at en parametrisk analyse vil være bedre end en ikke-parametrisk. Det er derfor resultaterne af analyserne på de logtransformerede data, hovedvægten er lagt på i denne rapport. Resultaterne opnået ved analyse på de originale data og på rangen af data vil dog også ganske kort blive omtalt.

3 Metoder

De metoder, der blev benyttet til dette projekt, kan inddeles i tre kategorier: Univariate, ikke-superviserede og superviserede. I dette kapitel gennemgås hver af de anvendte metoder indenfor de tre grupper.

3.1 Univariate

Som navnet antyder, er univariate metoder analyser af endimensionale variable. En univariat analyse kan altså afgøre for alle proteiner eller pletter på gelen, hvorvidt hver enkelt plet på gelen er forskellig for de givne grupper, men den kan ikke fortælle noget om eventuelle korrelationer eller sammenhænge mellem de enkelte proteiner. Det har dog været standard hidtil at benytte univariate metoder til analyse af 2d-elektroforese data, og i dette afsnit beskrives kort to sådanne metoder.

3.1.1 Ikke-parametrisk ensidet variansanalyse

Da det ikke altid er muligt at bestemme en underliggende fordeling for data, kan det, som nævnt, være en fordel i sådanne tilfælde at benytte en ikke-parametrisk test, en fordelingsfri test. En test er fordelingsfri, når den bygger på en stikprøvefunktion med en fordeling, som under nulhypotesen er uafhængig af de oprindelige observationers fordeling [6]. Det forudsættes for fordelingsfrie tests, at observationerne er indbyrdes uafhængige og har samme fordeling. Til analyse af 2d-elektroforese data er det almindeligt at benytte en fordelingsfri test, nemlig en ikke-parametrisk ensidet variansanalyse.

Ved en ensidet variansanalyse undersøges det, om middelværdierne for de valgte grupper er ens, der testes for fuldstændig homogenitet. Hvis det ikke er muligt at bestemme den underliggende fordeling af data, er en almindelig ensidet variansanalyse ikke hensigtsmæssig, hvorimod en ikke-parametrisk ensidet variansanalyse er mere fornuftig.

Selvom det, som nævnt, blev fundet, at de logtransformerede data med god tilnærmelse kan siges at følge en normalfordeling, så det altså kan antages, at fordelingen af data

kendes, er der alligevel, på de originale data, udført en ikke-parametrisk ensidet variansanalyse, da dette er en standardmetode at benytte til denne type data.

Wilcoxons rang test er en fordelingsfri test, og det er denne, der er benyttet til analyse af de udleverede data. Ved Wilcoxons rang test betragtes ikke de originale data, men derimod rangen af dem og her undersøges ikke, om middelværdierne i de to grupper, X og Y, kan antages at være ens, men derimod om summen af rangene i grupperne er den samme. Hvis der er flere end to grupper, der ønskes sammenlignet, benyttes Kruskal-Wallis test [5].

I denne undersøgelse testes for hver proteinplet på gelen, dels om summen af rangene for gruppen med alle de stimulerede rotter kan antages at være den samme som summen af rangene for kontrolgruppen, dels om summen af rangene i de fire grupper (kontrolgruppen, epilepsi under udvikling, mild grad af epilepsi og svær grad af epilepsi) kan antages at være ens.

Proceduren npar1way i SAS laver en ikke-parametrisk ensidet variansanalyse for de to tilfælde ved at udføre henholdsvis Wilcoxons rang test og Kruskal-Wallis test.

3.1.1.1 Problem med ensidet variansanalyse

Problemet med at benytte en ensidet variansanalyse og teste hver plet for sig er, at der foretages lige så mange tests, som der er pletter på gelen, hvilket i dette tilfælde vil sige 1849 tests.

Ved ensidet variansanalyse testes hypotesen, H_0: 'Grupper er ens'. Signifikansniveauet for testet er α, hvilket betyder, at det accepteres i α % af tilfældene at forkaste en sand hypotese. Når der kun udføres et enkelt test, er der altså α % sandsynlighed for at forkaste en sand hypotese.

Når der udføres 1849 tests, vil det betyde, at der i α % af disse 1849 vil blive forkastet en sand hypotese. Det vil sige, at en del af de tests, der udviser signifikans i virkeligheden er fejl af type I (en sand hypotese forkastes). Hvis signifikansniveauet for eksempel er 5 %, vil hypotesen om ens grupper i 92 af de 1849 tests blive forkastet, selvom den er sand.

Ensidet variansanalyse vil derfor give et resultat, der skal tages med forbehold. Ovenstående gælder selvfølgelig under forudsætning af, at hypotesen virkelig er sand.

3.1.2 T-test

Som nævnt tidligere er data blevet logtransformeret og under antagelse af, at de tilnærmet følger en normalfordeling, kan der på dette datasæt foretages en uparret t-test. Ved en uparret t-test sammenlignes middelværdierne i de to uafhængige grupper, stimulerede rotter og kontrolgruppen. Først må det dog undersøges, hvorvidt variansen i de to grupper er ens. Herefter kan middelværdierne sammenlignes.

En t-test samt forudgående sammenligning af varianser kan udføres af proceduren ttest i SAS.

Der skal foretages 1849 t-tests, og dette giver samme problem som beskrevet for ikke-parametrisk ensidet variansanalyse – i α % af tilfældene vil en sand hypotese blive forkastet, og resultatet må altså også her tages med et vist forbehold.

3.2 Ikke-superviserede metoder

For ikke-superviserede metoder benyttes et givet sæt variable til at gruppere data bedst muligt. Det kan enten være en opdeling af observationerne, der ønskes fundet, eller det kan være variablene, der ønskes inddelt i grupper ud fra nogle fælles egenskaber. Problemet med ikke-superviserede metoder er, at det kan være svært at verificere resultaterne, da det ikke er muligt at bestemme for eksempel en fejlrate, der kan give et billede af metodens effektivitet. For dette projekt er det dog for enkelte af metoderne muligt at sammenligne den fundne gruppering af observationerne med den opdeling, som er givet af CPA.

3.2.1 Principal komponentanalyse

Principal komponentanalyse er en metode til at danne nye variable, **Y**, som er linearkombinationer af de oprindelige variable, **X**.

Dispersionsmatricen for de oprindelige variable bestemmes, og ud fra denne findes egenværdier, λ_i, og tilhørende egenvektorer, $\mathbf{p}_i$. Definitionen på den i'te principale akse er, som angivet i [7]:

> Ved den i'te principale akse for **X** forstås retningen hørende til egenvektoren $\mathbf{p}_i$ svarende til den i'te største egenværdi.

Definitionen på den i'te principale komponent fås herefter som:

> Ved den i'te principale komponent af **X** forstås **X**'s projektion $Y_i = \mathbf{p}_i' \mathbf{X}$ på den i'te principale akse.

Vektoren af principale komponenter fås altså som $\mathbf{Y} = \mathbf{P}' \mathbf{X}$, hvor $\mathbf{P} = (\mathbf{p}_1 \ldots \mathbf{p}_k)$.

Der gælder for de principale komponenter, at de er ukorrelerede, og at variansen på den i'te komponent er λ_i, det vil sige den i'te største egenværdi. Desuden er den første principale komponent den linearkombination af de oprindelige variable, der har den største varians, og den m'te principale komponent er den linearkombination af de oprindelige variable, som er ukorreleret med de m-1 første principale komponenter og har størst varians [7].

Da datasættet indeholder 1849 variable, vil en principal komponentanalyse medføre, at der kan dannes 1849 principale komponenter. For at undgå dette udføres den principale komponentanalyse ikke på det oprindelige datasæt, men derimod på det transponerede. Sammenhængen mellem en analyse af det oprindelige datasæt og en analyse af det transponerede forklares ved hjælp af Eckart-Young's sætning.

3.2.1.1 Eckart-Young's sætning

For en vilkårlig n×p matrix, **x**, med rangen r, eksisterer der ortogonale matricer **U** (p×r) og **V** (n×r) og positive tal $\gamma_1, \ldots, \gamma_r$, kaldet singulære værdier, således at

$$\mathbf{x} = \mathbf{V}\,\boldsymbol{\Gamma}\,\mathbf{U}' = \begin{bmatrix}\mathbf{v}_1 \Lambda\ \mathbf{v}_r\end{bmatrix}\begin{bmatrix}\gamma_1 & \Lambda & 0\\ \mathrm{M} & \mathrm{O} & \mathrm{M}\\ 0 & \Lambda & \gamma_r\end{bmatrix}\begin{bmatrix}\mathbf{u}_1'\\ \mathrm{M}\\ \mathbf{u}_r'\end{bmatrix} = \gamma_1\mathbf{v}_1\mathbf{u}_1' + \Lambda\ + \gamma_r\mathbf{v}_r\mathbf{u}_r'$$

hvor $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_r)$, $\mathbf{v_1}, \ldots, \mathbf{v_r}$ er søjlerne i **V** og $\mathbf{u_1}, \ldots, \mathbf{u_r}$ søjlerne i **U**. Denne sætning kaldes Eckart-Young's sætning. Der ønskes fundet en sammenhæng mellem **x**'s singulære værdier og egenværdiproblemerne for de to symmetriske matricer **xx'** (n×n) og **x'x** (p×p).

Det skal først nævnes, at der for en vilkårlig reel matrix **x** gælder, at **x'x** og **xx'** har ikke-negative egenværdier, og at rangen for de tre matricer er den samme [7].

Da rangen er r, og antallet af egenværdier forskellige fra 0 er lig rangen af matricen, har de to matricer **x'x** (p×p) og **xx'** (n×n) r positive egenværdier og henholdsvis (p-r) og (n-r) egenværdier lig 0.

For **xx'** gælder, da **U** er ortogonal, at

$$\mathbf{x}\,\mathbf{x}' = (\mathbf{V}\,\boldsymbol{\Gamma}\,\mathbf{U}')(\mathbf{U}\,\boldsymbol{\Gamma}'\,\mathbf{V}') = \mathbf{V}\,\boldsymbol{\Gamma}^2\,\mathbf{V}'$$

Det vil sige, egenværdierne ses at være $\gamma_1^2, \ldots, \gamma_r^2$, og de tilhørende egenvektorer er $\mathbf{v_1}, \ldots, \mathbf{v_r}$. For **x'x** findes egenværdierne ligeledes til $\gamma_1^2, \ldots, \gamma_r^2$, mens de tilhørende egenvektorer her er $\mathbf{u_1}, \ldots, \mathbf{u_r}$.

De ikke-negative egenværdier for de to matricer er altså ens, og sammenhængen mellem de tilhørende egenvektorer er ifølge Eckart-Young's sætning

$$\mathbf{V} = \mathbf{x}\,\mathbf{U}\,\boldsymbol{\Gamma}^{-1} \quad \text{og} \quad \mathbf{U} = \mathbf{x}'\,\mathbf{V}\,\boldsymbol{\Gamma}^{-1}$$

Det er altså muligt ud fra en analyse af det transponerede datasæt at opnå de ønskede resultater for det oprindelige datasæt. Eftersom det givne datasæt indeholder 1849 variable og kun 18 observationer, vil dette altså være en mere rimelig fremgangsmåde.

Et program, der kører en principal komponentanalyse i SAS på det transponerede datasæt, er udarbejdet i forbindelse med Line Conradsens eksamensprojekt [8] og venligst stillet til rådighed.

3.2.2 Kanonisk korrelationsanalyse

Når der haves to grupper af variable, der beskriver data på hver sin måde, kan kanonisk korrelationsanalyse benyttes til at finde en sammenhæng mellem disse. Dette gøres ved at danne linearkombinationer af hver af de to sæt af variable og dernæst finde det par, som har størst korrelation. Disse to variable kaldes det første par kanoniske variable, og korrelationen mellem dem kaldes den første kanoniske korrelation.

Benævnes det ene sæt variable $\mathbf{X}_1$ og det andet $\mathbf{X}_2$, findes det r'te par kanoniske variable, som det par af linearkombinationer $U_r = \mathbf{a}_r^{'}\mathbf{X}_1$ og $V_r = \mathbf{b}_r^{'}\mathbf{X}_2$ (hvor både U_r og V_r har variansen 1) der begge er ukorrelerede med de foregående r-1 par af kanoniske variable, og for hvilke det gælder, at korrelationen mellem dem er maksimal [2]. Hver kanonisk variabel er altså ukorreleret med alle kanoniske variable i de to sæt variable på nær den tilsvarende kanoniske variabel i det modsatte sæt.

Hvis de to sæt variable samles i den stokastiske variabel $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ og

dispersionsmatricen for denne variabel har udseendet $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$ kan den r'te kanoniske korrelation bestemmes som den r'te største rod, λ_r, af

$$\det\begin{pmatrix} -\lambda\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & -\lambda\mathbf{\Sigma}_{22} \end{pmatrix} = 0,$$

og koefficienterne **a** og **b** i de kanoniske variable opfylder følgende:

$$\begin{pmatrix} -\lambda_r \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda_r \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{pmatrix} = 0$$

$$\mathbf{a}_r' \boldsymbol{\Sigma}_{11} \mathbf{a}_r = 1$$

$$\mathbf{b}_r' \boldsymbol{\Sigma}_{22} \mathbf{b}_r = 1$$

For at få et indtryk af hvor god en opdeling af data den kanoniske korrelationsanalyse giver, kan hver af de kanoniske variable i den ene gruppe plottes mod den tilsvarende kanoniske variabel i den anden gruppe.

Den kanoniske korrelationsanalyse udføres i SAS med proceduren cancorr.

3.2.3 Faktoranalyse

Faktor analyse benyttes til at bestemme en underliggende faktorstruktur for data, det vil sige, det undersøges, hvordan variablene afhænger af nogle underliggende faktorer. For variablene **X** kan dette skrives:

$$\mathbf{X} = \mathbf{A}\,\mathbf{F} + \mathbf{G}$$

hvor **F** er vektoren af fælles faktorer også kaldet faktorværdierne, **A** indeholder faktorvægtene, der angiver, med hvilken vægt de enkelte faktorer beskriver de forskellige variable og **G** er vektoren af unikke faktorer, som er faktorer, der er specifikke for de enkelte variable. Både **X**, **F** og **G** antages at være stokastiske, og det forudsættes, at **F** og **G** er ukorrelerede, og at deres dispersionsmatricer er henholdsvis en enhedsmatrix og en diagonalmatrix [7]:

$$\mathrm{D}(\mathbf{F}) = \begin{pmatrix} 1 & \Lambda & 0 \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & \Lambda & 1 \end{pmatrix} = \mathbf{I} = \mathbf{I}_{\mathbf{m}}, \qquad \mathrm{D}(\mathbf{G}) = \begin{pmatrix} \delta_1 & \Lambda & 0 \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & \Lambda & \delta_k \end{pmatrix} = \boldsymbol{\Delta}$$

Ligeledes forudsættes det, at observationerne er standardiserede, så V(X) = 1 for alle i, hvilket betyder, at dispersionsmatricen for **X** og korrelationsmatricen for **X** er ens:

$$\mathrm{D}(\mathbf{X}) = \mathbf{R} = \begin{pmatrix} 1 & \Lambda & r_{1k} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ r_{k1} & \Lambda & 1 \end{pmatrix}$$

For j = 1, …, k gælder nu,

$$\mathrm{V}(\mathrm{X}_j) = a_{j1}^2 + \Lambda + a_{jm}^2 + \delta_j = h_j^2 + \delta_j = 1$$

h_j^2 kaldes kommunaliteter, og h_j angiver den brøkdel af X_j's varians, der stammer fra de m fælles faktorer. δ angiver tilsvarende den del af X_j's varians, der ikke stammer fra de m fælles faktorer.

Problemet består i at bestemme faktorvægtene, altså finde et skøn over matricen **A**. Ovenstående model **X** = **A F** + **G** betragtes, hvor **X** er k-dimensional og **F** m-dimensional. **X**'s korrelationsmatrix betegnes **R**, og **V** er den matrix, der fremkommer, når 1-tallerne i **R**'s diagonal erstattes med estimater af kommunaliteterne. Estimaterne vælges i intervallet $[r^2, 1]$, hvor r^2 er den multiple korrelationskoefficient mellem den pågældende variabel og de resterende. Normalt vælges enten r^2 eller 1. Desuden defineres

$$\mathbf{\Lambda}_*^{½} = \begin{pmatrix} \sqrt{\lambda_1} & \Lambda & 0 \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ 0 & & \sqrt{\lambda_m} \\ \mathrm{M} & & \mathrm{M} \\ 0 & \Lambda & 0 \end{pmatrix}$$

Den principale faktorløsning til estimationsproblemet er nu givet ved [7]:

$$\mathbf{P}\,\mathbf{\Lambda}_*^{½} = \left(\sqrt{\lambda_1}\mathbf{p}_1, \mathrm{K}\ , \sqrt{\lambda_m}\mathbf{p}_m\right),$$

hvor λ_i, i = 1, …, m, er de m største egenvektorer til **V**, og
hvor $\mathbf{p}_i$, i = 1, …, m, er de tilsvarende normerede egenvektorer.

3.2.3.1 Varimax rotation af faktorer

Da $\mathbf{P}\,\mathbf{\Lambda}_*^{½}$ er et estimat for matricen **A**, følger nedenstående udtryk umiddelbart

$$\mathbf{A}\,\mathbf{A}' \approx (\mathbf{P}\,\mathbf{\Lambda}_*^{½})(\mathbf{P}\,\mathbf{\Lambda}_*^{½})'$$

Multipliceres $\mathbf{P}\,\mathbf{\Lambda}_*^{½}$ med en vilkårlig m×m ortogonal matrix **Q**, fås:

$$(\mathbf{P}\,\mathbf{\Lambda}_*^{½}\,\mathbf{Q})(\mathbf{P}\,\mathbf{\Lambda}_*^{½}\,\mathbf{Q})' = (\mathbf{P}\,\mathbf{\Lambda}_*^{½})\,\mathbf{Q}\,\mathbf{Q}'(\mathbf{P}\,\mathbf{\Lambda}_*^{½})'$$
$$= (\mathbf{P}\,\mathbf{\Lambda}_*^{½})(\mathbf{P}\,\mathbf{\Lambda}_*^{½})' \approx \mathbf{A}\,\mathbf{A}'$$

Der kan altså findes vilkårligt mange estimater for **A**-matricen ved at multiplicere den principale faktorløsning med en ortogonal matrix. På denne måde kan **A**-matricen blive mere simpel, og fortolkningen af faktorerne kan derved blive lettere.

En af de mest anvendte måder at vælge **Q**-matricen på er at benytte Varimax kriteriet. Ved Varimax rotation tilstræbes det at få en faktorstruktur, hvor hver variabel hovedsagelig afhænger af kun en enkelt faktor [17]. En sådan faktorstruktur vil resultere i, at hver faktor repræsenterer en specifik egenskab. **Q** vælges her således, at

$$\sum_j m\left\{\sum_i\left(\frac{a_{ij}^2}{h_i^2}\right)^2 - \frac{1}{m}\left[\sum_i\left(\frac{a_{ij}^2}{h_i^2}\right)^2\right]^2\right\}$$

maksimeres [7]. Ovenstående vil bevirke, at mange af a_{ij}-erne bliver ca. 0, og mange bliver store, hvilket giver en simpel struktur for **A**-matricen, som vil være nem at fortolke.

Både principal komponentanalyse og faktoranalyse er datareducerende teknikker, men der er nogle væsentlige forskelle på de to analyser. Ved principal komponentanalyse er det variationen i data, der ønskes forklaret, mens det ved faktoranalyse derimod er korrelationen mellem variablene, der ønskes belyst. Derudover bliver de principale komponenter dannet som linearkombinationer af variablene, hvor det ved faktoranalyse er variablene, der er funktioner af de underliggende faktorer.

En anden væsentlig forskel mellem principal komponentanalyse og faktoranalyse er, at ved faktoranalyse multipliceres faktorerne med kvadratroden af den tilsvarende egenværdi. Længden af en faktor bliver herved proportional med den del af den totale varians, den forklarer.

Et program, der udfører faktoranalyse og Varimax rotation på de fundne faktorer, er blevet udarbejdet i samarbejde med min vejleder Bjarne Ersbøll.

3.2.4 Clusteranalyse

Clusteranalyse anvendes til at placere observationer eller variable i undergrupper eller clusters, sådan at medlemmerne i hver gruppe har nogle fælles egenskaber, som er forskellige fra de andre grupper.

Line Conradsen [8] benyttede i sit projekt clusteranalyse til opdeling af observationer, men det viste sig, at clusteranalyse anvendt på denne måde ikke var velegnet for den givne type af data. Clusteranalyse anvendes derfor i dette projekt til at gruppere variable og ikke observationer.

Der er mange måder at få grupperet variablene i clusters, men jeg vil her kun nævne den, som er blevet benyttet i dette projekt. For andre metoder, der anvendes til at danne clusters, kan henvises til [12, 8].

Ved denne analyse er, som udgangspunkt, alle variable samlet i den samme cluster, som herefter opdeles. Den cluster, der vælges til videre opsplitning, er den, der har den største egenværdi hørende til den anden principale komponent [16]. Hvis egenværdien hørende til den anden principale komponent er lille, betyder det, at denne clusters variable kan beskrives ved hjælp af den første principale komponent, det vil sige, der er ét fællestræk, der går igen for samtlige variable. Hvis den anden egenværdi derimod er stor, betyder det, at en enkelt principal komponent ikke er nok til at beskrive variablene, men at de bedst kan beskrives ved hjælp af to (eventuelt flere) forskellige egenskaber.

Den valgte cluster opdeles i to ved at bestemme de første to principale komponenter, rotere dem og dernæst tildele variablene til de to komponenter. Variablene tildeles den

komponent, med hvilken de har den største korrelation. Variablene opdeles altså efter, hvilket af to fællestræk i den forrige cluster der beskriver dem bedst.

Denne procedure fortsætter, indtil hver cluster kun har én egenværdi større end 1, hvilket indikerer, at samtlige variable i hver cluster kan beskrives ved hjælp af en enkelt principal komponent.

Proceduren varclus i SAS udfører grupperingen af variablene.

3.3 Superviserede metoder

Modsat ikke-superviserede metoder haves for superviserede metoder en eller flere outputvariable eller klassifikationsvariable, som inputvariablene har indflydelse på. Formålet er her at prædiktere værdien af responsvariablene ved hjælp af inputvariablene. I dette tilfælde er outputvariablen graden af epilepsi, det vil sige variablen epilep i datasættet. Det er for superviserede metoder muligt at lave et skøn over fejlraten ved prædiktionen og derved få en indikation af, hvor god metoden er.

3.3.1 Diskriminantanalyse

Når der er to eller flere grupper, der ønskes diskrimineret mellem, som tilfældet er med de opgivne data, kan diskriminantanalyse anvendes. Diskriminantanalysen udvælger et antal variable blandt samtlige, som opdeler data i de ønskede grupper bedst muligt. Samtidig bestemmes en klassifikationsregel, som er en funktion af de udvalgte variable, der kan klassificere fremtidige observationer i en af grupperne.

3.3.1.1 Diskrimination mellem to populationer

I det følgende forudsættes det, at de to grupper er normalfordelte og har samme dispersionsmatrix ($\pi_1 \approx \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ og $\pi_2 \approx \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$).

Når der skal diskrimineres mellem to eller flere grupper, må det antages, at der findes en tabsfunktion, som fortæller, hvor stort et tab der er ved at misklassificere en given observation. Hvis det desuden antages, at a priori sandsynligheden, p_i, for at få en observation fra gruppe i kendes, er løsningen (diskriminantfunktionen) til diskriminationsproblemet [7]:

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\tfrac{1}{2}\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\tfrac{1}{2}\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2-\log c$$

hvor

$$c=\frac{(\text{tab ved at vælge 1 givet 2 er sand})\,p_2}{(\text{tab ved at vælge 2 givet 1 er sand})\,p_1}$$

Hvis tabsfunktionen ikke er kendt, er det almindeligt at antage, at tabene ved misklassifikation er de samme. Tilsvarende gælder for a priorifordelingen, som også ofte sættes til at være de samme for alle grupper. Konstanten c sættes altså til 1 i de tilfælde, hvor tabsfunktionen og a priorifordelingen ikke kendes.

Haves to fordelinger med ukendte parametre, estimeres disse ved hjælp af observationer, og diskriminantfunktionen kan derefter bestemmes ud fra de estimerede fordelinger. Når a priorisandsynligheden for hver af grupperne desuden er ens, og misklassifikationstabet ligeledes er det samme for begge grupper, benyttes Mahalanobis afstand til at diskriminere mellem populationerne. Mahalanobis afstand fra en vilkårlig observation i til gruppe 1 og gruppe 2 er givet ved henholdsvis [17]:

$$(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)\quad\text{og}\quad(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)$$

Observationen placeres i den gruppe, som er tættest, målt ved Mahalanobis afstanden, det vil sige i den gruppe, for hvilken Mahalanobis afstanden er mindst. Observationen bliver altså placeret i gruppe 1 hvis:

$$\begin{aligned}&(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)\le(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)\\&\Leftrightarrow\mathbf{x}'\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)\ge\tfrac{1}{2}\hat{\boldsymbol{\mu}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1-\tfrac{1}{2}\hat{\boldsymbol{\mu}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2\end{aligned}$$

hvilket svarer fuldstændigt til diskriminantfunktionen ovenfor.

3.3.1.2 Flere end to populationer

Når der skal diskrimineres mellem flere end to grupper, benyttes samme fremgangsmåde, som når der diskrimineres mellem to grupper. Grupperne sammenlignes blot to og to, og den mest sandsynlige gruppe vælges herefter.

Det forudsættes igen, at alle grupper er normalfordelte og har samme dispersionsmatrix, $\pi_i \approx \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

Hvis det antages, at misklassifikationstabet er det samme for alle grupper, er diskriminantfunktionen for problemet følgende [7]:

$$S_i = \mathbf{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \tfrac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \log p_i$$

Hvis det antages, at a priorifordelingen er ens for alle grupper, kan de udelades fra beregningerne. S_i udregnes for hver gruppe, og grupperne sammenlignes to og to. Gruppe i vælges her frem for gruppe j, hvis $S_i > S_j$.

3.3.2 Stepvis diskriminantanalyse

Hvis det bedste sæt af variable til diskriminantanalysen ikke er kendt, er der tre forskellige måder at udvælge disse på: Forward selection, backward selection og stepwise selection.

Det antages, at data i hver population er normalfordelt og har den samme dispersionsmatrix. Ved forward selection er den første variabel, der medtages i diskriminantfunktionen, den der giver den bedste diskrimination mellem de givne grupper. Den næste variabel, der tilføjes, er den, der forøger diskriminationsevnen mest af de tilbageværende variable etc.

Der er flere måder at udvælge den næste variabel, der skal tilføjes. Den mest benyttede er Wilk's Λ, som er forholdet mellem indenfor-grupper kvadratsummen og den totale kvadratsum [17]:

$$\Lambda = \frac{SAK_{indenfor}}{SAK_{total}} = \frac{SAK_{indenfor}}{SAK_{mellem} + SAK_{indenfor}}$$

Ved hvert trin indsættes den variabel, der har den mindste Wilk's Λ. Den mindste værdi for Wilk's Λ ses af ovenstående udtryk at fås ved at minimere indenfor-grupper kvadratsummen og maksimere mellem-grupper kvadratsummen. Fordelingen af Wilk's Λ kan tilnærmes med en F-fordeling.

Mahalanobis afstand kan ligeledes benyttes som kriterium for, hvilken variabel der skal indsættes i diskriminantfunktionen. Hvor Wilk's Λ maksimerer den totale separation mellem alle grupper, er formålet med Mahalanobis afstand at sørge for separation mellem alle par af grupper. Ved hvert trin tilføjes altså her den variabel, der medfører den største stigning i separationen mellem de par af grupper, der er tættest på hinanden [17].

Ved backward selection startes med alle variable i diskriminantfunktionen og for hvert trin fjernes den variabel, som bidrager mindst til separationen mellem grupperne.

Stepwise selection er en kombination af de to foregående metoder, og det er den, der benyttes ved diskriminantanalysen i dette projekt. Der startes her som for forward selection uden variable, og den første, der medtages, er ligeledes den, der giver den bedst mulige diskrimination mellem grupperne. Herefter bliver der ved hvert trin enten fjernet eller tilføjet en variabel. En variabel, der allerede er i diskriminantfunktionen fjernes ligesom ved backward selection, hvis den ikke bidrager væsentligt til separationen mellem grupperne. Hvis der ikke fjernes en variabel i et givet trin, tilføjes i stedet den variabel, der øger separationen mellem grupperne mest.

Proceduren stepdisc benyttes til at udvælge variablene, og de fundne variable bruges derefter i proceduren discrim for at finde diskriminantfunktionen. Proceduren discrim kan ligeledes udføre krydsvalidering af data, hvor hver observation bliver klassificeret ved hjælp af en diskriminantfunktion beregnet ud fra de øvrige observationer. Krydsvalidering bliver beskrevet nærmere i afsnit 3.3.4 senere i dette kapitel.

3.3.3 Kanonisk diskriminantanalyse

Kanonisk diskriminantanalyse er en alternativ måde at foretage diskriminantanalyse på i tilfældet med flere end to populationer og benyttes især, som det meget nyttige visuelle værktøj den er.

Ved kanonisk diskriminantanalyse ønskes fundet en diskriminantfunktion, der maksimerer forholdet mellem variationen mellem grupper og variationen indenfor-grupper.

En mellem-grupper (between groups) matrix, **B**, defineres:

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})'$$

en indenfor-grupper (within groups) matrix, **W**:

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)'$$

og en matrix, **T** (total sum of squares), der er summen af de to ovenstående:

$$\mathbf{T} = \mathbf{B} + \mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}})(\mathbf{X}_{ij} - \overline{\mathbf{X}})'$$

Diskriminantfunktionen, **d**, bestemmes nu, sådan at

$$\varphi(\mathbf{d}) = \frac{\mathbf{d}' \mathbf{B} \mathbf{d}}{\mathbf{d}' \mathbf{W} \mathbf{d}}$$

maksimeres [7]. Den første kanoniske variabel findes herefter som $\mathbf{d}_1\mathbf{x}$.

Efter den første diskriminantfunktion, $\mathbf{d}_1$, er fundet, bestemmes nu en ny diskriminantfunktion, $\mathbf{d}_2$, der ligeledes maksimerer ovenstående, men som samtidig er ortogonal til $\mathbf{d}_1$. Tilsvarende findes $\mathbf{d}_3$, der skal være ortogonal til både $\mathbf{d}_1$, og $\mathbf{d}_2$, $\mathbf{d}_4$, der

skal være ortogonal til $\mathbf{d}_1$, $\mathbf{d}_2$ og $\mathbf{d}_3$, osv. indtil antallet af kanoniske variable enten er lig med antallet af originale variable eller lig med antallet af grupper, der ønskes diskrimineret mellem, minus en, alt efter hvilket af disse to der er mindst.

Når de kanoniske variable er fundet, kan observationerne projiceres ned på ($\mathbf{d}_1$, $\mathbf{d}_2$) planet, hvilket giver et billede af, hvor godt grupperne er separeret. Det er i ($\mathbf{d}_1$, $\mathbf{d}_2$) planet, at punkterne separeres bedst muligt. Ofte projiceres observationerne desuden ned på ($\mathbf{d}_1$, $\mathbf{d}_3$) planet og på ($\mathbf{d}_2$, $\mathbf{d}_3$) planet.

De kanoniske variable findes ved hjælp af proceduren candisc, og de kan herefter plottes.

3.3.4 Krydsvalidering

Som nævnt kort under afsnittet om diskriminantanalyse kan krydsvalidering benyttes til at klassificere en observation ud fra analyse af de øvrige observationer, det vil sige, krydsvalidering er en måde at teste, hvor godt et fundet resultat er. En misklassifikationsrate, eller fejlrate, kan bestemmes ved krydsvalidering som antallet af misklassificerede observationer delt med antallet af observationer i alt. Størrelsen af fejlraten er hermed et udtryk for, hvor god den fundne model er. Ved almindelig diskriminantanalyse kan de bedste variable udvælges ved hjælp af stepdisc, og dernæst kan en diskriminantfunktion findes med discrim, som samtidig udfører krydsvalidering. På denne måde er den observation, der ønskes klassificeret, dog stadig med ved udvælgelsen af de bedste variable og kan derfor ikke betragtes som en ny variabel, der ønskes klassificeret.

Foruden at udføre krydsvalidering med discrim, blev et program benyttet, som blev udviklet i forbindelse med Line Conradsens eksamensprojekt [8]. Dette program fjerner en observation og udvælger dernæst de bedste variable og bestemmer en diskriminantfunktion baseret på de resterende observationer. Til sidst klassificeres den fjernede observation ved hjælp af den fundne diskriminantfunktion. Dette gøres for samtlige observationer, hvorefter fejlraten kan bestemmes. Denne form for krydsvalidering, hvor en enkelt observation udelades og klassificeres på basis af analyse af de øvrige observationer, kaldes leave-one-out krydsvalidering [12]. Krydsvalidering foretaget med discrim er altså også leave-one-out krydsvalidering.

Frem for kun at fjerne en enkelt observation og klassificere denne, fjernes ofte flere – dette kaldes k-fold krydsvalidering. Her opdeles datasættet i k lige store dele, og den ene af disse holdes uden for analysen. Den ønskede analyse af data udføres på de k-1 andre dele, og herefter klassificeres den k'te gruppes observationer. De værdier, der almindeligvis vælges for k, er 5 eller 10, det vil sige, datasættet deles i enten 5 eller 10 dele.

Problemet med 5- eller 10-fold krydsvalidering opstår, når data ikke indeholder ret mange observationer, hvilket kan betyde, at fejlraten bliver overestimeret – i mit tilfælde, hvor der kun er 18 rotter, vil 5-fold krydsvalidering bevirke, at tre til fire observationer skal fjernes, og analysen udføres på de resterende kun 14 eller 15 observationer, hvilket medfører en større usikkerhed på resultatet. Et andet problem er, hvordan observationerne udvælges blandt de 18 – hvis det er tilfældig udvælgelse, som foreslået i [12], er der en risiko for, at de observationer, der fjernes, stammer fra samme gruppe. Da datasættet er så lille, kan dette betyde, at en hel klasse fjernes fra analysen.

Ulempen ved at benytte leave-one-out krydsvalidering er derimod, at da de N datasæt, der benyttes til analyserne, er så ens, kan fejlraten have stor varians. En anden ulempe kunne være, at analysen skal gennemføres N gange, men dette er i mit tilfælde ikke så stort et problem, da N her kun er lig 18.

I dette projekt er kun benyttet leave-one-out krydsvalidering, da antallet af observationer er så småt.

3.3.5 Regressionsanalyse

Regressionsanalyse er en anden måde at diskriminere mellem grupper på. Ved diskriminantanalyse skelnes blot mellem grupperne, mens der ved regressionsanalyse ligeledes tages højde for graden af epilepsi. Dette skal forstås på den måde, at gruppen af rotter med svær grad af epilepsi følger efter gruppen med mild grad, som følger efter gruppen med epilepsi under udvikling, som igen følger efter den raske gruppe. Graden af epilepsi, rækkefølgen af grupperne, betyder altså noget.

3.3.5.1 Logistisk regression

Det ønskes undersøgt, hvordan graden af epilepsi afhænger af de givne variable ved hjælp af logistisk regression. Logistisk regression benyttes ofte, når responsvariablene er enten binære, ordinale, som tilfældet er i dette projekt, eller nominale. Det er derudover et nyttigt værktøj, når fordelingen af data ikke kendes, da en kendt fordeling af variablene ikke er en forudsætning for at udføre analysen [17]. Målet med logistisk regression er at finde sammenhængen mellem de uafhængige variable og sandsynlighederne for at tilhøre hver af de givne grupper. Samtidig sørges der for, at hver af disse sandsynligheder ligger mellem 0 og 1, og at summen af dem er 1 [12].

Når der skelnes mellem to klasser, kan disse sandsynligheder findes som:

$$p_i = p(\mathbf{x}_i) = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta}' \mathbf{x}_i))}$$

da dette er en voksende funktion med asymptoterne p = 0 og p = 1 [7]. **x** er de forklarende variable.

logit er defineret som

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

hvorved fås, at den logistiske regressionsmodel kan skrives som:

$$\text{logit}(p_i) = \alpha + \boldsymbol{\beta}' \mathbf{x}_i$$

som er en lineær model i logits.

Når der skelnes mellem k+1 ($k \geq 1$) klasser, baseres modellen på de kumulative sandsynligheder for responsgrupperne i stedet for på de individuelle sandsynligheder, og modellen får derved følgende udseende:

$$\text{logit}(\Pr(Y \leq i \,|\, x) = \alpha_i + \boldsymbol{\beta}' \mathbf{x}, \quad 1 \leq i \leq k$$

hvor α_i er k intercept parametre [16]. Parametrene α_i og $\boldsymbol{\beta}$ findes ved maximum likelihood estimation, jævnfør [7].

Ligesom ved diskriminantanalyse er det bedste sæt af variable ikke nødvendigvis kendt, og disse kan derfor udvælges på de samme tre måder, som beskrevet i afsnittet om diskriminantanalyse: Ved forward selection, backward selection og ved stepwise selection.

Den logistiske regressionsanalyse blev udført dels med proceduren proc logistic og dels med SAS Enterprise Miner's 'Regression Node'.

3.3.6 Klassifikationstræer

En tredje måde at diskriminere mellem grupper på er ved hjælp af ikke-parametriske klassifikationstræer, en kendt fordeling af data er altså heller ikke for denne metode forudsat. Her er prædiktionsreglerne givet i form af binære beslutningstræer. Der er to hovedmål ved konstruktionen af beslutningsreglerne, det første er at konstruere den mest præcise beslutningsregel som muligt, og det andet er at konstruere den beslutningsregel, der giver størst indsigt i problemet [4].

De k variable betegnes $x_1, x_2, \ldots, x_k$. Klassifikationstræet opbygges ved først at undersøge splits af formen $x_1 < C$, hvor C er en konstant i intervallet mellem minimum af x_1 og maksimum af x_1. C ændres, til det bedst mulige split er fundet. Ved det bedst mulige split forstås den deling, som adskiller de givne klasser mest. Herefter undersøges alle de øvrige variable på tilsvarende måde, og det bedste split af alle de fundne beholdes som den første knude.

Det kriterium, der almindeligvis benyttes, for hvilket split der er det bedste, er Gini kriteriet. Størrelsen af Gini indekset er et udtryk, hvor godt et split er. Gini indekset er givet ved:

$$i(t) = \sum_{j \neq i} p(j|t)\, p(i|t)$$

hvor $p(j|t)$ er de estimerede klassesandsynligheder ved knude t [3].

Ovenstående Gini indeks kan ligeledes skrives

$$i(t) = 1 - \sum_j p^2(j|t)$$

Det bedste split af de fundne findes som det split, der giver det laveste Gini indeks i(t), for jo bedre klasserne er separeret, jo tættere vil $p(j|t)$ ligge på enten 0 eller 1. Det ideelle vil være, at der i en knude kun findes en enkelt klasse – for denne klasse vil p være lig 1, og for de andre klasser vil p være lig 0, hvorved Gini indekset vil få den lavest tænkelige værdi, nemlig 0.

Data deles nu i to efter beslutningsreglen fundet i den første knude, og ovenstående procedure gentages i hver af de to nye grene. Herved deles det oprindelige datasæt i fire grupper, og for hver knude gentages proceduren. Denne fremgangsmåde fortsættes, indtil et stort træ er blevet opbygget med få observationer i hver slutknude. Herefter 'beskæres' træet. Hvilken del af træet der skæres af, afgøres ved at beregne estimerede fejlrater i hver gren, det vil sige, det estimeres, hvor stor en del af data der misklassificeres i hver gren. Dette gøres typisk ved hjælp af et testdatasæt, hvor det vides, hvilke klasser observationerne tilhører. Andelen af misklassificerede observationer vil typisk være stor for store deltræer, aftage som deltræet bliver mindre og herefter stige igen, når deltræet bliver for småt. Det træ, der giver den laveste fejlrate, beholdes.

En anden måde at teste klassifikationstræet på er ved krydsvalidering, som blev beskrevet i afsnit 3.3.4.

Klassifikationstræ for data opbygges ved hjælp af SAS Enterprise Miner's 'Tree Node'.

Alle de benyttede programmer samt flowdiagram for analyser foretaget i Enterprise Miner kan ses i bilag 1.

4 Resultater

I dette kapitel vil de vigtigste resultater for de tre grupper af metoder blive gennemgået og diskuteret, og derefter vil resultaterne af de ikke-superviserede og de superviserede metoder blive sammenlignet.

4.1 Resultater af univariate metoder

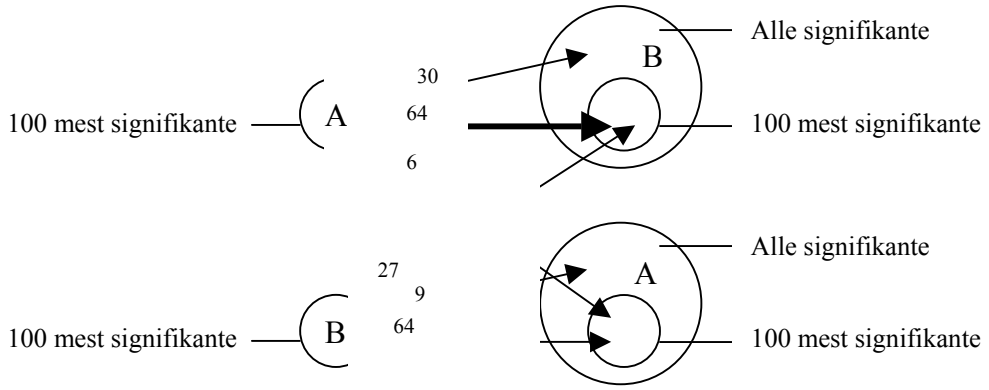
Som nævnt i afsnit 3.1 benyttes univariate metoder til, for hver enkelt variabel (plet på gelen), at sammenligne de givne grupper. Da det er standard til denne type data at benytte univariate metoder, har jeg for at have et sammenligningsgrundlag for mine øvrige analyser udført, dels ikke-parametrisk ensidet variansanalyse på de originale data, dels en t-test på de logtransformerede. For den ikke-parametriske variansanalyse er først gruppen af syge rotter blevet sammenlignet med gruppen af raske, og dernæst er de fire grupper (kontrolgruppen, under udvikling, mild grad af epilepsi og svær grad af epilepsi) blevet sammenlignet. For alle tre analyser er benyttet et signifikansniveau på 5 %.

For at lette læsningen af det følgende vil de to ikke-parametriske variansanalyser i dette afsnit blive benævnt A og B. De grupper, der sammenlignes i de to analyser, er:

A: ’Kontrolgruppe’ og ’syg’
B: ’Kontrolgruppe’, ’under udvikling’, ’mild’ og ’svær’.

Resultaterne af de tre analyser kan ses i bilag 2-4, hvor pletterne på gelen er ordnet efter signifikans. For A blev fundet 235 signifikante variable, for B blev fundet 251, og for t-testen blev fundet 247 signifikante pletter.

De 100 mest signifikante pletter fundet ved de to ikke-parametriske variansanalyser blev sammenlignet, og det viste sig, at 64 ud af de 100 var de samme. Der blev altså fundet 36 variable kun for A, og 36 variable blev fundet kun for B. I figur 4 på næste side er sammenhængen mellem resultaterne af de to analyser, A og B, illustreret:



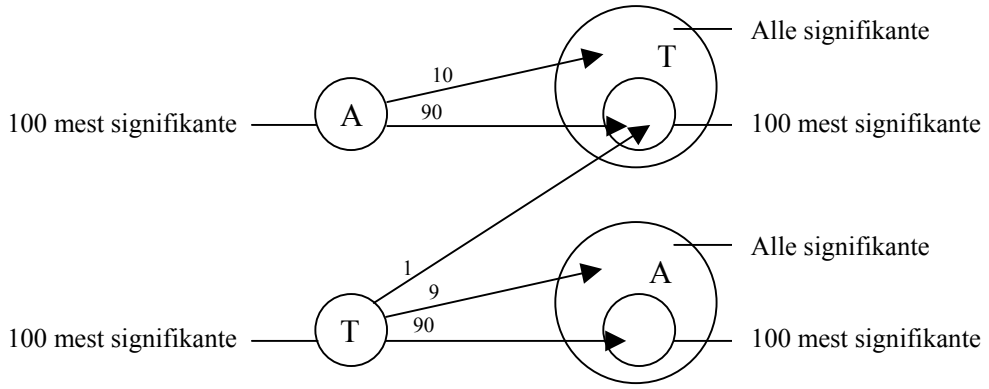
Figur 4: Ringene til venstre repræsenterer de 100 mest signifikante pletter fundet ved henholdsvis A og B, de store ringe til højre repræsenterer alle signifikante pletter fundet ved B og A, de mindre ringe inde i de to store gør det ud for de 100 mest signifikante pletter fundet ved de to analyser. Pilene fra ringene til venstre angiver, hvor mange af de 100 mest signifikante pletter der blev fundet blandt tilfældene til højre i figuren. For eksempel blev 64 af de 100 mest signifikante pletter fundet ved A også fundet blandt de 100 mest signifikante ved B (fremhævet pil i figur).

Som figur 4 viser, blev 30 af de 36 variable, som kun A fandt blandt de 100 mest signifikante, også fundet som signifikante ved B. De er bare ikke blandt de 100 mest signifikante. Der er altså kun 6 ud af de 100 pletter, som kun blev fundet som signifikante ved A og ikke ved B (pil fra A til A).

Af figuren ses omvendt, at det kun er 9 ud af de 36 pletter, fundet kun ved B, som også er signifikante ved A, og hele 27 pletter blev kun fundet signifikante ved B.

Da A kun skelner mellem epilepsi og ej epilepsi, mens B skelner mellem flere grader af epilepsi kan det forventes, at der bliver fundet forskellige signifikante variable ved de to analyser. Hvis der er forskel på to grupper i B, kan det skyldes en forskel i graden af epilepsi, og dette vil ikke nødvendigvis kunne ses, når grupperne med epilepsi slås sammen til en. Derimod er det mere rimeligt, når der er forskel på den syge og den raske gruppe, at denne forskel stadig er gældende, når den syge gruppe deles op i undergrupper.

De 100 mest signifikante pletter fundet ved t-testen blev dernæst sammenlignet med de 100 mest signifikante pletter for A, den ikke-parametriske variansanalyse med en syg og en rask gruppe. Det viste sig her, at der var 90 ud af de 100 variable, der var de samme og dermed kun 10 variable, der adskilte sig ved de to analyser. På næste side ses figur 5, som viser sammenhængen mellem resultatet af A og t-testen (T):



Figur 5: Ringene til venstre repræsenterer igen de 100 mest signifikante pletter, men denne gang fundet ved henholdsvis A og T (t-testen), de store ringe til højre repræsenterer alle signifikante pletter fundet ved T og A, de mindre ringe inde i de to store gør det ud for de 100 mest signifikante pletter fundet ved de to analyser. Pilene fra ringene til venstre angiver, hvor mange af de 100 mest signifikante pletter der blev fundet blandt tilfældene til højre i figuren.

Som det ses af figur 5, var de 10 variable fundet blandt de 100 mest signifikante for A alle blandt de 247 signifikante pletter fundet ved t-testen (T). Tilsvarende var 9 ud af 10 af de variable, som kun t-testen fandt, blandt de 235 signifikante pletter, som blev fundet ved A. Den sidste variabel, som kun t-testen og ikke A fandt (pil fra T til T), har dog vist sig at være nr. 3 på listen over de mest signifikante pletter for B.

De tre univariate analyser giver altså resultater, der stemmer meget godt overens.

4.2 Resultater af ikke-superviserede metoder

4.2.1 Inddeling af rotter i grupper

Rotterne blev af CPA inddelt i de fire allerede nævnte grupper, kontrolgruppe, epilepsi under udvikling, mild grad af epilepsi og svær grad af epilepsi, på basis af målinger af fire forskellige variable, der giver oplysning om, hvor svær grad af epilepsi den enkelte rotte har udviklet. De fire variable er henholdsvis HAFD nummer, High Activity Frequency Discharge, som er et mål for kortslutning i hjernen, altså en måling af hvor stor skade det elektriske chok har forårsaget, SE varighed, der er et mål for varigheden af

epilepsianfaldene, SE antal, der fortæller, hvor mange anfald hver rotte har haft og SE dag, der angiver, hvilken dag det første anfald indtraf. Nedenstående tabel 2 viser værdierne af de fire variable for de 12 stimulerede rotter:

Rottens nr.	HAFD nummer	SE varighed	SE antal	SE dag	Grad af epilepsi ifølge CPA
R1	116	8.2167	1	9	Mild
R2	287	21.4333	5	8	Svær
R3	81	7.9333	0		Under udvikling
R4	91	9.7833	0		Under udvikling
R5	88	7.7000	15	13	Svær
R6	120	5.9333	24	10	Svær
R7	95	3.7667	3	9	Mild
R8	87	8.7333	2	8	Mild
R9	102	2.8500	2	8	Mild
R10	38	10.4000	0		Under udvikling
R11	0	0	0		Ikke reageret
R12	85	6.3167	0		Under udvikling

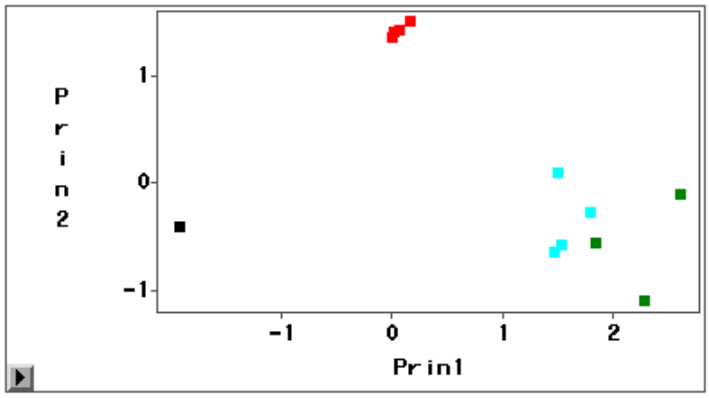
Tabel 2: Værdierne af de fire variable, HAFD, SE varighed, SE antal og SE dag, der er bestemmende for, hvilken grad af epilepsi den enkelte rotte menes at have.

Det fremgår, at der ingen målinger er for rotte nummer 11, der altså, som tidligere nævnt, ikke har reageret på stimuleringen. Som sagt kunne dette skyldes, at elektroderne ikke har siddet ordentligt eller eventuelt, at den pågældende rotte af en eller anden grund er resistent overfor behandlingen.

For at undersøge om de enkelte rotter er blevet placeret i den mest hensigtsmæssige gruppe, eller der eventuelt er en anden opdeling af rotterne, der er bedre, udførtes en principal komponentanalyse på de fire variable. Som for alle analyser i dette projekt er både de logtransformerede data samt de originale og rangen af data blevet undersøgt. For at kunne logtransformere data måtte der først lægges 0.5 til alle tal, hvor tallet 0.5 er valgt, da dette er halvdelen af det mindste tal i tabellen, jævnfør afsnit 2.4 om transformation af data. Herefter blev logaritmen til alle tal beregnet.

For overhovedet at kunne lave en analyse af ovenstående tabel må de manglende værdier for R3, R4, R10 og R12 erstattes af en meningsfuld talværdi. Da SE dag angiver, hvilken dag den pågældende rotte får det første anfald, og det kendetegnende ved de fire nævnte rotter netop er, at de ikke har haft synlige anfald, tildeles de en værdi, der er højere end den højeste værdi for SE dag for de andre rotter. For at denne værdi ikke skal styre analysen af data vælges en værdi, der ligger tæt på de øvrige værdier for SE dag, men som stadig er højere. Jeg har foretaget den principale komponentanalyse med forskellige værdier af SE dag for de fire nævnte rotter, og det viser sig, at den opdeling af rotterne, der findes, ikke afhænger væsentligt af, hvilken værdi for SE dag der benyttes.

I nedenstående figur er plottet værdien af den første og den anden principale komponent for de logtransformerede data for hver rotte. De principale komponenter er baseret på de fire variable, HAFD nummer, SE varighed, SE antal og SE dag. Her er benyttet SE dag-værdien 20 for de fire omtalte rotter:

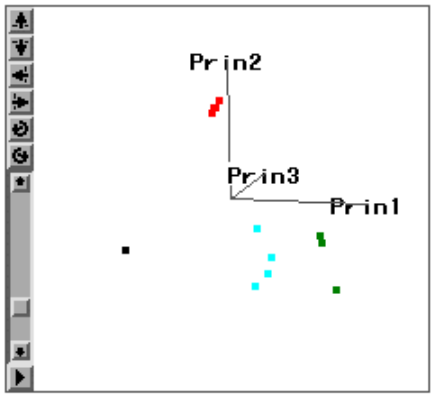


Figur 6: Plot af de to første principale komponenter for logtransformerede data, hvor de principale komponenter er baseret på de fire variable HAFD nummer, SE varighed, SE antal og SE dag. Sort: kontrolgruppe, rød: epilepsi under udvikling, blå: mild grad af epilepsi og grøn: svær grad af epilepsi.

Som det fremgår af figur 6, kan rotterne opdeles i grupper. De grupper, der her er markeret, er dem, som CPA har opgivet, se tabel 2, hvor den sorte gruppe er kontrolgruppen, den røde er epilepsi under udvikling, den blå er mild grad af epilepsi og den grønne er svær grad af epilepsi.

Inddrages desuden den tredje principale komponent, kan det på figur 7 på næste side endnu tydeligere ses, at den af CPA valgte inddeling er yderst fornuftig. I figuren er for

hver rotte plottet værdien af den første, den anden og den tredje principale komponent, og koordinatsystemet er blevet drejet, så opdelingen af rotterne klarest fremgår:



Figur 7: Plot af de tre første principale komponenter. Tydelig opdeling i de fire grupper.

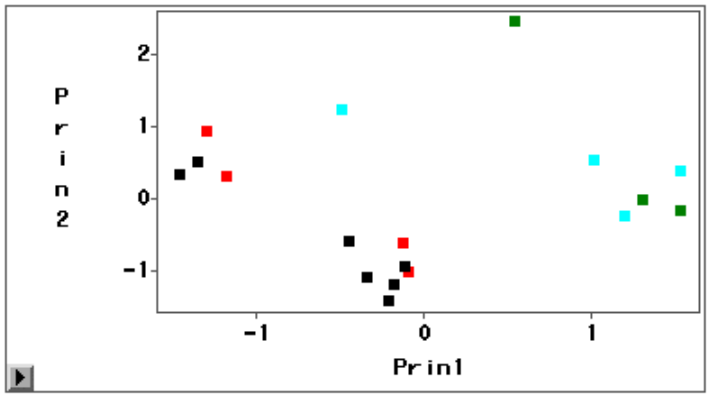
Når den principale komponentanalyse udføres på de originale data samt på rangen af data, findes den samme opdeling.

4.2.2 Resultat af principal komponentanalyse og faktoranalyse

Efter at have fået bekræftet at CPA's opdeling af rotterne i de fire grupper er den mest hensigtsmæssige, blev den principale komponentanalyse udført på datasættet med de 1849 variable.

Som beskrevet i afsnit 3.2.1, udførtes principal komponentanalyse på det transponerede datasæt i stedet for på det oprindelige på grund af det store antal variable. Dette blev gjort ved hjælp af Eckart-Young's sætning. Principal komponentanalyse blev først benyttet til at finde ud af, hvordan observationerne grupperer sig, og om denne gruppering stemmer overens med den opgivne fordeling af rotterne i de fire grupper.

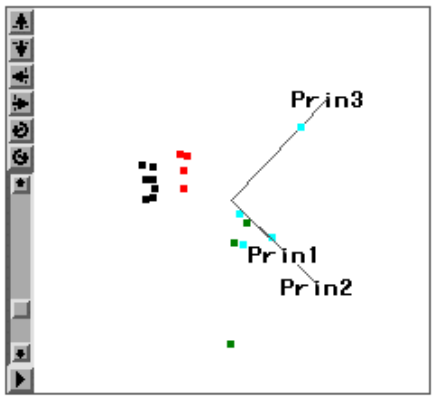
For de 18 observationer blev der for de logtransformerede data fundet grupperinger, der viste sig som på næste side. Opdelingen af observationerne er her illustreret ved hjælp af de to første principale komponenter:



Figur 8: Plot af de første to principale komponenter for logtransformerede data. Den principale komponentanalyse er her baseret på de 1849 variable. De fire grupper adskilles ikke, men ses på kontrolgruppen (sort) og epilepsi under udvikling (rød) under et og de to grupper med epilepsi i 'udbrud' (blå og grøn) under et er der en rimelig opdeling.

Som figur 8 viser, bliver de fire grupper ikke adskilt, men hvis der ses på grupperne to og to, kontrolgruppen (sort) og epilepsi under udvikling (rød) som den ene og mild (blå) og svær (grøn) grad af epilepsi som den anden, er der rimelig klar opdeling. Gruppen med epilepsi under udvikling ligner altså mere den raske gruppe end nogen af de to grupper med epilepsi i 'udbrud'.

Hvis den tredje principale komponent medtages fås en noget bedre opdeling af grupperne, som det kan ses i figur 9 nedenfor:



Figur 9: Plot af de tre første principale komponenter for logtransformerede data. Kontrolgruppe (sort) og epilepsi under udvikling (rød) adskilles pænt. Mild (blå) og svær (grøn) grad af epilepsi adskilles ikke.

Det fremgår af figur 9, at kontrolgruppen og gruppen med epilepsi under udvikling (henholdsvis sort og rød) nu bliver pænt adskilt fra hinanden og fra grupperne med epilepsi, mens det stadig ikke er muligt at skelne mellem mild (blå) og svær (grøn) grad af epilepsi ud fra de principale komponenter. Selv hvis både fjerde og femte principale komponent medtages, bliver opdelingen ikke forbedret – det er stadig ikke muligt at skelne mellem de to grupper med epilepsi.

Selvom det ikke er muligt at adskille de to grader af epilepsi, kan det dog være nyttigt at kunne skelne mellem rask, epilepsi under udvikling og epilepsi. Hvis formålet er at forstå, hvordan epilepsi udvikles, er det måske ikke så vigtigt at vide, hvilken grad af epilepsi rotten har, men derimod om den har epilepsi i 'udbrud' eller kun under udvikling. Hvis der kan findes væsentlige forskelle mellem disse grupper, kan det måske på længere sigt være muligt at forhindre yderligere udvikling af epilepsi.

For hverken de originale data eller for rangen af data findes en lige så god opdeling. I begge disse tilfælde bliver kontrolgruppen og gruppen med epilepsi under udvikling ikke adskilt særligt tydeligt, men de to grupper under et adskilles rimeligt fra de to grupper med epilepsi i 'udbrud' under et.

4.2.2.1 Gruppering af variable

Tilsvarende blev principal komponentanalyse, kombineret med faktoranalyse, benyttet til at finde grupperinger blandt variablene, blandt pletterne på gelerne. Egenvektorerne for det transponerede datasæt blev fundet ved hjælp af den principale komponentanalyse, og ud fra disse blev egenvektorerne for det oprindelige datasæt dernæst bestemt med Eckart-Young's sætning. Egenvektorerne blev normeret, så de alle havde længden 1, hvorefter de blev Varimax roteret. De roterede faktorer får herved alle lige stor vægt. Formålet med faktoranalysen var at finde grupper af pletter, der varierer på samme måde eller sagt på en anden måde finde grupper af proteiner, der reagerer på samme vis.

De 1849 pletter på gelen er ikke alle forskellige proteiner, nogle af dem er modifikationsprodukter af det samme protein. Proteinerne kan for eksempel have et forskelligt antal uladede sidegrupper (for eksempel glucosemolekyler) koblet på, og da proteinerne ved anden dimension i 2d-elektroforese adskilles efter deres molekylvægt, vil det samme protein ofte udmønte sig i mere end en plet på gelen. Pletterne vil være det

samme protein, men vil ikke have samme molekylvægt. En anden mulighed kan være, at proteinerne har et forskelligt antal ladede sidegrupper (for eksempel fosfatgrupper, der er meget lette i forhold til proteinet selv) koblet på, hvilket vil medføre, at det isoelektriske punkt bliver rykket. Igen bliver resultatet, at det samme protein viser sig som flere pletter på gelen.

Ovenstående kan være en årsag til, at der er god grund til at antage, at der vil være grupper af pletter på gelen, der opfører sig på samme måde. Det må derudover være rimeligt at forestille sig, at der er nogle proteiner, der reagerer ens.

Der er mange forskellige metoder til at afgøre, hvor mange egenvektorer eller faktorer der skal roteres, det vil sige, hvor mange grupper variablene bør opdeles i. Den jeg har valgt at benytte, er Horn's metode [14], og den beskrives ganske kort her.

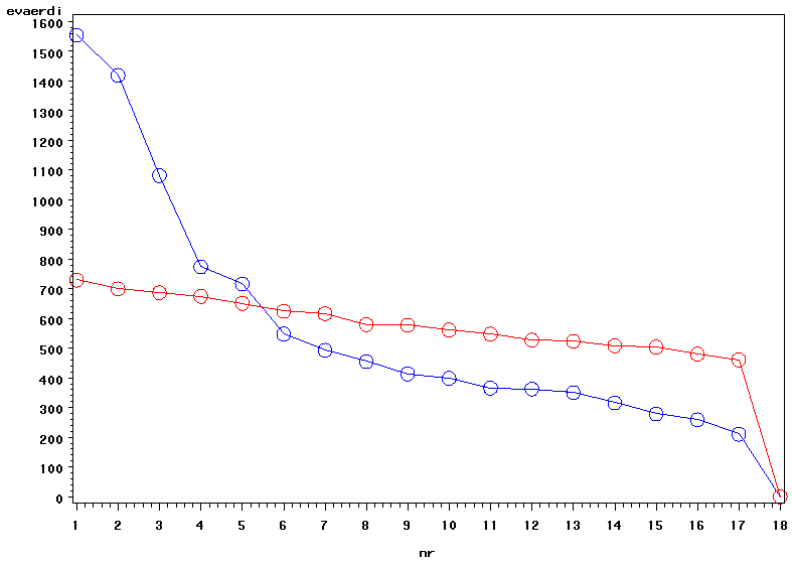
Horn's metode

Ved Horn's metode udføres principal komponentanalyse dels på det datasæt, der arbejdes på og dels på dette datasæt permuteret. Det permuterede datasæt fremkommer ved at permutere hver variabel i det oprindelige datasæt, det vil sige ved vilkårligt at blande observationerne for hver søjle i datasættet. Værdierne hørende til hver variabel vil herved stadig befinde sig i samme søjle, men i andre rækker end det oprindelige datasæt. Det permuterede datasæt vil på denne måde have samme totale varians, da variansen indenfor hver enkelt variabel forbliver den samme, men sammenhængen mellem variablene og observationerne vil være tabt.

For hver af de to datasæt bestemmes egenværdierne, som dernæst plottes som funktion af egenværdiens nummer. Dette kaldes et scree-plot. De to scree-plots tegnes i samme figur.

Ved Horn's metode udvælges de principale komponenter for det oprindelige datasæt, der indeholder mere information end de principale komponenter for det permuterede datasæt, og dette kan på scree-plottet aflæses ved skæringen mellem de to kurver.

I figur 10 på næste side ses scree-plot for både det logtransformerede datasæt og dets permuterede. Den blå kurve er egenværdierne for det oprindelige datasæt, og den røde er egenværdierne for det permuterede.



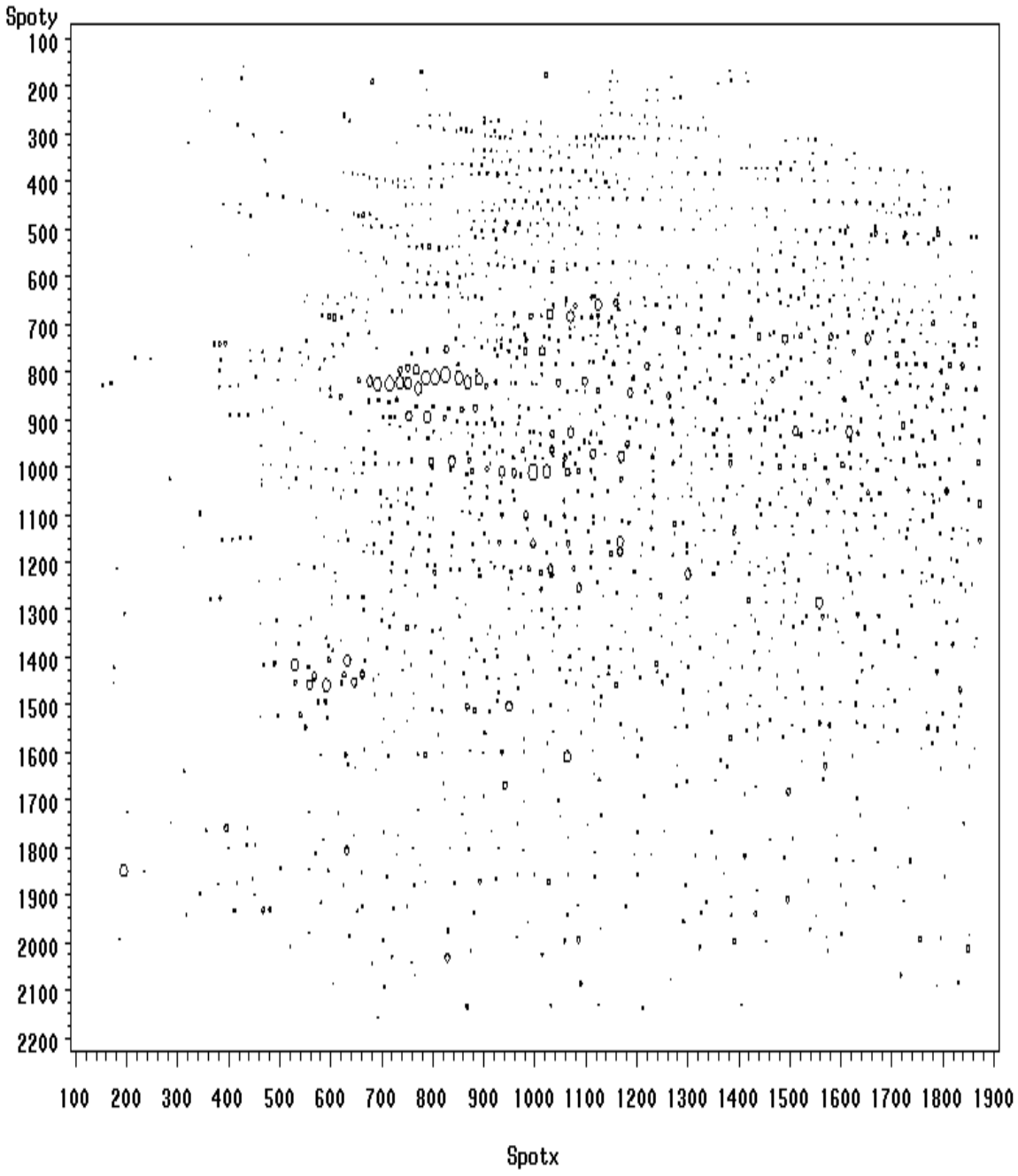
Figur 10: Scree-plot for logtransformerede data og dets permuterede.

Det fremgår af figur 10, at der for de logtransformerede data skal benyttes 5 principale komponenter, da egenværdien, og dermed variansen, for den sjette principale komponent, ifølge figuren er mindre end egenværdien hørende til den tilsvarende principale komponent for det permuterede datasæt. Når der medtages 5 faktorer i analysen, bliver der taget højde for godt 55 % af den samlede variation i data.

En tilsvarende undersøgelse er lavet for de originale data og for rangen af data, og her er antallet af principale komponenter (og dermed egenvektorer), der skal benyttes, fundet til henholdsvis 5 og 6. De tilhørende scree-plots kan ses i bilag 5.

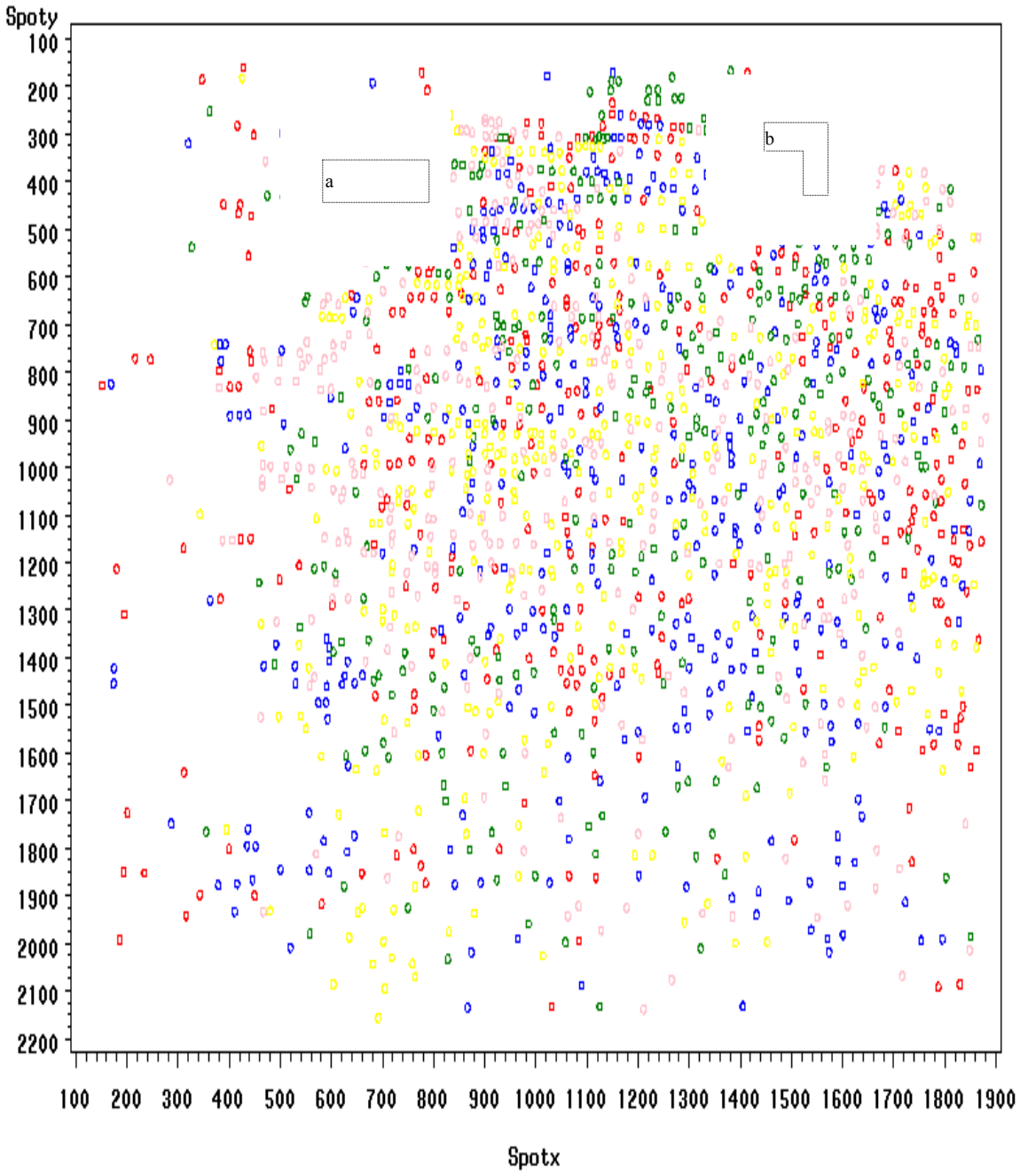
Der blev udført faktoranalyse med 5 faktorer for de logtransformerede data, 5 faktorer for de originale og 6 faktorer for rangen af data, og i alle tre tilfælde blev resultatet illustreret på et plot af gelen. Hver plet på gelen er her farvet efter, hvilken faktor der har størst indflydelse på pletten. For de logtransformerede data vil der altså være fem farver på plottet af gelen, og hver af disse farver repræsenterer en gruppe. På plottet af gelen er der ikke taget hensyn til intensiteten af pletten. Hvis størrelsen af pletterne skal afspejle intensiteten, bliver nogle af pletterne så små, at det ikke er muligt at se, hvilken farve de har, og dermed afgøre, hvilken faktor der har størst indflydelse på de pågældende pletter. For alligevel at kunne få et indtryk af hvor stor intensiteten af pletterne er, viser figur 11

nedenfor et plot af pletternes placering på gelen, og deres forskellige størrelser er et udtryk for størrelsen af intensiteten:

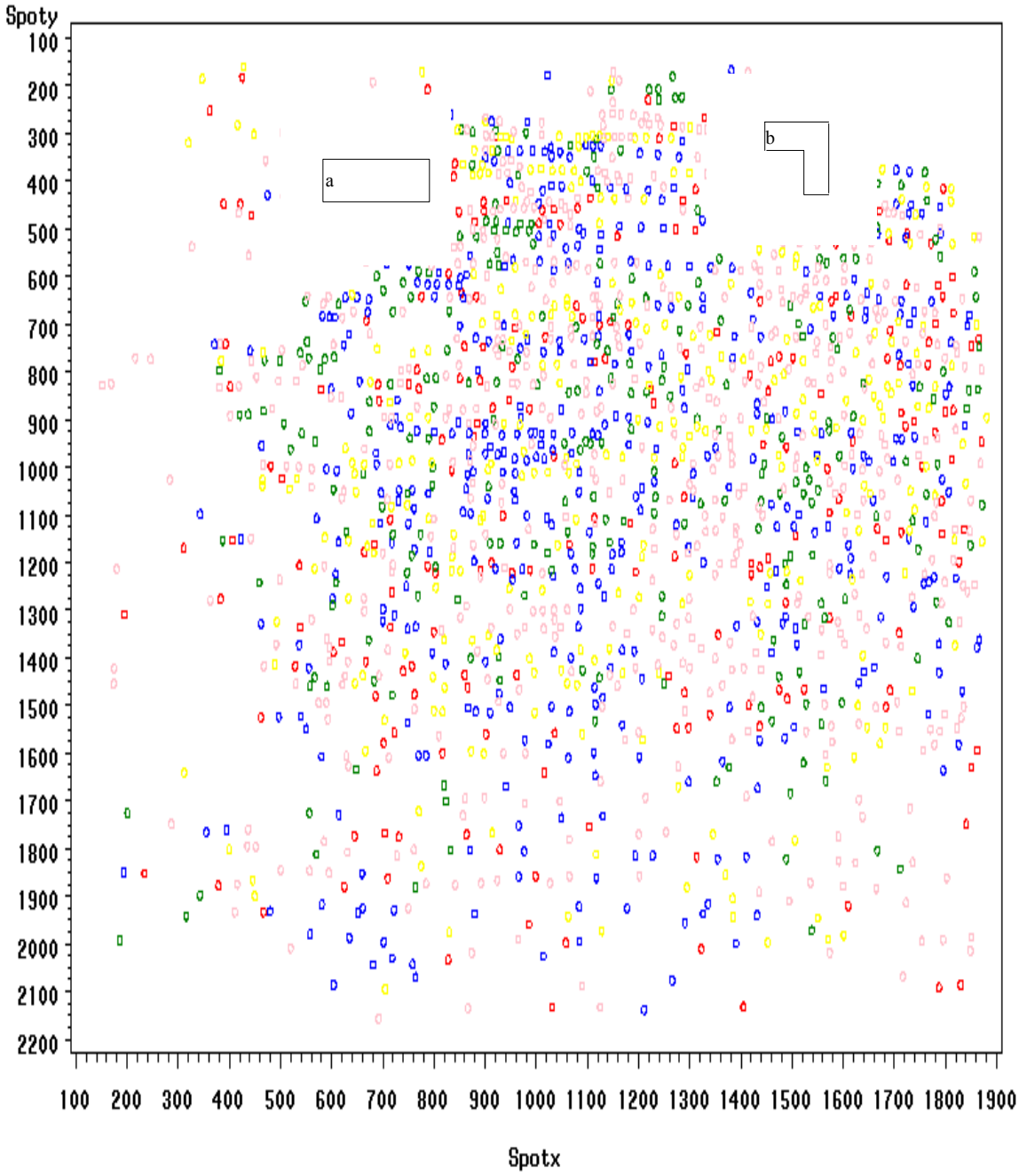


Figur 11: Plot af pletternes placering på gelen. Størrelsen af de enkelte pletter er udtryk for størrelsen af intensiteten. X-aksen er udtryk for det isoelektriske punkt, pI, og y-aksen er udtryk for molekylvægten for proteinerne.

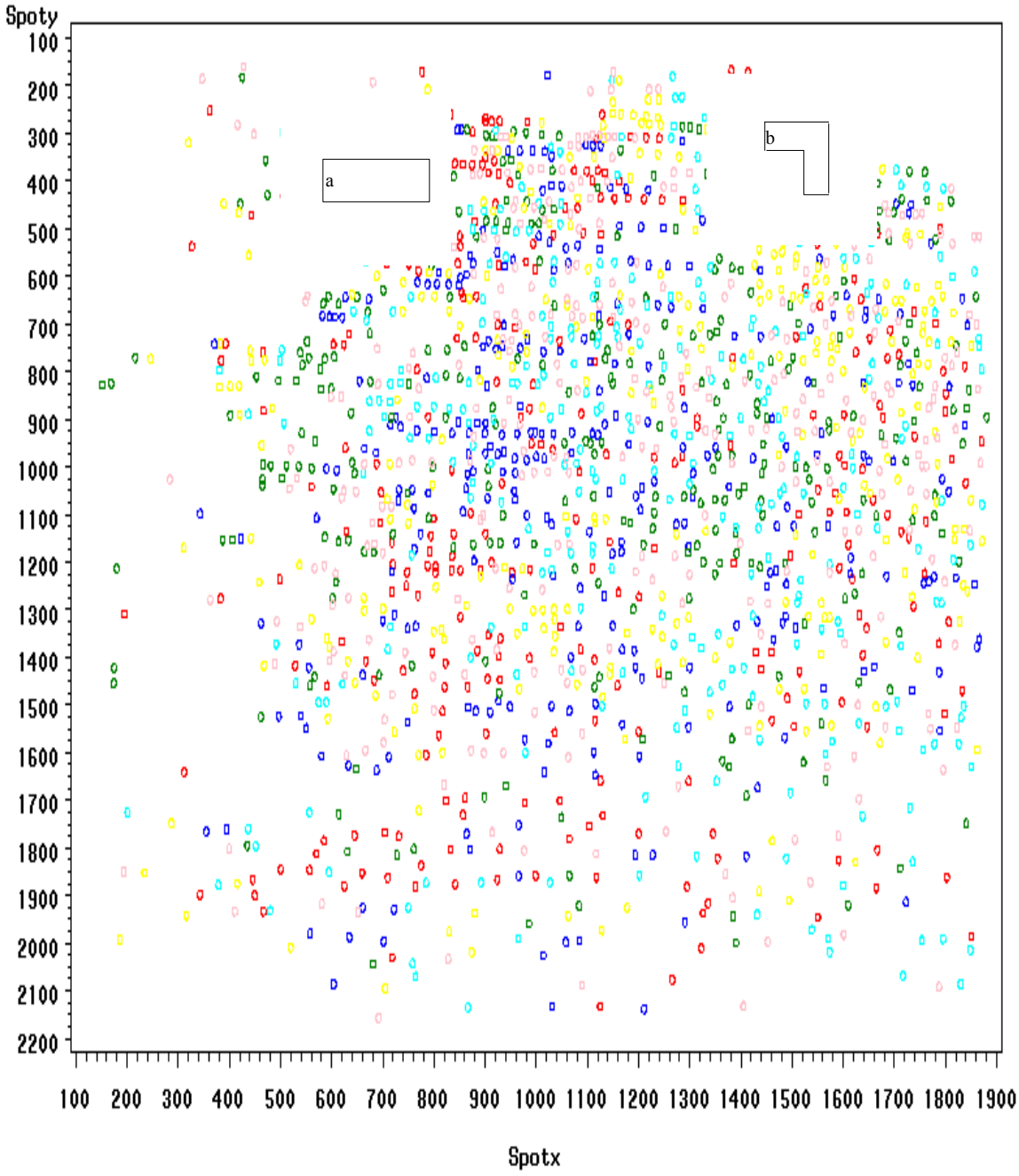
På de følgende tre figurer 12-14 kan resultatet af faktoranalyserne af henholdsvis de logtransformerede data, de originale og rangen af data ses:



Figur 12: Plot af 5 faktorers indflydelse på pletterne på gelen. Logtransformerede data. Pletterne er farvet efter, hvilken af de 5 grupper, givet ved faktorerne, de tilhører. Gruppe 1: rød, gruppe 2: blå, gruppe 3: grøn, gruppe 4: gul og gruppe 5: lyserød.



Figur 13: Plot af 5 faktorers indflydelse på pletterne på gelen. Originaldata. Pletterne er farvet efter, hvilken af de 5 faktorer de afhænger mest af. Faktor 1 har størst indflydelse på alle de røde pletter, de blå pletter tilhører gruppen givet ved faktor 2, de grønne hører til gruppen givet ved faktor 3, faktor 4 har størst indflydelse på alle de gule pletter, og faktor 5 har størst indflydelse på alle de lyserøde.



Figur 14: Plot af 6 faktorers indflydelse på pletterne på gelen. Rangen af data. Pletterne er farvet efter, hvilken af de 6 grupper, givet ved faktorerne, de tilhører. Gruppe 1: rød, gruppe 2: blå, gruppe 3: grøn, gruppe 4: gul, gruppe 5: lyseblå og gruppe 6: lyserød.

På figur 12-14 er det tydeligt at se, at det er forskellige grupperinger af variablene, der blev fundet for de tre datasæt. Det kan umiddelbart være svært at sige, hvilken af de tre opdelinger der er den bedste, men en måde at komme dette lidt nærmere på kan være at kigge på pletter, der ligger ved siden af hinanden, enten på linie eller over hinanden. Som forklaret tidligere vil der på gelen være flere forskellige modifikationer af det samme protein, hvilket for eksempel ville kunne ses netop som en række pletter på linie ved siden af hinanden eller over hinanden. Det vil være nærliggende at antage, at disse modifikationer vil reagere nogenlunde ens, og at det dermed vil være den samme faktor, der vil have størst indflydelse på dem.

For eksempel ligger den gruppe af pletter, markeret som kassen a, i de tre figurer, på en nogenlunde lige linie med enkelte pletter på en linie umiddelbart under. Dette kunne tyde på et protein, som har et forskelligt antal ladede sidegrupper koblet på, hvorved de enkelte modifikationer får forskelligt isoelektrisk punkt. De pletter, der ligger umiddelbart under, kunne være modifikationer af proteinet med forskellige uladede sidegrupper.

For både logtransformerede data og for originaldata er det, som det fremgår af figur 12 og 13, den samme faktor (markeret med henholdsvis blå og grøn), der har størst indflydelse på størstedelen af pletterne i denne gruppe, hvorimod den samme gruppe af pletter er under indflydelse af flere forskellige faktorer for rangen af data, se figur 14.

Et andet eksempel kan gives med den vandrette linie af pletter og den lodrette linie i forlængelse, markeret som kassen b, i de tre figurer. For de logtransformerede data ses det af figur 12, at hele denne gruppe er påvirket af den samme faktor (blå), mens den vandrette linie af pletter for rangen af data er påvirket af en faktor (lyseblå), og den lodrette er påvirket af flere forskellige. For de originale data er både den vandrette og den lodrette række af pletter påvirket af flere forskellige faktorer.

Der kan ligeledes findes grupper (rækker) af pletter, hvor de originale data eller rangen af data giver en bedre opdeling end de logtransformerede, men overordnet synes jeg, det virker, som om de logtransformerede data giver den bedste gruppering. Dette er også det mest rimelige, da en forudsætning for at udføre faktoranalyse er, at data er normalfordelte, og netop de logtransformerede data kan, som nævnt, med god tilnærmelse siges at følge en normalfordeling.

Faktoranalyse blev desuden udført med to til seks faktorer på både logtransformerede data, originaldata og rangen af data, og et plot af faktorernes indflydelse på pletterne blev lavet for alle tilfælde. Af disse plots kunne det også se ud til, at de logtransformerede data giver en bedre gruppering af variablene end både originaldata og rangen af data gør.

4.2.2.2 Faktoranalyse sammenholdt med ensidet variansanalyse

Det blev undersøgt, om der kunne siges at være en sammenhæng mellem grupperingen af variablene fundet ved faktoranalysen og resultatet af den ikke-parametriske ensidede variansanalyse, hvor fire grupper (kontrolgruppen, epilepsi under udvikling, mild grad af epilepsi og svær grad af epilepsi) blev sammenlignet. Af bilag 3 fremgår, hvilke pletter på gelen der blev fundet som de mest signifikante ved variansanalysen, og ud for hver enkelt plet er desuden anført, hvilken faktor der har den største indflydelse på den pågældende plet. Det viser sig for de logtransformerede data, at blandt de 100 mest signifikante pletter er det for 95 af disses vedkommende den samme faktor, der har størst indflydelse, og af de 150 mest signifikante pletter er det kun 19, der er placeret i andre grupper. I tabel 3 nedenfor er angivet de tilsvarende tal for de originale data og for rangen af data.

	Log-transformeret		Original		Rang	
Antal mest signifikante pletter	100	150	100	150	100	150
Antal pletter i samme gruppe	95	131	96	136	97	130
Antal i andre grupper	5	19	4	14	3	20

Tabel 3: Antallet af pletter blandt henholdsvis de 100 mest signifikante og de 150 mest signifikante pletter, fundet ved ikke-parametrisk ensidet variansanalyse (fire grupper), som placeres i den samme gruppe (givet ved den samme faktor) ved faktoranalyse.

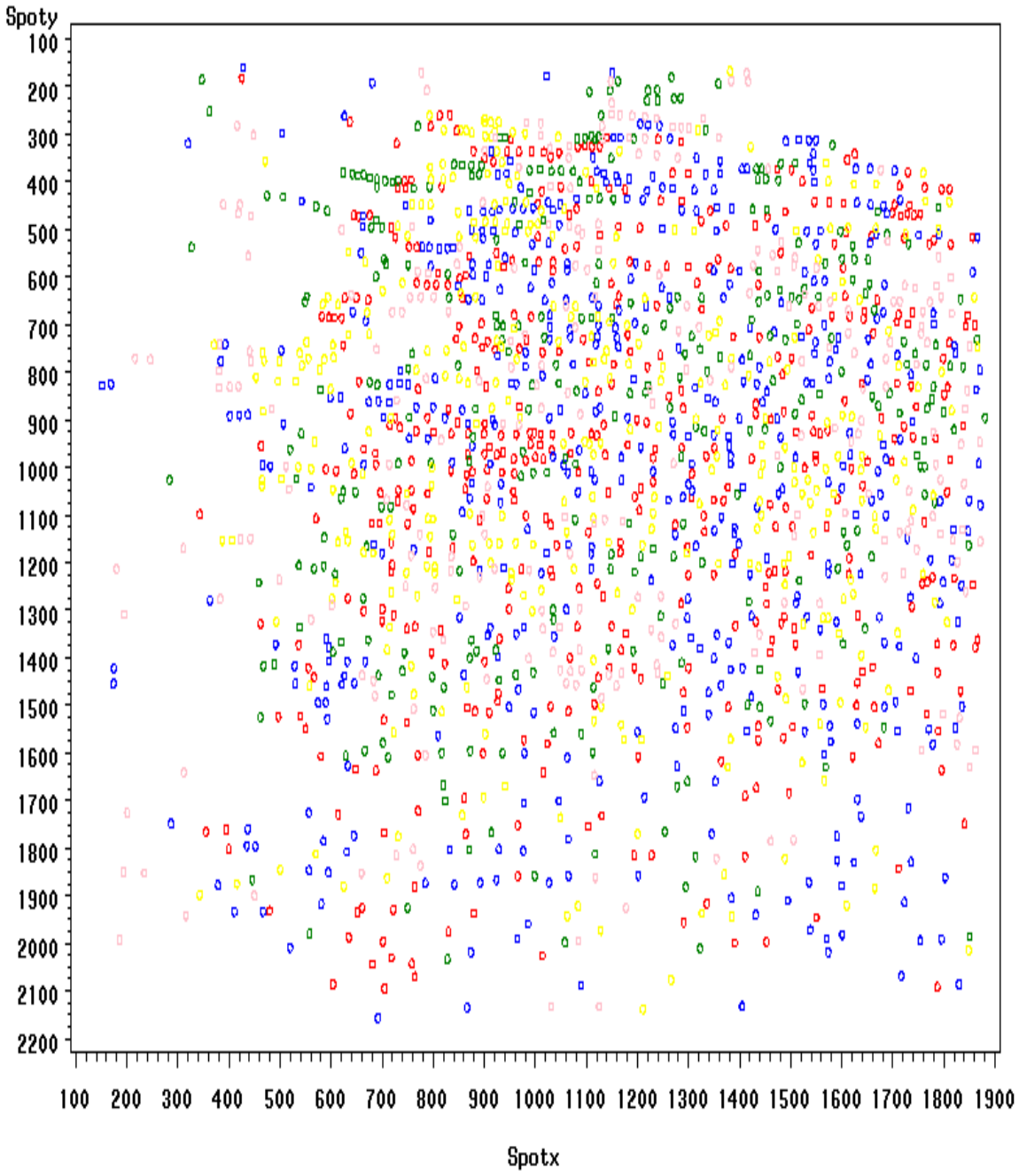
Som det fremgår af tabel 3, placerer faktoranalysen altså en meget stor del af de pletter, som den ensidede variansanalyse finder som de mest signifikante, i den samme gruppe. Disse pletter må derfor variere på samme måde og dermed have nogenlunde samme reaktionsmønster.

Det samme mønster gør sig gældende, når der ses på resultatet af den ensidede variansanalyse, hvor der kun skelnes mellem syg og rask, og når der ses på resultatet fra t-testen, se bilag 2 og 4. Også her placerer faktoranalysen en meget stor del af de mest signifikante pletter i den samme gruppe.

4.2.2.3 Rotation af principal faktorløsning

Faktoranalysen med efterfølgende Varimax rotation af de normerede egenvektorer, gennemgået i sidste afsnit, var baseret på kovariansmatricen for data frem for på korrelationsmatricen. Ved at basere analysen på kovariansmatricen opnås, at variablene indgår med en vægt svarende til deres varians. En variabel med stor varians vil altså indgå i analysen med større vægt end en variabel med lille varians. Hvis analysen i stedet blev baseret på korrelationsmatricen (det vil sige, på standardiserede data), ville alle variable derimod indgå i analysen med lige stor vægt, hvilket kunne tænkes at give et lidt anderledes resultat.

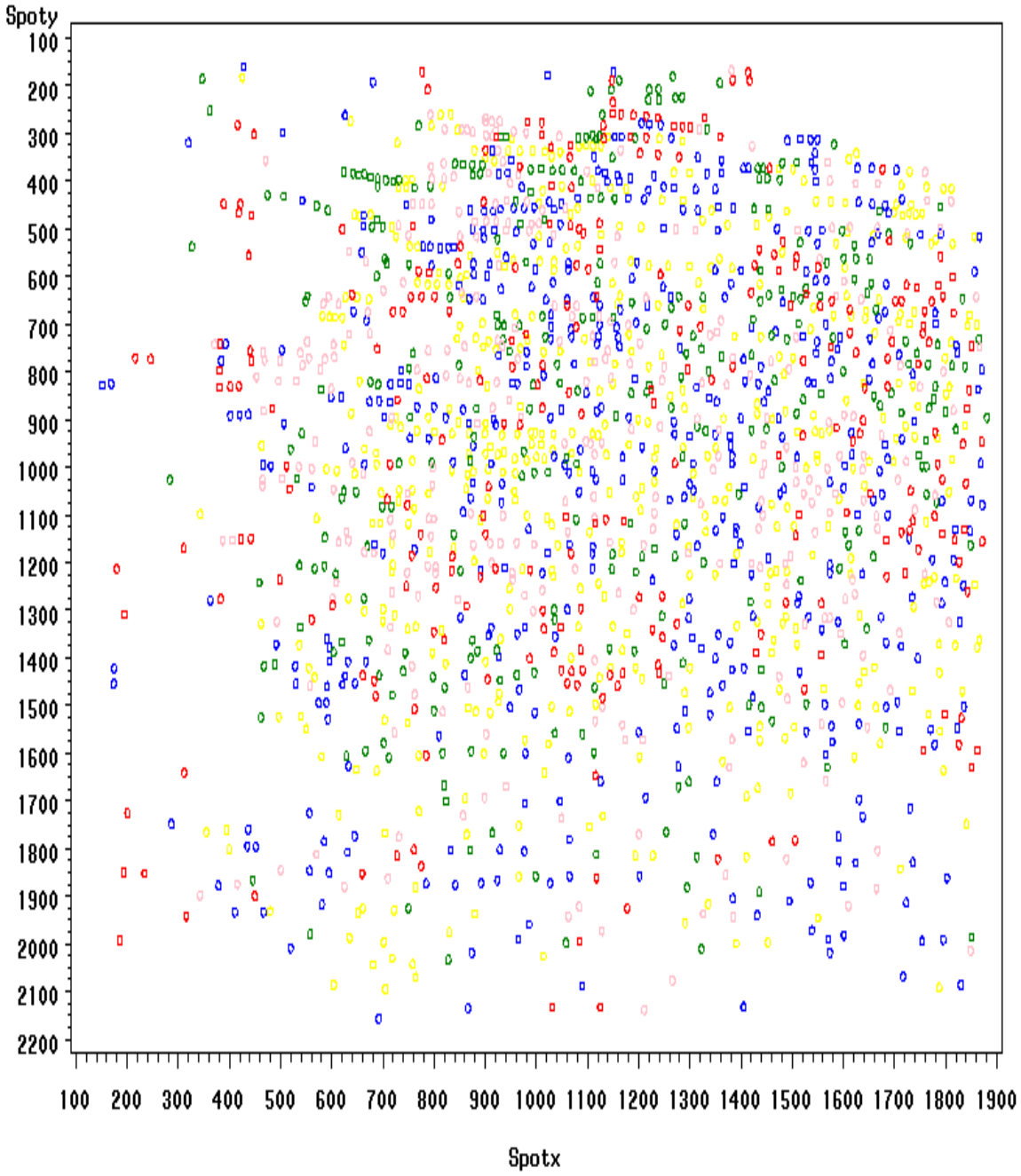
For at kunne sammenligne de to forskellige måder at gribe en faktoranalyse an på, blev den principale faktorløsning for de logtransformerede data Varimax roteret. I figur 15 ses et plot af gelen, hvor hver plet (ligesom på de tidligere plots) er farvet efter, hvilken faktor der har størst indflydelse på pletten:



Figur 15: Den principale faktorløsning er her blevet Varimax roteret, og pletterne på plottet er farvet efter, hvilken af de roterede faktorer de afhænger mest af. Faktor 1 har størst indflydelse på alle de røde pletter, faktor 2 har størst indflydelse på alle de blå pletter, de grønne pletter hører til gruppen givet ved faktor 3, gruppen givet ved faktor 4 omfatter alle de gule pletter og alle de lyserøde pletter tilhører gruppen givet ved faktor 5.

Umiddelbart ligner figur 15 ikke helt figur 12 (roterede egenvektorer for de logtransformerede data), der er visse lighedspunkter, men også en hel del forskelle. Men

hvis der ændres på farverne for nogle af faktorerne (det vil sige bytter om på rækkefølgen af faktorerne), fås et noget andet billede, se figur 16 nedenfor:



Figur 16: Samme faktorløsning som for figur 15 (den principale faktorløsning er blevet Varimax roteret), men rækkefølgen af faktorerne er blevet ændret en smule. Pletterne er stadig farvet efter, hvilken af de roterede faktorer de afhænger mest af. Men gruppen givet ved faktor 1 er nu alle de gule pletter, gruppen givet ved faktor 2 er igen alle de blå pletter, de grønne pletter hører ligesom før til gruppen givet ved faktor 3, gruppen givet ved faktor 4 omfatter nu alle de lyserøde pletter og alle de røde pletter tilhører nu gruppen givet ved faktor 5.

Sammenlignes figur 16 nu med figur 12 ses, at der er meget stor lighed mellem dem. Det er ikke alle pletter på figur 16, der nu har samme farve som de tilsvarende pletter på figur 13, men det er nogenlunde de samme grupper, de 1849 variable bliver inddelt i ved de to forskellige varianter af faktoranalysen. Der bliver altså fundet stort set det samme resultat ved de to analyser, der er blot byttet lidt om på rækkefølgen af faktorerne ved den sidst gennemførte analyse i forhold til den første. I tabel 4 nedenfor er en oversigt over, i hvilken rækkefølge de 5 faktorer blev fundet for henholdsvis den først gennemførte faktoranalyse, hvor de normerede egenvektorer blev roteret, og den anden gennemførte faktoranalyse, hvor den principale faktorløsning blev roteret. I kolonnen til venstre er angivet de 5 faktorer fra analysen af kovariansmatricen (rotation af normerede egenvektorer), og i kolonnen til højre er angivet, hvilken faktor fra analysen af korrelationsmatricen (rotation af principal faktorløsning) disse 5 grupper nogenlunde svarer til:

Analyse af kovariansmatrix	Analyse af korrelationsmatrix
1	5
2	2
3	3
4	1
5	4

Tabel 4: Oversigt over sammenhængen mellem de to varianter af faktoranalysen. Til venstre i tabellen er angivet grupperne (givet ved faktorer) fundet ved rotation af de normerede egenvektorer og til højre er angivet, hvilke grupper (fundet ved rotation af den principale faktorløsning) disse svarer nogenlunde til.

Grupperne givet ved faktor 1, 4 og 5 fra analysen af kovariansmatricen svarer altså nogenlunde til grupperne givet ved henholdsvis faktor 5, 1 og 4 fra analysen af korrelationsmatricen, mens grupperne givet ved faktor 2 og 3 fra analysen af kovariansmatricen nogenlunde svarer til de to grupper ligeledes givet ved faktor 2 og 3 fra analysen af korrelationsmatricen. Rækkefølgen af faktorerne er ændret en smule.

Ved rotationen af de normerede egenvektorer havde alle faktorer lige stor vægt, hvilket ikke er tilfældet for analysen af den principale faktorløsning, hvor faktorerne er vægtet efter de tilhørende egenværdier eller rettere kvadratet på egenværdien. Den første faktor

(hørende til den største egenværdi) har altså størst vægt i analysen. Det har vist sig at være denne faktor, der har størst indflydelse på en meget stor del af de mest signifikante pletter på gelen. 96 ud af de 100 mest signifikante pletter på gelen (fundet ved variansanalysen med fire grupper) tilhører gruppen givet ved denne faktor.

Det viser sig, når man undersøger kommunaliteterne, at de variable, der tilhører gruppen givet ved denne første faktor, forklarer ca. 28 % af faktorernes varians, det vil sige over en fjerdedel. Variablene, hørende til gruppen givet ved faktor 2, forklarer ligeledes mere end en fjerdedel af den samlede varians for faktorerne, nemlig ca. 26 %. I nedenstående tabel 5 er angivet den procentdel af faktorernes samlede varians, der kan forklares ved de enkelte grupper:

Roteret principal faktorløsning	Andel af faktorernes totale varians variablene i gruppen forklarer
1	28.04 %
2	26.02 %
3	15.75 %
4	16.21 %
5	13.97 %

Tabel 5: Oversigt over den procentdel af faktorernes totale varians, der bliver forklaret ved variablene i hver gruppe. Fundet på basis af analysen af korrelationsmatricen.

Som nævnt kan over en fjerdedel af faktorernes variation altså forklares ved variablene i gruppen givet ved faktor 1, som netop er den gruppe, størstedelen af de mest signifikante variable tilhører.

4.2.3 Clusteranalyse på variable

Som nævnt i afsnit 3.2.4 blev clusteranalyse i dette projekt benyttet til at gruppere variable og ikke observationer. Af de 1849 variable blev der for de logtransformerede data dannet 240 clusters ved hjælp af proceduren varclus, hvor det for hver cluster var gældende, at den anden egenværdi var mindre end 1. Den mindste af grupperne indeholdt kun 2 variable, mens den største indeholdt 23. Resultatet af clusteranalysen på de logtransformerede data kan ses i bilag 6. For originaldata og rangen af data blev ligeledes

dannet 240 clusters. Antallet af variable i hver cluster varierede dog for disse to datasæt henholdsvis fra 1 til 29 og fra 2 til 28. Det var altså ikke de samme grupper, der blev dannet for de tre datasæt.

På grund af det store antal variable, og dermed det store antal clusters der blev dannet, var det ikke muligt at få genereret scores for hver enkelt cluster. Det vil sige, det var ikke muligt at få genereret nye variable, der beskrev de enkelte clusters og ved hjælp af disse variable få forenklet eller reduceret datasættet.

Varclus proceduren kan, i tilfælde hvor antallet af variable eller antallet af dannede clusters er væsentligt mindre, generere et datasæt, der indeholder de ønskede score koefficienter. Dette datasæt indeholder udover en variabel for hver af de oprindelige variable yderligere tre variable, _NCL_, _TYPE_ og _NAME_. _NCL_ angiver antallet af clusters, _TYPE_ angiver, hvilken type statistisk resultat den enkelte observation indeholder, og _NAME_ indeholder navnet på clusteren (for eksempel clus1). For variablen _TYPE_ gælder, at den kan antage 12 forskellige værdier, hvoraf otte gentages, hver gang der dannes en ny cluster. For eksempel er den ene af disse otte værdier, _TYPE_ antager, 'members', og denne observation i det genererede datasæt vil indeholde oplysning om antallet af medlemmer i de dannede clusters. I det genererede datasæt vil der altså, for hver gang der er blevet dannet en cluster, være en observation, hvor variablen _TYPE_ har værdien 'members'. For tre af de otte _TYPE_-værdier, som gentages for hver ny dannet cluster, gælder desuden, at de gentages yderligere lige så mange gange, som der er clusters dannet indtil da. Den ene af disse tre er 'score', som for hver dannet cluster indeholder de ønskede scorekoefficienter. Når der for eksempel kun er dannet to clusters, vil der her være to observationer med _TYPE_-værdien 'score' – den ene indeholdende scorekoefficienterne for alle variablene hørende til cluster 1 og den anden indeholdende scorekoefficienterne for alle variablene i cluster 2. For alle de oprindelige variable vil der altså, i observationerne med _TYPE_-værdien 'score', være angivet en scorekoefficient – hvis variablen ikke er medlem af den pågældende cluster vil denne koefficient have værdien 0. Antallet af observationer, der genereres ved dannelse af en ny cluster, afhænger derfor af, hvor mange clusters der er blevet dannet hidtil. Det vil for eksempel sige, at når cluster nr. 2 dannes, vil der blive genereret $3 \cdot 2 + 5 = 11$ observationer, mens der når cluster nr. 30 dannes, vil blive genereret $3 \cdot 30 + 5 = 95$ observationer.

Ovenstående vil derfor resultere i et datasæt med ualmindelig mange observationer, som det vil tage utrolig lang tid at generere. Resultatet af clusteranalysen på variablene kan derfor bedst beskrives ved at sammenholde de fundne grupper med resultatet af dels de univariate metoder, dels faktoranalysen.

4.2.3.1 Sammenligning med resultat af univariate metoder og faktoranalyse

Ved sammenligningen mellem den ensidede variansanalyse og faktoranalysen viste det sig, at det var den samme faktor, der havde størst indflydelse på en meget stor del af de 0mest signifikante pletter på gelen. Det var derfor nærliggende at undersøge, om de mest signifikante pletter blev placeret i den eller de samme clusters. For de logtransformerede data kan i tabel 6 nedenfor ses en oversigt over, hvilke clusters de 100 mest signifikante pletter (fra variansanalysen med fire grupper, fra variansanalysen med to grupper og fra t-testen) fordeler sig i, hvor mange variable disse clusters indeholder, samt hvor mange af disse variable der er blandt de første 100 pletter:

Cluster nr.		Antal variable i cluster	Variansanalyse med fire grupper	Variansanalyse med to grupper	T-test
			Antal variable bl. 100 signifikante	Antal variable bl. 100 signifikante	Antal variable bl. 100 signifikante
Clusters fælles for de tre analyser	**2**	23	22	21	22
	20	18	11	2	4
	40	8	4	4	5
	97	6	1	3	3
	98	8	2	4	3
	102	16	8	12	13
	169	10	7	9	9
	176	20	17	19	20
	216	5	1	3	3
Clusters ej fælles	60	12	5		
	146	15	10		
	179	5		3	3

Tabel 6: Oversigt over hvilke clusters de 100 mest signifikante pletter (fra variansanalysen med fire grupper, variansanalysen med to grupper og t-testen) fordeler sig i, antallet af medlemmer i disse clusters

og det antal af de enkelte clusters medlemmer der er blandt de 100 mest signifikante pletter. Udover de i tabellen nævnte er der 23 andre clusters, der hver bidrager med en enkelt eller to variable blandt de 100 mest signifikante. Heraf er 4 fælles for alle tre analyser, 6 er fælles for variansanalysen med to grupper og t-testen, 6 er alene for variansanalysen med fire grupper, 5 er for variansanalysen med to grupper, og de sidste 2 er for t-testen. For de fem fremhævede clusters gælder, at den samme faktor (fra faktoranalysen hvor de normerede egenvektorer blev roteret) har størst indflydelse på samtlige medlemmer i gruppen.

Som det fremgår af tabel 6, er det relativt få clusters, de 100 mest signifikante pletter hører til, så ligesom faktoranalysen indikerede, ser det også her ud til, at der er nogle fælles egenskaber for nogle af de mest signifikante pletter.

For især et par af de i tabellen anførte clusters er det meget påfaldende, så stor en del af medlemmerne der er blandt de 100 mest signifikante pletter. For eksempel er der i cluster nr. 2 i alt 23 variable, og kun 1 (hhv. 2 og 1) variabel er ikke blandt de 100 mest signifikante pletter. Tilsvarende er der for cluster nr. 176 med 20 variable kun 3 (hhv. 1 og 0) variable, der ikke er blandt de 100 mest signifikante, og for cluster nr. 169 med 10 medlemmer er det ligeledes kun 3 (hhv. 1 og 1).

Af tabel 6 fremgår desuden, at 9 clusters er fundet blandt de 100 mest signifikante pletter for begge variansanalyser samt for t-testen, 2 clusters er specifikke for variansanalysen, hvor fire grupper blev sammenlignet, og 1 cluster er fundet kun for variansanalysen, hvor to grupper blev sammenlignet og for t-testen. Især den ene af de to clusters, der er specifikke for variansanalysen med fire grupper, er iøjnefaldende, idet hele 10 ud af 15 variable i denne cluster (nr. 146) er blandt de 100 mest signifikante.

Ovenstående kunne tyde på, at variablene i de 9 fælles clusters hovedsageligt adskiller syg (epilepsi) og rask (ej epilepsi), mens variablene i de 2 clusters (nr. 60 og nr. 146) for variansanalysen med fire grupper nok mere skelner mellem forskellige grader af epilepsi.

For de tre allerede nævnte grupper, nr. 2, nr. 169 og nr. 176, gælder desuden, at den samme faktor (fundet ved faktoranalysen hvor de normerede egenvektorer blev roteret) har størst indflydelse på alle medlemmer af gruppen. Det samme er tilfældet for de to clusters nr. 20 og nr. 102, samtlige variable i disse clusters tilhører den samme gruppe givet ved denne faktor. Disse fem clusters er fremhævet i tabel 6. Derudover gælder for de fire clusters nr. 40, nr. 216, nr. 146 og nr. 179, at kun et enkelt af hver gruppes medlemmer afviger, den samme faktor har altså størst indflydelse på alle medlemmer på nær et i hver af disse clusters.

For alle de i tabellen anførte clusters forholder det sig sådan, at det er den samme faktor (faktor 4, roterede normerede egenvektorer), der har størst indflydelse på flest medlemmer i hver gruppe, hvilket jo stemmer overens med, at det var denne faktor, der dominerede blandt de 100 mest signifikante pletter.

Dette kunne tyde på, at der er en sammenhæng mellem faktorer og clusters, og for at afgøre om det kunne være tilfældet, blev et udpluk af alle de fundne clusters undersøgt. I nedenstående tabel 7 fremgår, hvilke faktorer der har størst indflydelse på variablene i et udpluk af de 240 clusters:

Cluster nr.	Antal variable i cluster	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
6	12	1	2		9	
9	16		13			3
23	10			10		
30	7	5			2	
77	11	7	3			1
94	10	1			3	6
103	11					11
143	13			12	1	
208	6		5		1	

Tabel 7: Oversigt over hvilke faktorer der har størst indflydelse på variablene i ni forskellige clusters. For hver cluster er angivet antallet af variable i clusteren, samt hvor mange variable der tilhører hver af de fem grupper givet ved faktorerne.

Det fremgår af tabel 7, at der for hver cluster er én faktor, som dominerer billedet. Eftersom faktoranalysen inddelte variablene i fem grupper efter nogle fællestræk mellem dem, og clusteranalysen inddelte variablene i væsentligt flere grupper (240), virker det rimeligt at antage, at der er en vis sammenhæng mellem de dannede clusters og de fem større grupper. At én faktor vil være dominerende i hver cluster giver derfor i meget høj grad mening.

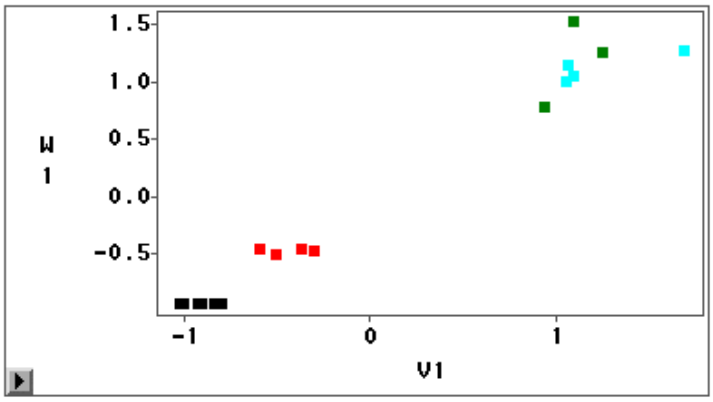
4.2.4 Resultat af kanonisk korrelationsanalyse

De to sæt variable, der ønskes benyttet til den kanoniske korrelationsanalyse, er de fire variable givet i tabel 2 (HAFD nummer, SE varighed, SE antal og SE dag) som det ene og proteinpletterne som det andet. Kanonisk korrelationsanalyse kan dog ikke anvendes direkte på det konkrete problem, da der i det ene sæt variable er 1849 variable og samtidig er der, som tidligere nævnt, kun 18 observationer at bygge analyserne på. Med så stort et antal variable og så lille et antal observationer bliver det muligt at konstruere en linearkombination af det ene sæt variable, som er lig med en linearkombination af det andet sæt variable, hvilket jo ikke siger ret meget. Antallet af variable, totalt for de to sæt, må derfor ikke overstige antallet af observationer.

I stedet for en direkte kanonisk korrelationsanalyse har jeg derfor valgt at benytte det lille sæt variable, der beskriver epilepsianfaldene, som det ene sæt variable og nogle af de principale komponenter som det andet sæt.

For de logtransformerede data blev den kanoniske korrelationsanalyse altså udført med de fire variable givet ved tabel 2 samt de fem første principale komponenter, som Horn's metode anbefalede. Det største antal sæt af kanoniske variable, der kan dannes, er lig antallet af variable i det mindste af de to grupper af variable, det vil sige, i dette tilfælde blev der dannet fire par kanoniske variable.

På figur 17 på næste side er den første kanoniske variabel for den ene gruppe variable (tabel 2) plottet som funktion af den første kanoniske variabel for den anden gruppe variable (principale komponenter):



Figur 17: Plot af det første par kanoniske variable for logtransformerede data. Sort: kontrolgruppe, rød: epilepsi under udvikling, blå: mild grad af epilepsi og grøn: svær grad af epilepsi. Kontrolgruppen og epilepsi under udvikling er tydeligt adskilt fra hinanden og fra de to syge grupper.

Ligesom ved principal komponentanalyse adskilles kontrolgruppen (markeret med sort i figur 17) og gruppen af rotter med epilepsi under udvikling (rød) pænt fra hinanden og fra gruppen af rotter med epilepsi i 'udbrud'. Grupperne med mild (blå) og svær (grøn) grad af epilepsi er ikke blevet separeret, som det også var tilfældet ved den principale komponentanalyse.

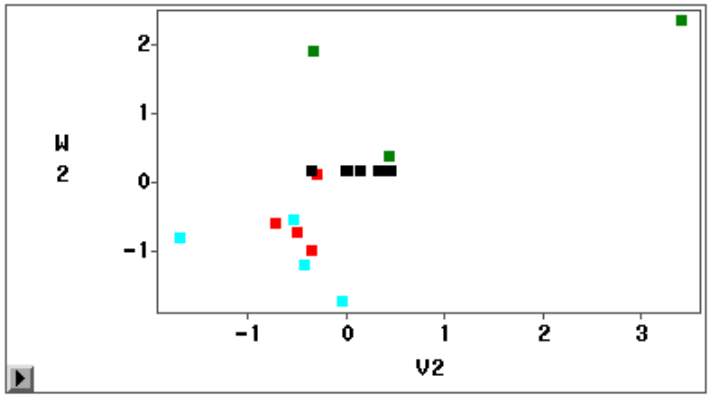
Denne kanoniske korrelationsanalyse er egentlig baseret på resultatet af den principale komponentanalyse, da det jo er de principale komponenter, der benyttes som det ene sæt variable, hvorved antallet af variable reduceres betydeligt. At resultaterne af de to analyser ligner hinanden, kan måske have noget med dette at gøre, men det kan lige såvel være en bekræftelse af, at de to grupper med epilepsi i 'udbrud' er svære at adskille ud fra de givne data. Resultatet af den kanoniske korrelationsanalyse bygger jo på en vægtning af de fem første principale komponenter, hvor opdelingen af observationerne fundet ved den principale komponentanalyse alene var baseret på de tre første principale komponenter og de egenskaber, disse tre repræsenterer.

I tabel 8 på næste side er anført egenværdierne samt den andel af den totale varians, som egenværdierne bidrager med:

Egenværdi	Andel af total varians
32.2498	0.9640
0.7704	0.0230
0.4193	0.0125
0.0150	0.0004

Tabel 8: Egenværdierne og den andel af den totale varians som de bidrager med.

Som det ses, tager det første par kanoniske variable højde for mere end 96 % af den totale variation i data. Da de to grupper af rotter med epilepsi i 'udbrud' ikke kan adskilles med det første par kanoniske variable, må variationen i data for disse to grupper være meget lille i forhold til den totale variation i data. Ved at medtage det andet par kanoniske variable vil der blive taget højde for yderligere 2.3 %, altså i alt knap 99 %, af den totale variation, så det kunne måske bevirke, at også de to grupper af rotter med epilepsi kunne separeres. For at undersøge om dette er tilfældet, er det andet par kanoniske variable blevet plottet i figur 18 nedenfor:



Figur 18: Plot af det andet par kanoniske variable for logtransformerede data. Sort: kontrolgruppe, rød: epilepsi under udvikling, blå: mild grad af epilepsi og grøn: svær grad af epilepsi. Mild grad af epilepsi og svær grad af epilepsi adskilles nu tydeligt fra hinanden.

Som det ses på figur 18, bliver de to grupper med epilepsi i 'udbrud' nu tydeligt adskilt. Det er altså muligt ved hjælp af de første to par kanoniske variable, baseret på de to sæt variable givet ved tabel 2 og ved de første fem principale komponenter, at skelne mellem de fire grupper: kontrolgruppen, epilepsi under udvikling, mild grad af epilepsi og svær grad af epilepsi.

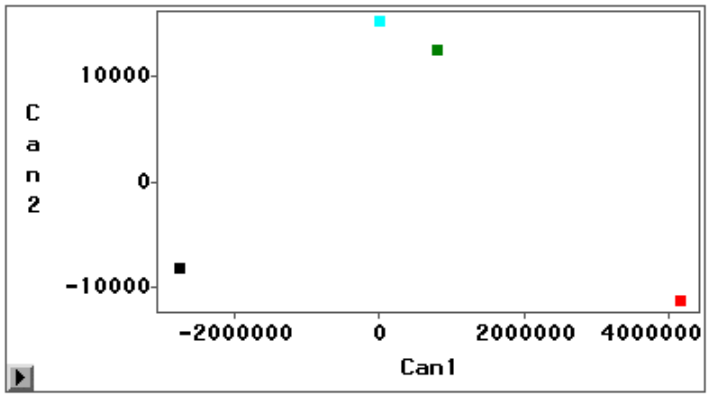
For originaldata og rangen af data var resultatet knap så overbevisende – der skulle tre par kanoniske variable til, før de fire grupper var helt adskilt for de originale data, mens det for rangen af data, selv ikke med alle fire par kanoniske variable, kunne lade sig gøre at skille de to grupper med epilepsi i 'udbrud' ad.

4.3 Resultater af superviserede metoder

4.3.1 Resultat af diskriminantanalyse samt stepvis og kanonisk diskriminantanalyse

Resultaterne for diskriminantanalyse, stepvis diskriminantanalyse og kanonisk diskriminantanalyse bliver gennemgået under et. Disse analyser hænger meget sammen, idet den stepvise diskriminantanalyse først udvalgte de relevante variable, som diskriminantanalysen dernæst bestemte diskriminantfunktionen ud fra. Den kanoniske diskriminantanalyse blev benyttet til at visualisere resultatet.

Som udgangspunkt blev den stepvise diskriminantanalyse gennemført uden begrænsning på, hvor mange variable der skulle medtages i diskriminantfunktionen. Som signifikansniveau for at tilføje variable til diskriminantfunktionen blev benyttet default værdien, der er 15 %. Den samme værdi blev benyttet som signifikansniveau for at beholde variable. For både logtransformerede, originaldata og for rangen af data blev 15 variable medtaget i diskriminantfunktionen. Resultatet kan ses i bilag 7. Der var ingen af de variable, som blev inkluderet i modellen, der siden blev fjernet igen. De 15 variable, som blev udvalgt af den stepvise diskriminantanalyse, blev benyttet til at udføre den kanoniske diskriminantanalyse, og resultatet for de logtransformerede data kan ses i figur 19 på næste side:

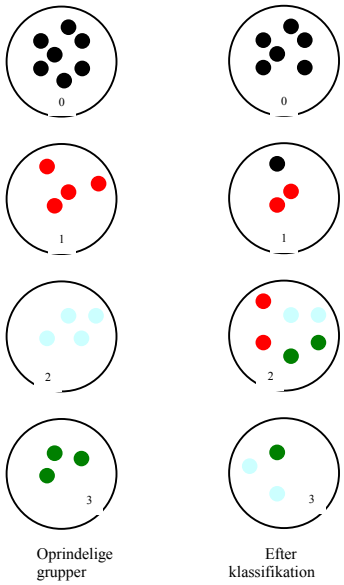


Figur 19: Resultat af kanonisk diskriminantanalyse for logtransformerede data. 15 variable udvalgt af stepvis diskriminantanalyse. Sort: kontrolgruppe, rød: epilepsi under udvikling, blå: mild grad af epilepsi, grøn: svær grad af epilepsi.

Som det meget tydeligt fremgår af figur 19, er denne opdeling ualmindelig god, de fire grupper ligger så spredt, at der overhovedet ingen tvivl er om, hvilken gruppe de enkelte observationer tilhører. Dette bekræftes også af krydsvalideringen foretaget med discrim proceduren, hvor alle observationer blev korrekt klassificeret ved hjælp af en diskriminantfunktion baseret på de fundne 15 variable. Krydsvalideringen med discrim tager, som nævnt, ikke højde for, at alle variable har medvirket ved udvælgelsen af de 15 variable, men foretages kun i forbindelse med bestemmelse af diskriminantfunktionen.

Sandsynligheden for at den fundne opdeling vil være nyttig til klassifikation af nye observationer er dog ikke særlig stor. Opdelingen er så god, at modellen nok er blevet for specifik for de opgivne data, og det er tvivlsomt, om den kan benyttes som et generelt redskab til klassifikation.

Den stepvise diskriminantanalyse blev derfor kombineret med krydsvalidering, det vil sige, krydsvalideringen blev ligeledes foretaget ved udvælgelsen af variable. Både den stepvise diskriminantanalyse samt bestemmelsen af diskriminantfunktionen blev altså foretaget i alt 18 gange, hvor en observation hver gang blev udeladt for til sidst at blive klassificeret ved hjælp af diskriminantfunktionen fundet ud fra de øvrige 17 observationer. Hver gang blev udvalgt 15 variable, og det blev på denne måde fundet, at hele 7 ud af 18 rotter blev misklassificeret, som det fremgår af figur 20, næste side:



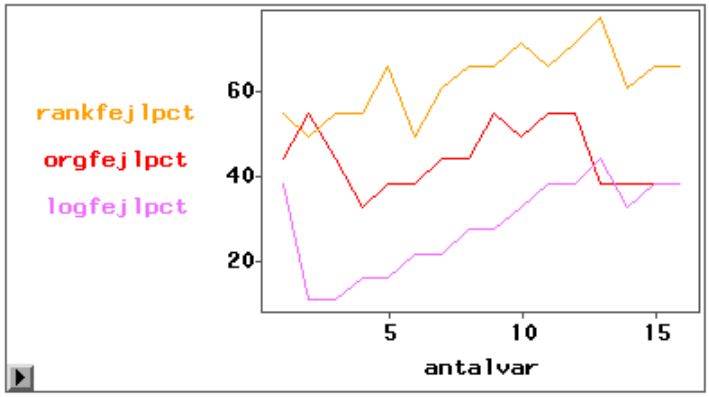
Figur 20: Til venstre ses de oprindelige fire grupper: den raske gruppe 0 (sort), epilepsi under udvikling 1 (rød), mild grad af epilepsi 2 (blå) og svær grad af epilepsi 3 (grøn). Til højre ses hvilke af de fire grupper, 0, 1, 2 eller 3, observationerne blev placeret i ved krydsvalidering. 15 variable var inkluderet i diskriminantfunktionen.

Det ses, at det hovedsageligt er rotter fra de tre syge grupper, der misklassificeres.

Et helt tilsvarende billede fås ved analyse af originaldata og rangen af data. De fire grupper adskilles utrolig godt, hvis der ingen krydsvalidering udføres, men foretages krydsvalideringen ved variabeludvælgelsen fås et ikke særligt overbevisende resultat. Det kunne altså tyde på, at en mere realistisk model kunne findes ved at reducere antallet af variable i diskriminantfunktionen.

For at finde det antal variable, og dermed den diskriminantfunktion, der er det mest hensigtsmæssige at benytte til diskrimination mellem de fire grupper, blev den stepvise diskriminantanalyse kombineret med krydsvalidering. Analysen blev foretaget et antal gange, hvor der første gang kun blev udvalgt en enkelt variabel til diskriminantfunktionen, og for hver efterfølgende gang blev inkluderet yderligere en variabel i funktionen. For hver gang analysen blev udført, blev fejlraten bestemt som antallet af misklassificerede observationer delt med antal observationer i alt. Den bedste diskriminantfunktion blev herefter fundet som den funktion, der gav den laveste fejlrate.

Nedenstående figur 21 viser fejlraten som funktion af antallet af variable i diskriminantfunktionen:



Figur 21: Fejlrate som funktion af antal variable for logtransformerede data (lilla), originaldata (rød) og rangen af data (gul).

Figur 21 viser, at de logtransformerede data gav færrest fejl uanset antallet af variable i diskriminantfunktionen – i et enkelt tilfælde, med 13 variable i funktionen, kom antallet af fejl for de originale data dog lige under. Forløbet af kurven for de logtransformerede data følger også fuldstændigt, hvad man måtte forvente. For et lille antal variable er fejlraten høj, den falder dernæst, som antallet af variable bliver større, og funktionen bliver bedre til at diskriminere, og når antallet af variable bliver for stort, stiger fejlraten igen, da modellen herefter er blevet overestimeret. Kurverne for de originale data og rangen af data følger i store træk et tilsvarende forløb, men er dog ikke så 'pæne'.

Det er desuden meget iøjnefaldende ved figur 21, at analysen af rangen af data giver fejlrater på 50 % og opefter. Det er altså ikke hensigtsmæssigt at benytte diskriminantanalyse på rangen til at separere grupperne.

For de logtransformerede data blev den mindste værdi af fejlraten fundet, som det fremgår af figur 21, når enten 2 eller 3 variable er inkluderet i diskriminantfunktionen. Fejlraten er her på kun 11.1 %. Det er altså mest hensigtsmæssigt kun at inkludere 2 eller 3 variable i diskriminantfunktionen. Medtages flere end 3 variable, vil modellen blive overestimeret, altså for specifik for de opgivne data. Jo flere variable der medtages, jo dårligere bliver modellen som diskriminator for nye observationer. Dette fremgår også af tabel 9 på næste side, hvor antallet af rigtigt klassificerede observationer, som antallet af

variable i modellen forøges, kan aflæses. Der er i tabellen kun medtaget op til 10 variable i modellen, da antallet af rigtigt placerede observationer blot fortsætter med at falde, som det også fremgår af figur 21:

Antal variable	Antal rigtigt klassificeret	Andel rigtigt klassificeret
1	11	0.6111
2	16	0.8889
3	16	0.8889
4	15	0.8333
5	15	0.8333
6	14	0.7778
7	14	0.7778
8	13	0.7222
9	13	0.7222
10	12	0.6667

Tabel 9: Antallet af rigtigt klassificerede observationer samt andelen af alle observationer der er klassificeret rigtigt

Som det ses, er der for både 2 og 3 variable i modellen 16 rotter, der bliver placeret i den rigtige gruppe, hvilket er det højeste antal rigtigt klassificerede.

Der er to metoder, der kan benyttes til at udvælge, hvilke to variable der skal inkluderes i modellen. Den ene er at benytte den stepvise diskriminantanalyse direkte og sætte som krav, at kun 2 variable skal medtages. Den anden er at se på de variable, der blev udvalgt, når den stepvise diskriminantanalyse blev kombineret med krydsvalidering. For hver af de 18 gange analysen blev udført, blev der jo udvalgt to variable, diskriminantfunktionen blev bestemt ud fra disse to, og den udeladte observation blev dernæst klassificeret ved hjælp af denne. Der er altså blevet fundet 18×2 variable.

De 2 variable, der blev udvalgt ved den første metode, var plet nr. 1567 (matchno 16110) og plet nr. 843 (matchno 894), og det viser sig ved den anden metode, at i 11 af de 18 tilfælde bliver begge disse variable ligeledes udvalgt, og i 6 af de resterende 7 tilfælde bliver den ene af disse to variable udvalgt. I tabel 10, næste side er en oversigt over,

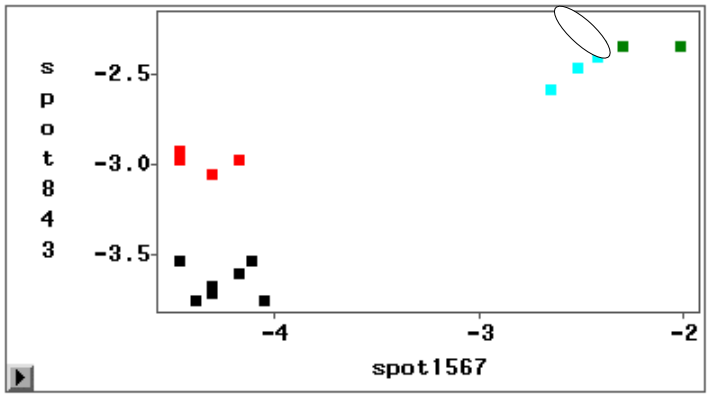
hvilke variable der blev udvalgt ved den anden metode, samt antallet af gange hver variabel blev inkluderet i modellen som den første og som den anden variabel:

Variabel	Antal gange som nr. 1	Antal gange som nr. 2
Spot635	1	0
Spot843	2	12
Spot1043	0	3
Spot1567	14	0
Spot1589	1	3

Tabel 10: Oversigt over hvilke variable der blev udvalgt ved stepvis diskriminantanalyse kombineret med krydsvalidering. Det er endvidere angivet, hvor mange gange hver variabel blev inkluderet i modellen som henholdsvis den første og den anden variabel.

Som tabel 10 viser, er det helt overvejende de to variable nr. 1567 og nr. 843, som også blev fundet ved direkte stepvis diskriminantanalyse, der dominerer billedet. Det mest oplagte må altså være at anvende disse to variable. De 2 variable er blevet identificeret af CPA som 2 forskellige modifikationsprodukter af proteinet 'glial fibrillary acidic protein delta', og det giver ifølge Stephen J. Fey, CPA, god mening at forbinde dette protein med epilepsi. Ifølge Stephen J. Fey er det desuden yderst rimeligt, at 2 modifikationer af et protein forklarer mere i relation til epilepsi (er bedre til at adskille grupperne) end blot et enkelt modifikationsprodukt af proteinet. Alle modifikationer af proteinerne spiller en eller anden rolle i organismen, så hver eneste modifikation er altså vigtig.

Diskriminationen mellem grupperne blev visualiseret ved at plotte de to fundne variable (nr. 1567 og nr. 843) mod hinanden. I figur 22 på næste side ses opsplitningen i de fire grupper:



Figur 22: Opsplitning i de fire grupper ved hjælp af diskriminantanalyse, de to variable (nr. 1567 og nr. 843) inkluderet i diskriminantfunktionen er plottet mod hinanden. Sort: kontrolgruppe, rød: under udvikling, blå: mild epilepsi, grøn: svær epilepsi.

Som ved principal komponentanalyse (og kanonisk korrelationsanalyse) bliver kontrolgruppen meget tydeligt adskilt fra epilepsi under udvikling, og begge disse grupper adskilles også klart fra grupperne med epilepsi. Der er en ikke helt så tydelig skelnen mellem de to grupper med henholdsvis mild og svær grad af epilepsi, men opdelingen mellem de to grupper er dog væsentligt tydeligere end den, som blev fundet ved principal komponentanalyse. Opdelingen illustreret på figur 22 underbygges desuden af Mahalanobis afstand mellem de fire grupper, se tabel 11 nedenfor:

Gruppe nr.	0	1	2	3
0	0	43.77285	282.89038	350.32496
1	43.77285	0	181.81675	242.83863
2	282.89038	181.81675	0	4.44174
3	350.32496	242.83863	4.44174	0

Tabel 11: Mahalanobis afstand mellem de fire grupper, kontrolgruppe (0), epilepsi under udvikling (1), mild grad af epilepsi (2) og svær grad af epilepsi (3).

Det fremgår her, at afstanden fra både kontrolgruppen (0) og gruppen med epilepsi under udvikling (1) til de to grupper med epilepsi (2 og 3) er stor, afstanden mellem kontrolgruppen og epilepsi under udvikling er ikke helt så stor, men stadig udtalt, mens afstanden mellem de to grupper med epilepsi er meget lille. Mahalanobis afstand bekræfter altså resultatet af både den principale komponentanalyse og kanonisk korrelationsanalyse, hvor det for det første viste sig, at gruppen med epilepsi under

udvikling mere ligner den raske gruppe end grupperne med epilepsi i 'udbrud' og for det andet, at det var vanskeligt at adskille de to sidstnævnte grupper.

Da proceduren discrim blev kørt med krydsvalidering for at finde diskriminantfunktionen, krydsvalideringen foregår altså alene ved bestemmelse af diskriminantfunktionen, blev to observationer fra de to grupper med epilepsi da også misklassificeret. En rotte med mild epilepsi blev klassificeret som hørende til gruppen med svær epilepsi, mens en rotte med svær epilepsi blev klassificeret som hørende til gruppen med mild epilepsi. De to misklassificerede observationer er markeret med en ellipse omkring i figur 22.

Diskriminantfunktionen med 2 variable for de logtransformerede data blev fundet til:

Variable	0	1	2	3
Konstant	-1040	-846.43051	-417.14624	-369.57378
Spot1567	-191.23045	-195.74928	-113.36955	-99.99364
Spot843	-347.36761	-283.02463	-228.59306	-221.90151

Tabel 12: Diskriminantfunktionen med 2 variable for de logtransformerede data.

Som tabel 12 viser, er koefficienterne for variablen spot843 meget lave for alle fire grupper, men der ses at være en stigning fra gruppe til gruppe med den laveste værdi for gruppe 0. Jo lavere værdi af variablen 843 en observation har, jo 'lavere' en gruppe vil den blive placeret i ved diskriminantanalysen. Det hænger selvfølgelig også sammen med værdien af den anden variabel spot1567, men lidt af det samme gør sig gældende her, så hvis en observation har meget lave værdier af både spot843 og spot1567 vil den blive placeret i gruppe 0, mens en observation med høje værdier af disse to variable vil blive placeret i gruppe 2 eller 3, hvilket ligeledes fremgår af figur 22. Som det ses, er forskellen mellem koefficienterne for de to grupper 2 og 3 ikke særlig stor, hvilket igen stemmer overens med, at disse to grupper er svære at adskille.

Når 3 variable blev medtaget i analysen (antallet af korrekt klassificerede var i dette tilfælde lige så højt som for kun 2 variable i modellen), var det, ved den direkte stepvise diskriminantanalyse, de samme 2 variable, nr. 1567 og nr. 843, samt variabel nr. 366 (matchno 385), der blev udvalgt. I tabel 13 på næste side ses en oversigt over, hvilke variable der blev udvalgt ved den stepvise diskriminantanalyse kombineret med

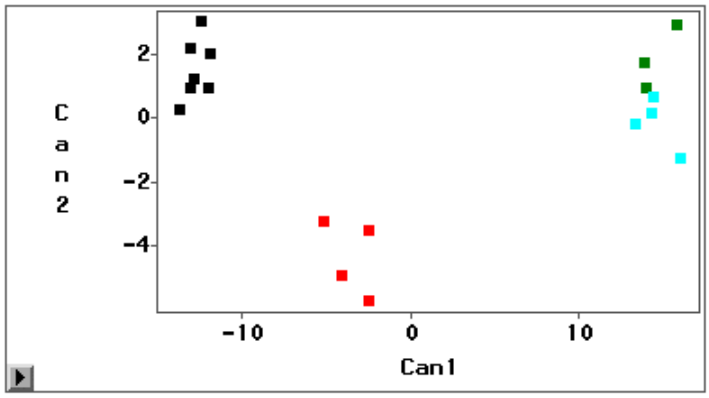
krydsvalidering (her blev udvalgt 18×3 variable), samt antallet af gange hver variabel blev udvalgt som den første, den anden og den tredje variabel:

Variabel	Antal gange som nr. 1	Antal gange som nr. 2	Antal gange som nr. 3
Spot355	0	0	1
Spot366	0	0	5
Spot445	0	0	2
Spot616	0	0	1
Spot619	0	0	1
Spot635	1	0	0
Spot728	0	0	1
Spot757	0	0	1
Spot843	2	12	0
Spot1043	0	3	0
Spot1482	0	0	1
Spot1567	14	0	0
Spot1589	1	3	1
Spot1750	0	0	4

Tabel 13: Oversigt over hvilke variable der blev udvalgt ved stepvis diskriminantanalyse kombineret med krydsvalidering. Det er endvidere angivet, hvor mange gange hver variabel blev inkluderet i modellen som henholdsvis den første, den anden og den tredje variabel.

Det er ud fra tabel 13 igen helt tydeligt, at variabel nr. 1567 og nr. 843 dominerer billedet, men som det fremgår, er der da heller ikke sket nogen forandringer for, hvilke variable der udvælges som henholdsvis den første og den anden variabel. Det er kun ved udvælgelsen af den tredje variabel, at nye variable optræder. Det ses, at den variabel, der optræder flest gange som den tredje udvalgte variabel, er nr. 366, som ligeledes blev udvalgt af den direkte stepvise diskriminantanalyse.

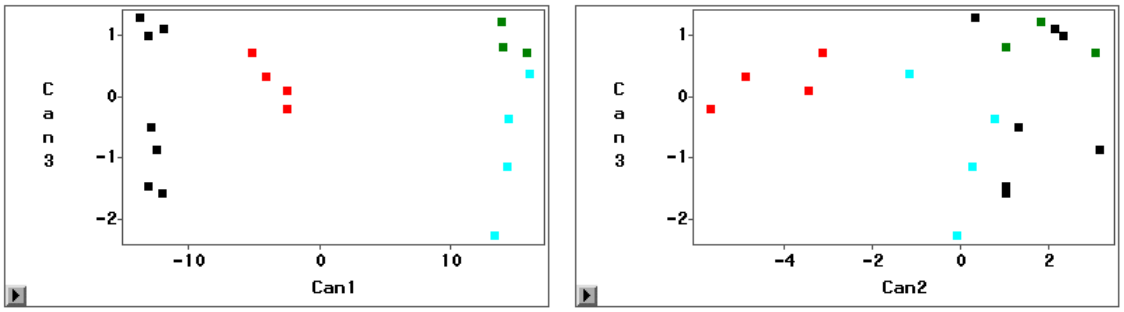
Ved hjælp af kanonisk diskriminantanalyse med de tre variable, nr. 1567, nr. 843 og nr. 366, i analysen blev opdelingen visualiseret, hvilket kan ses i figur 23 nedenfor, hvor observationerne er blevet projiceret ned i planet givet ved den første og den anden kanoniske variabel:



Figur 23: Opdelingen i de fire grupper fundet ved hjælp af kanonisk diskriminantanalyse. Tre variable, nr. 1567, nr. 843 og nr. 366, er medtaget i analysen. Sort: kontrolgruppe, rød: under udvikling, blå: mild epilepsi, grøn: svær epilepsi.

Som figur 23 viser, bliver de fire grupper adskilt rimelig godt. Som tidligere er kontrolgruppen og gruppen med epilepsi under udvikling adskilt fuldstændigt fra hinanden og fra de to grupper med epilepsi i 'udbrud'. Det er stadig svært at adskille de to grupper med epilepsi, men skellet mellem dem er dog blevet noget klarere.

Projiceres observationerne desuden ned i planet givet ved den første og den anden kanoniske variabel samt i planet givet ved den anden og den tredje kanoniske variabel, bliver opdelingen i de to grupper med epilepsi endnu tydeligere:



Figur 24: Opsplitningen i de fire grupper visualiseret ved hjælp af kanonisk diskriminantanalyse baseret på variablene nr. 1567, nr. 843 og nr. 366. Til venstre er observationerne projiceret ned i planet givet ved den første og den tredje kanoniske variabel, og til højre er observationerne projiceret ned i planet givet ved den anden og den tredje kanoniske variabel. De to grupper med epilepsi bliver nu (især på figuren til højre) tydeligt adskilt.

Som figur 24 viser, bliver de to grupper med epilepsi i 'udbrud' nu adskilt tydeligt fra hinanden. Dette fremgår især, når der ses på figuren til højre (planet udspændt af den anden og den tredje kanoniske variabel), men også figuren til venstre viser en klarere skelnen mellem de to grupper end tidligere fundet.

Dette bekræftes også af Mahalanobis afstand mellem grupperne, se tabel 14 nedenfor:

Gruppe nr.	0	1	2	3
0	0	119.31978	750.32424	747.37672
1	119.31978	0	348.96145	369.05002
2	750.32424	348.96145	0	7.31158
3	747.37672	369.05002	7.31158	0

Tabel 14: Mahalanobis afstand mellem de fire grupper, kontrolgruppe (0), epilepsi under udvikling (1), mild grad af epilepsi (2) og svær grad af epilepsi (3).

Afstanden fra de to grupper, 0 og 1, til de to grupper med epilepsi, 2 og 3, er, som det ses meget stor, især afstanden fra 0 til 2 og 3 er meget markant. Afstanden mellem de to grupper 0 og 1 er ligeledes betydelig, selvom den ikke er helt så stor som afstanden til 2 og 3, mens afstanden mellem de to grupper med epilepsi er lille. Denne afstand er dog ikke helt så lille som den, der blev fundet ved analysen, hvor kun 2 variable blev inkluderet i modellen (se tabel 11), hvilket altså stemmer overens med, at disse to grupper bliver bedre adskilt med den kanoniske diskriminantanalyse med 3 variable.

Når der udføres krydsvalidering alene ved bestemmelsen af diskriminantfunktionen og altså ikke ved variabeludvælgelsen, er der da heller ingen observationer, der misklassificeres.

Diskriminantfunktionen med 3 variable blev fundet til:

Variable	0	1	2	3
Konstant	-1759	-1274	-550.25795	-531.76131
Spot1567	-288.22718	-270.52277	-155.08742	-146.04291
Spot843	-849.80358	-670.34583	-444.68857	-460.43335
Spot366	272.61449	210.15488	117.25030	129.42392

Tabel 15: Diskriminantfunktionen med 3 variable for logtransformerede data.

Ligesom for diskriminantfunktionen med 2 variable viser diskriminantfunktionen med 3 variable, at lave værdier af både variabel nr. 843 og nr. 1567 vil pege i retning af en lavere grad af epilepsi. Koefficienterne for disse to variable stiger (stort set) her, som graden af epilepsi stiger. Koefficienterne for den sidste variabel, nr. 366, falder derimod, som graden af epilepsi stiger, hvilket vil sige, at jo højere en værdi for denne variabel en observation har (det er selvfølgelig afhængigt af værdierne for de to øvrige variable, som samtidig må have lave værdier), jo lavere en gruppe vil observationen blive placeret i.

For begge de to variable, nr. 1567 og nr. 843, gælder, at de er blandt de 50 mest signifikante pletter på gelen fundet ved ikke-parametrisk ensidet variansanalyse, hvor fire grupper blev sammenlignet. Den sidste variabel, nr. 366, blev ikke fundet blandt de signifikante variable ved variansanalysen. Det er dog også meget rimeligt, at en variabel ikke i sig selv udviser signifikans, men at den i kombination med andre variable derimod har en indflydelse. Ved de univariate metoder tages jo ikke højde for samspillet, korrelationerne, mellem variablene.

For de originale data og for rangen af data blev den laveste fejlrate fundet med henholdsvis 4 og 6 variable i diskriminantfunktionen. Fejlraterne var i disse to tilfælde 33.3 % og 50 %. De fire variable fundet for de originale data er nr. 275 (matchno 290), nr. 616 (matchno 647), nr. 1481 (matchno 12765) og nr. 1381 (matchno 10092). De seks variable fundet for rangen af data er nr. 843 (matchno 894), nr. 1589 (matchno 16610), nr. 616 (matchno 647), nr. 1754 (matchno 22867), nr. 1400 (matchno 10692) og nr. 728 (matchno 766).

4.3.2 Resultat af regressionsanalyse

Som nævnt i afsnit 3.3.5 blev proceduren logistic benyttet til at udføre stepvis logistisk regression på de givne data. Som signifikansniveau for at tilføje og beholde variable i modellen blev benyttet den samme værdi, som blev benyttet ved diskriminantanalysen, nemlig 15 %. Det viste sig dog meget hurtigt, at der var problemer forbundet med udførelsen af analysen. Den logistiske regression blev først forsøgt gennemført på hele datasættet med de fire grader af epilepsi: rask, under udvikling, mild grad og svær grad. For både logtransformerede data, originaldata og rangen af data resulterede det i en

advarsel fra SAS-systemet om, at fuldstændig separation af data var fundet, og at maximum likelihood estimatet derfor ikke eksisterede.

Fuldstændig separation af data vil sige, at der for alle observationer i datasættet gælder, at sandsynligheden for at anbringe en observation i den korrekte responsgruppe, under de givne omstændigheder, er 1 [18]. Problemet omkring fuldstændig separation af data i forbindelse med logistisk regression opstår for datasæt med få observationer. Når der er få observationer, er sandsynligheden langt større for, at de kan separeres totalt, end når der er mange observationer, for som antallet af observationer øges, vil sandsynligheden for at finde sæt af separerede datapunkter gå mod 0 [1].

Til dette projekt er der kun givet 18 observationer, så det er et meget lille datasæt at arbejde med, hvilket altså gør det muligt at finde en opdeling, der placerer alle observationerne i de korrekte grupper. Derudover er der meget få observationer i hver af de givne grupper, hvilket ligeledes kan give større usikkerhed på resultatet.

En anden ting, der spiller ind, er, at antallet af variable samtidig er så stort, at 'curse of dimensionality' bliver yderst aktuelt. De 18 observationer er fordelt i et 1849-dimensionelt rum, så tætheden af observationer er ikke ligefrem overvældende!

Det blev undersøgt, om en sammenlægning af de fire grupper i to større grupper, hvorved antallet af observationer i hver gruppe jo bliver højere, eventuelt kunne modvirke problemet med fuldstændig separation. Resultaterne af dels principal komponentanalyse, dels kanonisk korrelationsanalyse og dels diskriminantanalyse antyder, at den mest fornuftige sammenlægning af grupperne vil være at se på kontrolgruppen og epilepsi under udvikling under et og som den anden gruppe have epilepsi i 'udbrud'. Resultatet af denne undersøgelse blev dog, ligesom resultatet hvor alle fire grupper var medtaget, at fuldstændig separation af data blev fundet.

Herefter blev antallet af variable reduceret ligesom ved kanonisk korrelationsanalyse, hvor nogle af de principale komponenter blev benyttet som variable i stedet for de oprindelige. Den logistiske regression blev først forsøgt udført med alle de 18 principale komponenter som variable både med fire grupper og med to grupper, og endnu engang blev resultatet, at maximum likelihood estimatet ikke eksisterede. I det første tilfælde (fire grupper) fordi grupperne blev adskilt 'quasi-complete', det vil sige, næsten alle observationer havde en sandsynlighed på 1 for at blive placeret i den korrekte gruppe

[18], og i det andet tilfælde fordi grupperne blev separeret fuldstændigt. At adskillelsen er fuldstændig for analysen med to grupper og kun næsten fuldstændig for analysen med fire er rimeligt, da to grupper i princippet kan adskilles i en dimension, hvilket ikke er muligt med flere grupper. Der skal normalt flere variable til at adskille flere grupper end til at adskille to.

For at undersøge om det overhovedet var muligt at få lavet en logistisk regression, blev et mindre antal af de principale komponenter benyttet som variable, og for de logtransformerede data kunne det, med fire grupper, lade sig gøre, hvis kun 6 principale komponenter blev inkluderet i analysen. Heraf blev 4 (principal komponent nr. 1, nr. 2, nr. 3 og nr. 6) medtaget i den logistiske model. Med to grupper var det kun muligt at få opstillet en model, hvis en enkelt principal komponent (nr. 1) blev taget med i analysen. Dette underbygger også antagelsen om, at der skal flere variable til at adskille fire grupper, end der skal til at adskille kun to grupper.

I tabel 16 nedenfor ses en oversigt over, hvilke grupper observationerne med størst sandsynlighed blev placeret i, dels ved analysen med fire grupper og dels ved analysen med to grupper:

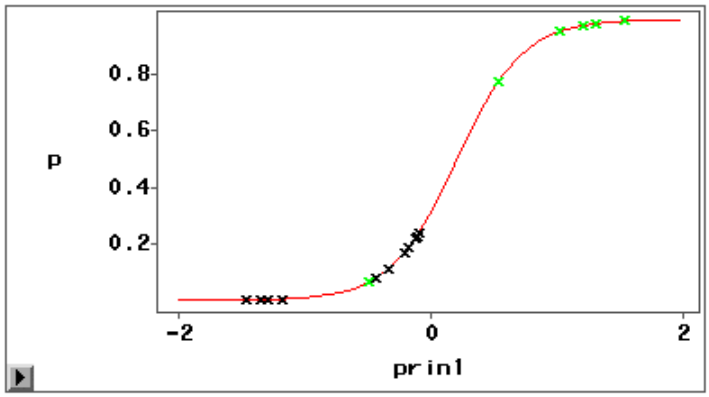
Placeret i / Opr. gruppe	Kontrol	Under udv.	Mild	Svær	'Rask'	Syg
Kontrol	7				11	
Under udv.		4				
Mild			2	2	1	6
Svær			1	2		

Tabel 16: Oversigt over hvilke grupper observationerne placeres i ved regressionsanalysen på de logtransformerede data, hvor nogle af de principale komponenter fungerer som uafhængige variable. I tabellen ses både resultatet fra analysen med fire grupper (venstre del af tabellen) og fra analysen med to grupper (højre del af tabellen). For analysen med to grupper er kontrolgruppen og gruppen med epilepsi under udvikling slået sammen til en 'rask' gruppe, og de to grupper med epilepsi i 'udbrud' er slået sammen til en syg gruppe.

Som det ses af tabellen, bliver langt størstedelen af observationerne klassificeret korrekt, for regressionen med to grupper bliver kun en enkelt observation misklassificeret, og dette sker med kun en enkelt variabel i analysen. For regressionen med fire grupper bliver tre observationer misklassificeret. To observationer fra gruppen med mild grad af epilepsi

vil med størst sandsynlighed blive placeret i gruppen med svær grad af epilepsi, og en observation fra gruppen med svær grad af epilepsi vil med størst sandsynlighed blive placeret i gruppen med mild grad. Endnu engang ses altså, at de to grupper med epilepsi i 'udbrud' er svære at adskille.

Resultatet for analysen med de to grupper kan illustreres ved en logistisk kurve, hvor de enkelte observationers placering desuden er markeret, se nedenstående figur 25:



Figur 25: Logistisk kurve for logtransformerede data. Regressionsanalysen blev foretaget med den første principale komponent som uafhængig variabel, og der blev i analysen skelnet mellem to grupper. De enkelte observationers sandsynlighed, p, for at blive placeret i den syge gruppe er markeret med kryds på kurven. Sandsynligheden for at blive placeret i den 'raske' gruppe kan findes som 1-p. Den 'raske' gruppe er markeret med sort og den syge med lysegrøn.

Ligesom det fremgår af tabel 16, viser også figur 25, at kun en enkelt af observationerne fra den syge gruppe (lysegrøn) har større sandsynlighed for at blive placeret i den 'raske' gruppe end i den syge.

For de originale data og rangen af data blev ligesom for de logtransformerede data fundet, at der til analysen med fire grupper skulle bruges flere variable end til analysen med to grupper. Med fire grupper blev benyttet henholdsvis 2 og 4 principale komponenter i den endelige model, med to grupper for rangen af data blev benyttet en enkelt principal komponent, som tilfældet også var for de logtransformerede data, men med to grupper for de originale data blev kun et intercept inkluderet i modellen og ingen principale komponenter overhovedet!

Konklusionen af ovenstående må være, at der er alt for få observationer i analysen og for at kunne lave en fornuftig logistisk regressionsanalyse af problemet, må der flere observationer til. Med flere observationer at basere analysen på, kunne man håbe på at få et mere realistisk resultat.

Den logistiske regression (eller forsøget på logistisk regression!) slår meget tydeligt fast, hvor forsigtig man bør være med fortolkningen af analyser på datasæt af denne type, hvor antallet af variable er så stort og antallet af observationer så smått. Det vil altid være muligt at adskille observationerne i nogle givne grupper, men om de kriterier, der benyttes til dette, kan overføres til at gælde for nye observationer, altså mere generelt, er ikke nødvendigvis sikkert.

4.3.2.1 Regression med Enterprise Miner

Den stepvise logistiske regressionsanalyse blev ligeledes forsøgt udført i Enterprise Miner med 'Regression Node', og også her blev signifikansniveauet for at tilføje og beholde variable i modellen sat til 15 %. Det fulde datasæt med fire grader af epilepsi blev her undersøgt.

Modsat analysen udført med proc logistic resulterede kørslen af regressionen i Enterprise Miner ikke i en advarsel om fuldstændig separation af data og dermed manglende maximum likelihood estimat. I Enterprise Miner blev derimod fundet, at en enkelt variabel (samt intercepts) blev inkluderet i modellen. Dette var tilfældet for både logtransformerede data, originaldata og rangen af data, men det var dog ikke den samme variabel, der blev udvalgt i de tre analyser. For de logtransformerede data var det plet nr. 843 (matchno 894), der blev udvalgt, hvilket er en af de 2 variable, der blev udvalgt ved diskriminantanalysen, og som er identificeret af CPA som et protein, der giver god mening i forbindelse med epilepsi. For originaldata og for rangen af data blev henholdsvis variabel nr. 1221 (matchno 1313) og variabel nr. 1573 (matchno 16307) udvalgt.

Som nævnt i afsnit 3.3.5 har den logistiske model følgende udseende:

$$\text{logit}(\Pr(Y \le i \mid x) = \alpha_i + \boldsymbol{\beta}'\mathbf{x}, \quad 1 \le i \le k$$

og i tabel 17 nedenfor er for de logtransformerede data anført maximum likelihood estimaterne for parametrene i modellen:

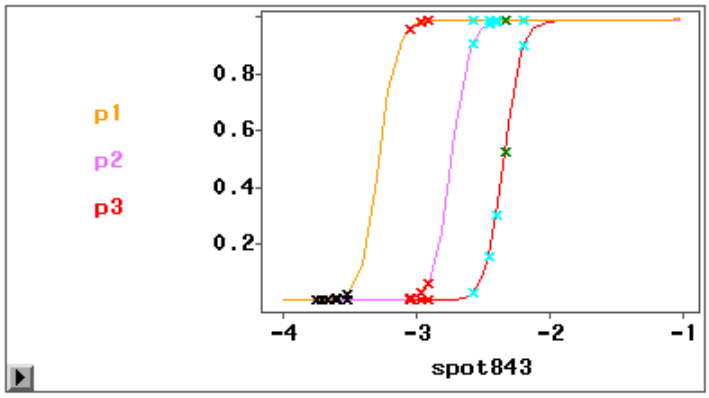
Parameter	Estimat
α_1: Intercept1	48.9307
α_2: Intercept2	40.7636
α_3: Intercept3	34.9328
β: Spot843	14.9540

Tabel 17: Maximum likelihood estimater for parametrene i modellen for de logtransformerede data.

Ud fra parameterestimaterne i tabel 16, kan de 3 logistiske kurver, der markerer overgangen fra en gruppe til den næste, tegnes. De tre kurver er fundet ved hjælp af støttepunkter for p_i, som beregnes med formlen nedenfor:

$$p_i = p(\text{spot843}) = \frac{1}{1+\exp(-(\alpha_i + \beta\,\text{spot843}))}$$

I figur 26 nedenfor er de 3 resulterende kurver for de logtransformerede data plottet:



Figur 26: Logistiske kurver fundet med Enterprise Miner's 'Regression Node' for logtransformerede data. Gul kurve adskiller kontrolgruppen (0) og epilepsi under udvikling (1), lilla kurve adskiller epilepsi under udvikling og mild grad af epilepsi (2), og rød kurve adskiller mild grad af epilepsi og svær grad af epilepsi (3).

I figur 26 er med krydser markeret sandsynlighederne for, at observationerne placeres i de forskellige grupper. Sandsynligheden for at blive placeret i gruppe 0 findes som $1-(p_1+p_2+p_3)$, sandsynligheden for at blive placeret i gruppe 1 er $p_1-(p_2+p_3)$,

sandsynligheden for at blive placeret i gruppe 2 er p_2-p_3, mens sandsynligheden for at blive placeret i gruppe 3 findes som p_3. Kontrolgruppens medlemmer er markeret med sort, under udvikling med rød, mild epilepsi med blå og svær grad af epilepsi med grøn. Det ses, at alle rotter fra kontrolgruppen har en værdi for variablen spot843 på mellem -4 og ca. -3.5, og som de 3 logistiske kurver viser, er det i dette interval for variablen gruppe 0 (kontrolgruppen), der dominerer. Observationerne placeres i denne gruppe med en sandsynlighed på næsten 1. Fra omkring -3.3 til omkring -2.8 er det gruppe 1 (under udvikling), der dominerer billedet, og de fire observationer fra denne gruppe vil da også med størst sandsynlighed blive placeret her. Fra ca. -2.8 til ca. -2.4 er sandsynligheden størst for at blive placeret i gruppe 2 (mild grad af epilepsi). Dette interval er, som det fremgår af figur 26, noget mindre end det interval, der dækker over gruppe 1, og som det ses, er der for to ud af de fire observationer i denne gruppe en betragtelig sandsynlighed for, at observationen placeres i gruppe 3 i stedet for i gruppe 2, og for en enkelt af de fire er sandsynligheden for at blive placeret i gruppe 3 langt større end sandsynligheden for at blive placeret i den korrekte gruppe. Fra omkring -2.4 og opefter er sandsynligheden for at blive placeret i gruppe 3 (svær grad af epilepsi) størst. De tre rotter, der tilhører denne gruppe, har alle samme værdi af variablen spot843, hvilket er grunden til, at der kun er en enkelt grøn observation markeret i figuren. For alle tre observationer i denne gruppe gælder altså, at sandsynligheden for at blive korrekt placeret i gruppe 3 kun er en anelse større end sandsynligheden for at blive placeret i gruppe 2. Ovenstående er udelukkende baseret på værdien af variabel spot843.

Det må understreges, at disse resultater skal tages med et forbehold, da en tilsvarende kørsel af en logistisk regression i det almindelige SASstat, som nævnt, ikke kunne gennemføres. Om der i Enterprise Miner er blevet taget højde for de problemer, der opstår ved kørslen i SASstat, eller om de tidligere nævnte advarsler blot ignoreres, er ikke til at sige, men det mest sandsynlige er nok, at Enterprise Miner stopper, når den når til det punkt, hvor fuldstændig separation opdages, og det resultat, der gives, er fra iterationen før. Man kunne forestille sig, at Enterprise Miner blot undlader at informere brugeren om problemet.

Resultatet fundet ved regressionsanalysen i Enterprise Miner stemmer dog meget godt overens med resultatet fra den stepvise diskriminantanalyse. Ved diskriminantanalysen blev fundet, at der skulle 2 til 3 variable til at adskille de fire grupper, og den ene af disse var plet nr. 843, som netop er den variabel, der blev udvalgt ved regressionsanalysen. For denne variabel forholdt det sig ved diskriminantanalysen sådan, at jo lavere en værdi den

havde, jo 'lavere' en gruppe ville observationen blive placeret i. Placeringen af observationen afhang selvfølgelig også af værdien af den anden variabel 1567 og den tredje variabel 366. Ved regressionsanalysen blev tilsvarende fundet, at for lave værdier af variabel nr. 843 blev observationen med størst sandsynlighed placeret i gruppe 0, og jo højere værdien af denne variabel var, jo højere en gruppe blev observationen placeret i.

For diskriminantanalysen med 2 variable i modellen blev ved krydsvalidering desuden fundet, at en observation fra gruppen med mild grad af epilepsi (2) blev klassificeret som tilhørende gruppen med svær grad af epilepsi (3), og ved regressionsanalysen blev netop fundet, at en observation fra gruppe 2 (den samme som blev misklassificeret ved diskriminantanalysen) med langt større sandsynlighed ville blive placeret i gruppe 3.

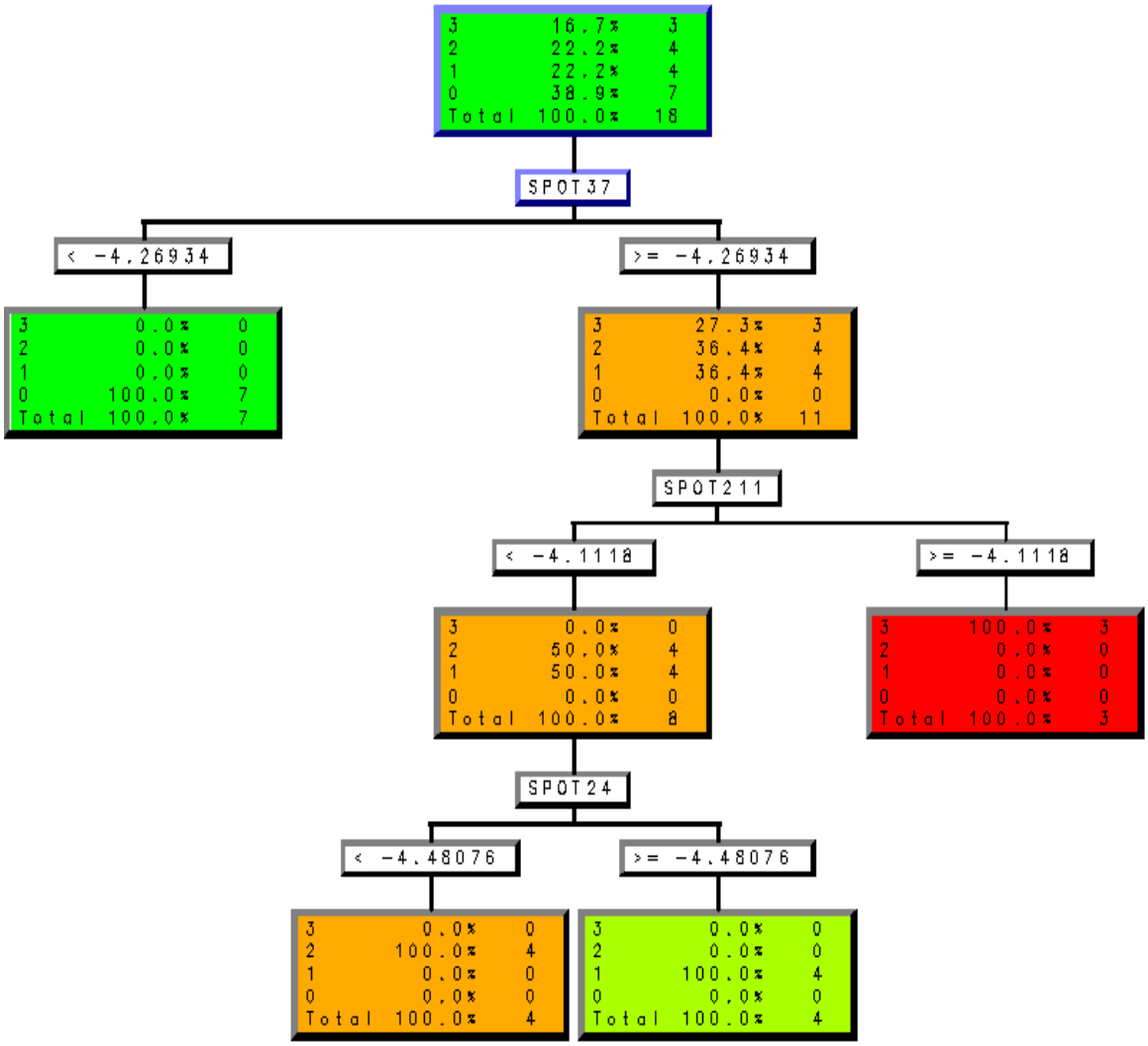
En observation fra gruppen med svær grad af epilepsi blev ved diskriminantanalysen med krydsvalidering ligeledes klassificeret som hørende til gruppen med mild grad af epilepsi, mens resultatet fra regressionsanalysen blot fortæller, at for de tre observationer i denne gruppe er sandsynligheden for at blive placeret i den korrekte gruppe kun lidt større, end den er for at blive placeret i gruppen med mild grad af epilepsi.

Resultaterne fra de to analyser stemmer altså nogenlunde overens, men jeg vil dog endnu engang pointere, at resultatet af regressionsanalysen må tages med forbehold.

4.3.3 Klassifikationstræer med Enterprise Miner

Med SAS Enterprise Miner's 'Tree Node' kan klassifikationstræer, som er beskrevet i [3, 4] og gennemgået i afsnit 3.3.6, tilnærmes. I 'Tree Node' er det blandt andet muligt at specificere, at træet skal være binært, det vil sige, at der fra hver knude kun udgår to grene, at hvert split findes ved Gini kriteriet, og at et testdatasæt benyttes til validering af træet, hvilket er nogle af de ting, der karakteriserer de omtalte klassifikationstræer. Da Enterprise Miner's 'Tree Node' er beregnet til store datasæt, vil testdatasættet normalt være en mindre del af det datasæt, der arbejdes på. I mit tilfælde er datasættet desværre så småt, at det ikke er muligt at benytte en del af det til validering af træet. Derudover er det ikke muligt i 'Tree Node' at udføre krydsvalidering, så træet bliver altså ikke beskåret, som det ellers er meningen. Der er altså elementer ved denne analyse, der gør, at der kan være tvivl om resultatet.

For både logtransformerede data, originaldata og rangen af data fås præcis det samme resultat. I alle tre tilfælde blev fundet, at de fire grupper blev adskilt fuldstændigt med kun 3 forgreninger. De 3 forgreninger blev foretaget ved variabel nr. 37 (matchno 40), nr. 211 (matchno 224) og nr. 24 (matchno 27). I nedenstående figur 27 ses resultatet for de logtransformerede data – det eneste, der adskiller denne figur fra tilsvarende plots af resultatet for originaldata og rangen af data, er, at split-værdierne for hver variabel er forskellige, da værdierne for de enkelte variable jo er forskellige fra datasæt til datasæt.



Figur 27: Klassifikationstræ for logtransformerede data. Mørkegrøn er kontrolgruppen, lysegrøn er epilepsi under udvikling, orange er mild grad af epilepsi, og rød er svær grad af epilepsi. Alle observationer er klassificeret korrekt. I hver kasse er angivet, hvor mange observationer fra hver gruppe der er i den pågældende knude. Over hver kasse er angivet beslutningsreglen for det givne split.

Som figur 27 viser, sker første forgrening ved variabel spot37, og her skilles hele kontrolgruppen (mørkegrøn) fra de tre stimulerede grupper. Gini indekset i startknuden (alle observationer stadig samlet) er $1-(0.167^2+0.222^2+0.222^2+0.389^2)=0.722$ og det split, der blev valgt som det første, var det split, der medførte det laveste Gini indeks i de følgende knuder. I knuden hvor kontrolgruppen er blevet skilt fra de øvrige grupper, er Gini indekset $1-(0+0+0+1^2)=0$, altså det lavest mulige Gini indeks. I knuden med de øvrige tre grupper er Gini indekset $1-(0.273^2+0.364^2+0.364^2+0)=0.660$, der ses at være lavere end Gini indekset for startknuden, hvilket jo var meningen. Ved næste forgrening, spot211, bliver gruppen med svær grad af epilepsi (rød) adskilt fra de to øvrige grupper, og ved den sidste forgrening, spot24, separeres gruppen med epilepsi under udvikling (lysegrøn) fra gruppen med mild grad af epilepsi (orange). Ved hver knude bliver værdien af Gini indekset mindre, for til sidst at være 0 i alle endeknuderne. Ingen af observationerne misklassificeres altså.

Da det ikke er muligt at teste træet (beskære det) enten ved hjælp af et testdatasæt eller med krydsvalidering, er sandsynligheden, for at dette træ vil være godt til at klassificere nye observationer, ikke særlig stor. Træet er utrolig godt til at skille de opgivne data i de fire ønskede grupper, men som nævnt under afsnittet om regressionsanalyse, vil det altid være muligt at finde kriterier, der medfører en given opdeling, når datasættet indeholder så mange variable og så få observationer. Derudover er Enterprise Miner's 'Tree Node' som nævnt beregnet til store datasæt, altså til datasæt med mange observationer (normalt flere tusinde), hvilket jo ikke er opfyldt i dette tilfælde, så der må antages at være stor usikkerhed på resultatet.

Sammenlignes resultatet fra klassifikationstræet med resultatet af den ikke-parametriske ensidede variansanalyse, hvor fire grupper blev sammenlignet, viser det sig, at alle de 3 variable fundet i klassifikationstræet er blandt de signifikante variable fundet ved variansanalysen. Plet nr. 37 og plet nr. 24 er oven i købet blandt de 50 mest signifikante pletter på gelen.

Sammenholdes klassifikationstræet derimod med den ikke-parametriske variansanalyse, hvor kontrolgruppen blev sammenlignet med den stimulerede gruppe under et, ses det, at kun plet nr. 37 er blandt de signifikante variable. Dette giver i høj grad mening, da det er forgreningen ved denne variabel, der skiller kontrolgruppen fra de resterende, og denne adskillelse sker i første trin. Ved de øvrige forgreninger bliver de tre grupper med

forskellige grader af epilepsi skilt fra hinanden, og det er derfor rimeligt, at de er signifikante ved variansanalysen med fire grupper, men ikke ved variansanalysen med to grupper.

4.4 Sammenligning og diskussion af resultater af ikke-superviserede og superviserede metoder

I det følgende afsnit vil resultaterne af de ikke-superviserede og de superviserede metoder blive sammenlignet og diskuteret. Dette er gjort ved at gennemgå de fundne resultater af de superviserede metoder et efter et og sammenholde disse med resultaterne af de ikke-superviserede metoder.

Først ses på resultatet af diskriminantanalysen med 2 variable i modellen. De 2 variable, der blev fundet af den stepvise diskriminantanalyse til at indgå i diskriminantfunktionen, var nr. 843 (matchno 894) og nr. 1567 (matchno 16110) (begge disse er, som nævnt, blandt de 50 mest signifikante variable fundet ved variansanalysen), og det viser sig, at den faktor, der har størst indflydelse på variabel nr. 843, er den samme som den, der har størst indflydelse på nr. 1567. Derudover tilhører de begge cluster nr. 2, som er en af de clusters, der har næsten alle sine medlemmer blandt de 100 mest signifikante pletter på gelen fundet ved den ikke-parametriske variansanalyse. Kendetegnende for cluster nr. 2 var desuden, at den samme faktor havde størst indflydelse på samtlige variable i clusteren. De 2 variable varierer altså på samme måde, hvilket også er meget rimeligt, da de jo som tidligere nævnt er forskellige modifikationsprodukter af det samme protein.

Når endnu en variabel blev medtaget i diskriminantanalysen, blev nr. 366 (matchno 385) udvalgt af den stepvise diskriminantanalyse. Denne variabel var ikke blandt de signifikante variable fundet ved variansanalysen, og det viser sig, at den faktor, der har størst indflydelse på denne variabel, er en anden end den, der har størst indflydelse på de to først udvalgte variable. Variabel nr. 366 tilhører desuden en anden cluster end de to øvrige variable, nemlig cluster nr. 230. Det virker rimeligt, at de udvalgte variable hører til forskellige grupper (repræsenteret ved faktorer eller clusters), sådan at én variabel tager højde for variationen i én retning, mens en anden tager højde for variationen i en anden retning. Herved bliver der taget hensyn til flere forskellige egenskaber.

Ved regressionsanalysen i Enterprise Miner blev fundet, at en enkelt variabel blev anvendt til at skelne mellem de fire grupper, kontrolgruppen, under udvikling, mild og svær grad af epilepsi. For de logtransformerede data var denne variabel den samme som den ene af de to fundet ved diskriminantanalysen, nr. 843. For de originale data og rangen af data var det henholdsvis variabel nr. 1221 (matchno 1313) og nr. 1573 (matchno 16307), der blev benyttet. For disse tre variable gælder, at de alle ligger blandt de 20 mest signifikante pletter på gelen, og at den samme faktor har størst indflydelse på dem. Herudover tilhører de alle tre den samme cluster, både når der ses på clusters fundet ud fra de logtransformerede data, clusters fundet ud fra originale data og clusters fundet ud fra rangen af data. For de logtransformerede datas vedkommende tilhører de tre variable den allerede omtalte cluster nr. 2. Resultatet må dog, som tidligere nævnt på grund af det lave antal observationer, tages med forbehold.

Til sidst ses på resultatet fra klassifikationstræet. Her blev det samme resultat fundet for de tre datasæt, og de variable, der blev benyttet i træet, var nr. 37 (matchno 40), nr. 211 (matchno 224) og nr. 24 (matchno 27). Den samme faktor har størst indflydelse på plet nr. 37 og plet nr. 24, mens det er en anden faktor, der influerer mest på plet nr. 211.

Det viser sig endvidere, at de tre variable benyttet i klassifikationstræet tilhører tre forskellige clusters. Nr. 37 tilhører cluster nr. 176, der ligesom cluster nr. 2 er en af de clusters, hvor næsten alle medlemmer er blandt de 100 mest signifikante variable (fundet ved variansanalysen), og hvor den samme faktor har størst indflydelse på alle variable i clusteren. Nr. 211 tilhører cluster nr. 5, og variabel nr. 24 tilhører cluster nr. 146. Nr. 146 er en af de clusters, som var specifikke for variansanalysen, hvor fire grupper blev sammenlignet. Hele 10 ud af 15 medlemmer i denne cluster blev fundet blandt de 100 mest signifikante pletter, mens ingen af clusterens medlemmer altså blev fundet blandt de 100 mest signifikante pletter ved variansanalysen med to grupper og ved t-testen. Dette kunne tyde på, at variablene i cluster nr. 146 skelner mellem forskellige grader af epilepsi og ikke kun mellem syg og rask. At det er en variabel fra denne cluster, der skiller gruppen med epilepsi under udvikling fra gruppen med mild grad af epilepsi i klassifikationstræet, virker altså rimeligt. Der må dog også for denne analyse antages at være stor usikkerhed på resultatet.

De anvendte metoder har vist sig at være mere eller mindre velegnede til denne type af data, hvor der er mange variable og få observationer. I tabel 18 på næste side ses en kort opsummering af anvendeligheden af hver enkelt metode i forhold til dette projekts data:

Metode	Velegnet	Begrundelse
Ikke-superviserede metoder		
Principal komponentanalyse	Ja	Rimeligt velegnet til gruppering af observationer og som dimensionsreducerende værktøj.
Faktoranalyse	Ja	De fundne grupper af variable virker umiddelbart rimeligt fornuftige.
Clusteranalyse på variable	Ja	Udmærket værktøj til gruppering af variable, grupperne virker som for faktoranalysen fornuftige.
Kanonisk korrelationsanalyse	Ja	Velegnet til gruppering af observationer, når der er to sæt beskrivende variable.
Superviserede metoder		
Diskriminantanalyse	Ja	Model testes med krydsvalidering, hvorved resultatet bliver mere troværdigt.
Logistisk regressionsanalyse	Nej	For få observationer, fuldstændig separation af data forhindrer udførelse af analyse.
Klassifikationstræer	Nej	Alt for få observationer, skal bruge flere hundreder, så træet kan testes/beskæres.

Tabel 18: Oversigt over om de enkelte metoder er velegnede set i relation til dette projekts data.

De ikke-superviserede metoder har altså vist sig at være rimeligt velegnede, mens det af de superviserede metoder kun er diskriminantanalysen, der er egnet til den givne type data.

5 Konklusion

Formålet med dette projekt var at undersøge, om det med multivariat dataanalyse, ved anvendelse af et datamateriale med et meget lille antal observationer i forhold til antallet af variable, var muligt at finde nogle enkelte proteiner eller en gruppe af proteiner, som havde betydning for, om de undersøgte rotter havde epilepsi eller ej. De metoder, der blev benyttet, var dels nogle af de standardmetoder, der benyttes i dag (ikke-parametrisk ensidet variansanalyse og t-test) og dels en række multivariate metoder, der kan deles op i to hovedgrupper, ikke-superviserede og superviserede. Ikke-superviserede metoder benyttes til at gruppere data (enten observationerne eller variablene) bedst muligt uden at benytte oplysninger, om hvilken gruppe observationerne stammer fra, mens superviserede metoder udnytter de givne oplysninger om de forskellige klasser og bestemmer en klassifikationsregel baseret på dette.

Principal komponentanalyse og kanonisk korrelationsanalyse viste sig at være velegnede til at gruppere observationerne. I begge tilfælde var det muligt at adskille kontrolgruppen og gruppen med epilepsi under udvikling fra hinanden og fra de to grupper med epilepsi i 'udbrud' under et. Ved principal komponentanalyse blev til dette benyttet de tre første principale komponenter, og for kanonisk korrelationsanalyse blev benyttet det første par kanoniske variable, som tager højde for mere end 96 % af den totale variation i data. Ved at medtage det andet par kanoniske variable blev det desuden muligt at få adskilt de to sidstnævnte grupper. Variationen i data for de to grupper med epilepsi i 'udbrud' er altså meget lille i forhold til den totale variation i data.

Der blev med faktoranalyse og clusteranalyse på variablene fundet nogle grupper af proteiner, der ser ud til at være interessante i forbindelse med epilepsi. Ved faktoranalysen blev variablene inddelt i fem grupper, og det viste sig, at en meget stor del af de pletter, der blev fundet som signifikante ved de univariate metoder, tilhørte den samme gruppe (givet ved faktor 4, logtransformerede data, markeret med gul i figur 12). Mange af de signifikante variable må altså variere på nogenlunde samme vis. Det kunne tænkes, at eftersom der er mange i denne gruppe, der ser ud til at have forbindelse til epilepsi alene baseret på univariate metoder, er der måske endnu flere i denne gruppe, der, selvom de ikke i sig selv udviser signifikans, stadig kunne have en indflydelse.

Ved clusteranalysen blev variablene inddelt i væsentligt flere grupper, 240, og det viste sig, at der var en vis sammenhæng mellem grupperne fundet ved faktoranalysen og grupperne fundet ved clusteranalysen. For hver cluster gjaldt det, at der var én faktor, der var dominerende, størstedelen af medlemmerne i en cluster tilhørte altså den samme gruppe givet ved faktorerne. De mest signifikante pletter på gelen, fundet ved de univariate metoder, viste sig desuden at tilhøre relativt få clusters, hvor især cluster nr. 2, nr. 169 og nr. 176 må fremhæves, da næsten alle variablene i disse tre grupper var blandt de 100 mest signifikante. Samtlige medlemmer i disse tre grupper samt alle medlemmer i de to grupper, nr. 20 og nr. 102, som ligeledes er nogle af de clusters, hvis medlemmer er blandt de mest signifikante, tilhørte desuden den allerede nævnte gruppe givet ved faktor 4. Af clustrene fundet for de 100 mest signifikante variable, var der to grupper, der var specifikke for variansanalysen, hvor fire grupper blev sammenlignet, og altså ikke blev fundet for hverken variansanalysen med to grupper eller t-testen. Medlemmerne i disse to clusters, nr. 60 og nr. 146, kunne derfor tænkes at være gode til at skelne mellem forskellige grader af epilepsi og ikke blot mellem syg og rask.

Ved hjælp af de superviserede metoder er der blevet fundet nogle enkelte proteiner, der kunne se ud til at have en forbindelse til epilepsi. Ved diskriminantanalyse blev fundet at den bedste diskriminantfunktion indeholdt to til tre variable, nr. 843 og nr. 1567, som begge viste sig at tilhøre den omtalte cluster nr. 2, samt nr. 366 som tilhørte cluster nr. 230. De to første variable blev af CPA identificeret som forskellige modifikationsprodukter af 'glial fibrillary acidic protein delta', som ifølge CPA giver god mening at sætte i forbindelse med epilepsi. Variabel nr. 843 blev også fundet ved den stepvise logistiske regression i Enterprise Miner, hvor kun en enkelt variabel blev udvalgt til modellen. Det blev her fundet, at jo lavere værdien af variabel nr. 843 var, jo mindre var graden af epilepsi. Det samme resultat blev fundet ved diskriminantanalysen, hvor en lav værdi for variabel nr. 843 ligeledes pegede i retning af en mindre grad af epilepsi. Ved diskriminantanalysen afhang graden af epilepsi dog som sagt også af variabel nr. 1567 og nr. 366, så en lav værdi for nr. 843 ville ikke alene kunne gøre udslaget. Regressionsanalysen udført i Enterprise Miner skal dog tages med forbehold, da en tilsvarende kørsel med proc logistic i SASstat resulterede i en advarsel, om at maximum likelihood estimatet ikke eksisterede på grund af fuldstændig separation af data. Når antallet af observationer er så småt, og antallet af variable samtidig er så stort, vil det være muligt at finde en opdeling, der placerer alle observationer i de korrekte grupper, hvilket altså skete her. Det må formodes, at fuldstændig separation også opdages i

Enterprise Miner, at analysen stoppes, når dette sker, og at det givne resultat er fra iterationen, før fuldstændig separation opdages.

Den sidste superviserede metode, der blev benyttet, var klassifikationstræer opbygget i Enterprise Miner. Det træ, der blev fundet, var opbygget med kun tre variable, nr. 37, nr. 211 og nr. 24, og ved hjælp af disse blev de fire grupper adskilt fuldstændigt. Alle tre variable blev fundet blandt de signifikante for variansanalysen, hvor fire grupper blev sammenlignet, men kun nr. 37 blev fundet signifikant ved variansanalysen med to grupper. Dette stemmer overens med, at variabel nr. 37 blev benyttet til at skille kontrolgruppen fra de øvrige tre grupper, mens de to variable, nr. 211 og nr. 24 blev benyttet til at skille de tre syge grupper fra hinanden. De tre variable viste sig at tilhøre tre forskellige clusters, nr. 176, nr. 5 og nr. 146, hvor nr. 176 er en af de fem tidligere nævnte clusters, hvis medlemmer alle tilhørte den samme gruppe fundet ved faktoranalysen, og nr. 146 er en af de to clusters, som kun blev fundet for variansanalysen, hvor fire grupper blev sammenlignet. Der er dog stor usikkerhed på resultatet, for det første fordi Enterprise Miner's 'Tree Node' er beregnet til meget store datasæt (flere tusinde observationer), og for det andet fordi det fundne træ ikke er blevet beskåret, som det ellers burde, enten ved hjælp af et testdatasæt eller ved krydsvalidering, og metoden er derfor ikke særlig velegnet til denne type af data.

Ganske kort opsummeret vil det altså sige, at de ikke-superviserede metoder har vist sig ganske velegnede til at gruppere dels observationer, dels variable, mens det kun var diskriminantanalysen blandt de superviserede metoder, der har vist sig at være en egnet metode at anvende til dette projekts data.

Det tilbagevendende store problem ved analyserne i dette projekt er det meget lille antal observationer og det meget store antal variable – med kun 18 observationer at basere en undersøgelse af 1849 variable på må der påregnes en stor grad af usikkerhed på resultaterne. Antallet af variable er i sig selv et problem på grund af 'curse of dimensionality', så at gøre antallet af observationer større vil ikke nødvendigvis afhjælpe alle problemer, men en større mængde data ville kunne gøre resultaterne mere pålidelige, og usikkerheden på analyserne ville kunne mindskes. Det ville med flere observationer kunne blive muligt at benytte en del af sit datamateriale til validering af resultaterne.

Når det er sagt, er det dog som afslutning stadig værd at påpege, at de fundne resultater har vist sig at give god mening og stemme godt overens med hinanden.

Referencer

1. Albert, A. & Anderson, J. A.: On the existence of maximum likelihood estimates in logistic regression models, Biometrika Vol. 71 (1), p. 1-10, 1984.

2. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis, 2. udgave, John Wiley & Sons, Inc., 1971.

3. Breiman, Leo, Friedman, Jerome H., Olshen, Richard A. & Stone, Charles J.: Classification And Regression Trees, Wadsworth, Inc., 1984.

4. California Statistical Software, Inc.: An Introduction to CART Methodology, 48 sider, 961 Yorkshire Court, Lafayette, CA 94549, Copyright© 1985.

5. Conover, W. J.: Practical Nonparametric Statistics, 3. udgave, John Wiley & Sons, Inc., 1999.

6. Conradsen, Knut: En Introduktion til Statistik, Bind 1A-1B, 7. udgave, IMM, Lyngby, 1999.

7. Conradsen, Knut: En Introduktion til Statistik, Bind 2A-2B, 4. udgave, IMM, Lyngby, 1984.

8. Conradsen, Line: Statistiske analyser af to-dimensionale elektroforese-geler, IMM, Lyngby, 2002.

9. CPA, Centre for Proteome Analysis, information hentet fra www.proteome-analysis.dk

10. Donoho, David L.: Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality, Department of Statistics, Stanford University, August 8, 2000, http://www-stat.stanford.edu/~donoho.

11. Hanka, Rudolf & Harte, Thomas P.: Curse of Dimensionality: Classifying Large Multi-Dimensional Images with Neural Networks, Medical Informatics Unit, University of Cambridge, 1996, http://www.medinfo.cam.ac.uk/miu/papers/thomas/paper.html.

12. Hastie, Trevor, Tibshirani, Robert & Friedman Jerome: The Elements of Statistical Learning, Data Mining, Inference and Prediction, Springer-Verlag, 2001.

13. Hilger, Klaus Baggesen: Exploratory analysis of multivariate data, IMM, Lyngby, 2001.

14. Horn, J. L.: A rational and test for the number of factors in factor analysis, Psychometrika, 30, 179-185, 1965.

15. Pedersen, Lars: Analysis of two-dimensional electrophoresis gel images, IMM, Lyngby, 2002.

16. SAS Institute Inc., SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1-2, Cary, NC: SAS Institute Inc., 1989.

17. Sharma, Subhash: Applied Multivariate Techniques, John Wiley & Sons, Inc., 1996.

18. Stokes, Maura, E., Davis, Charles S., Koch, Gary G.: Categorical Data Analysis Using the SAS System, Cary, NC: SAS Institute Inc., 1995.

Filename: eks_proj_mia
Directory: C:\Documents and Settings\fkc\My Documents
Template: C:\Documents and Settings\fkc\Application Data\Microsoft\Templates\Normal.dot
Title: Indledning
Subject:
Author: Mia Skettrup
Keywords:
Comments:
Creation Date: 27-03-2003 14:12
Change Number: 3
Last Saved On: 27-03-2003 14:19
Last Saved By: Mia Skettrup
Total Editing Time: 5 Minutes
Last Printed On: 07-04-2003 13:25
As of Last Complete Printing
Number of Pages: 94
Number of Words: 23.714 (approx.)
Number of Characters: 135.171 (approx.)