Multivariat analyse av DCE-MRI-bilete av kreftsvulstar

Multivariate analysis of DCE-MRI images of cancer tumours

Turid Katrine Gjerstad Torheim

Universitetet for miljø- og biovitenskap
Institutt for matematiske realfag og teknologi
Masteroppgave 60 stp. 2011

Forord

Denne oppgåva avsluttar mitt studium Master i Matematiske realfag ved Universitetet for Miljø- og Biovitskap. Arbeidet tilsvarar 60 studiepoeng.

Cecilia Marie Futsæther har vore hovudrettleiaren min, og skal ha stor takk for god hjelp og oppfølging. Ho introduserte meg for temaet, og eg set stor pris på å få arbeide med noko som kjennast nyttig.

Knut Kvaal og Ole Mathis Opstad Kruse har også bidrege med gode råd, og ikkje minst programmeringshjelp. Takk til Ole Mathis og Ulf Geir Indahl for lån av Matlab-skript.

Takk til Eirik Malinen for forklaringar på vanskelege medisinske omgrep og fenomen som eg ikkje kunne noko om før eg byrja. Han skal også ha takk for å ha skaffa og bearbeida datasettet, saman med Erlend Kristoffer Frivold Andersen og Heidi Lyng.

Og til slutt takk til Amund, for korrekturlesing, "tech-support" og generelt tolmod.

Turid Katrine Gjerstad Torheim
Ås, 13.12.2011

Samandrag

Denne masteroppgåva byggjer på eit DCE-MRI (*Dynamic Contrast Enhaced Magnetic Resonance Imaging*) -studium av 88 pasientar med livmorhalskreft, gjennomført på Det Norske Radiumhospitalet (no ein del av Oslo universitetssykehus) i perioden 2001-2004. I DCE-MRI-undersøkingane målast den relative signalauken RSI frå vevet etter injisering av eit kontrastmiddel, og dette gjev ein tidsserie på 14 bilete. Målingane har blitt tilpassa ein farmakokinetisk modell kalla Brix-modellen, som reduserer tidsserien frå kvar voksel ned til tre modellparameterar. Alle pasientane har så fått behandling i form av stråleterapi, med jamleg oppfølging i etterkant. Målet med denne oppgåva er å undersøkje om parameterane frå Brix-modellen kan knyttast til behandlingsutfall i form av progresjonsfri overleving, det vil seie om pasienten vert frisk att eller ikkje. Analysane i denne oppgåva skil ikkje mellom tilbakefall i form av metastasar og lokalt tilbakefall. Til skilnad frå tidlegare studium, nyttar denne oppgåva multivariate statistiske metodar.

Dei multivariate metodane nytta i oppgåva er prinsipalkomponentanalyse (PCA), diskriminant analyse (LDA og QDA), klyngeanalyse, PLS, lineær regresjon, SIMCA og støttevektormaskiner (SVM).Vi kombinerer også LDA med ein variabelseleksjon, for å fjerne variablar som gjev lite informasjon. I analysane nyttar vi deskriptive statistiske parameterar, som til dømes gjennomsnitt, standardavvik og persentilverdiar, berekna ut i frå Brix-parameterane for kvar svulst. I èin analyse nyttar vi også histogramframstilling av Brix-parameterane over svulsten.

PCA syner at datasettet bestående av alder på pasienten, stadiet av sjukdom, svulstvolum og dei deskriptive statistiske parameterane kan reduserast til få prinsipalkomponentar utan å miste mykje informasjon. Det trengst berre åtte komponentar for å forklare over 90% av den totale variansen i dei 64 variablane. Inspeksjon av skårplott syner ikkje grupperingar som samsvarar med behandlingsutfall, det vile seie pasientar som vert friske og pasientar som får tilbakefall, med unntak av eitt av dei tredimensjonale skårplotta.

PLS med dei same forklaringsvariablane som i PCA og anten behandlingsutfall, stadium eller svulstvolum som respons, gjev forklart varians på 50% - 60% i kalibrering, men residualvarians på over 100% etter full kryssvalidering. Heller ikkje lineær regresjon med utvalde komponentar frå PCA-modellen forklarer behandlingsutfall godt.

Ikkje-overvaka klassifisering i form av K-means- og K-medians-klyngeanalyse gjev ikkje gruppeinndeling som samsvarar med utfallet av stråleterapi.

Overvaka klassifisering i form av LDA og QDA lukkast betre i å skilje mellom utfalla. LDA etter variabelseleksjon, der variablane som forklarer 90% av totalvariansen vert nytta som forklaringsvariablar, syner seg å klassifisere signifikant, med ein p-verdi på 0,011 for både tilbakefallspasientane og pasientane som vert friske att.

Dei ikkje-lineære metodane SIMCA og SVM er dei som gjev mest nøyaktig klassifikasjon, det vil seie dei som plasserer flest pasientar i riktig gruppe. SIMCA gjev ei nøyaktigheit på 91%, sensitivitet (andel riktig klassifiserte pasientar av pasientane som vart friske) på 100% og spesifisitet (andel riktig klassifiserte tilbakefallspasientar) på 78%, medan SVM gjev nøyaktigheit på 93%, sensitivitet på 96% og spesifisitet på 88%. SVM-modellen er også god

etter full kryssvalidering, med 88% nøyaktigheit, trass i mange støttevektorar.

Konklusjonen er at multivariate metodar kan vere nyttige i analyse av DCE-MRI-bilete, sidan dei gjev kvantitative mål på nøyaktigheita til klassifiseringane og gjer det mogleg å identifisere kva svulstar som vert feilklassifiserte. Det er også ein fordel at metodane automatisk tek omsyn til samspel mellom variablar.

Abstract

This master's thesis is based on a DCE-MRI (Dynamic Contrast Enhaced Magnetic Resonance Imaging) study of 88 patients with cervical cancer, performed at the Norwegian Radium Hospital (now a part of Oslo Universty Hospital) in the period from 2001 to 2004. The DCE-MRI examination measures the relative signal increase (RSI) from the tissue after injection of a contrast agent, and this gives a time series of 14 images. The measurements have been fitted to a pharmacokinetic model, the Brix model, and this reduces the time series from each voxel to three model parameters. All patients have been treated with radiotherapy, and have been followed up afterwards. The aim of this thesis is to examine whether the parameters from the Brix model can be associated with treatment outcome measured by progression free survival, that is whether the patient is cured from the cancer or not. We do not separate between locoregional and distant relapse. In contrast to earlier studies, this study uses multivariate statistical methods.

The multivariate methods used in this study are principal component analysis (PCA), discriminant analysis (LDA and QDA), cluster analysis, PLS, linear regression, SIMCA and support vector machines (SVM). We also combine LDA with a variable selection, on order to remove variables that provide little information. In the analyses we use descriptive statistical parameters, such as average, standard deviation and percentile values, calculated from the Brix parameters for each tumour. In one of the analyses we also use histogram values of the Brix parameters over each tumour.

PCA shows that the data set consisting of patient age, tumour stage, tumour volume, and the descriptive statistical parameters, can be reduced to few principal components without losing much information. We only need eight principal components to explain 90% of the total variance of the 64 variables. Inspection of score plots show no grouping consistent with treatment outcome, that is patients that are cured and patients with relapse, with one exeption in one of the three dimensional score plots.

PLS with the same explanatory variables as in PCA and either treatment outcome, tumour stage or tumour volume as response variable, gives explained variance of 50%-60% in calibration, but over 100% residual variance after full cross validation. Nor linear regression with chosen principal component from the PCA model can explain treatment outcome well.

Unsupervised classification, in the form of K-means and K-medians cluster analysis, does not give grouping consistent with treatment outcome.

Supervised classification, LDA and QDA, is more successful in separating the two treatment outcomes. LDA after a variable selection where the variabels needed to explain 90% of the total variance is used as explanatory variables, gives significant classification with p-value 0.011 for both patients with relapse and patients that were cured.

The nonlinear methods SIMCA and SVM gives the most accurate classification, that is they predict the correct treatment outcome for most patients. SIMCA gives accuracy 91%, sensitivity 100% (the fraction of cured patients correctly classified as cured) and specificity 78% (the fraction of correctly classified relapse patients), while SM gives accuracy 93%, sensitivity 96% and specificity 88%. The SVM model is still accurate after full cross

validation, then with 88% accuracy, despite having many support vectors.

The conclusion is that multivariate methods can be of use in analysis of DCE-MRI-images, because they give quantitative measurements on the accuracy of the classifications and provide the possibility to identify the tumours that are incorrectly classified. It is also an advantage that the metods automatically take into concideration the interaction between variables.

Innhald

1 Innleiing

Livmorhalskreft er ein sjukdom som rammar kring 270 kvinner i Noreg kvart år, *[1]*. Sjukdommen synast å vere nært knytta til infeksjonar av HPV, humant papillomavirus, *[2]*. Behandlingsformene for livmorhalskreft er kirurgi, kjemoterapi og stråleterapi, avhengig av kor langt utvikla sjukdommen er når den vert oppdaga. Dei fleste, om lag 73%, vert friske att, men ikkje alle er like heldige.

Avbildingsteknikkar som magnetresonanstomografi (MRI) nyttast i kreftdiagnostikk blant anna for å lokalisere og bestemme storleiken av svulsten, *[3]*. Teknikken nyttar dei magnetiske eigenskapane til atomkjernar til å danne eit bilete av vevet i kroppen. MRI kan skilje ulike typar mjukt vev, i motsetnad til røntgen, som skil mellom mjukt og hardt vev, *[3]*. Sidan kreftsvulstar består av mjukt vev, kan desse undersøkjast ved hjelp av MRI.

Dynamisk kontrastforsterka (DCE) MRI nyttar til skilnad frå vanleg MRI, eit kontrastmiddel, *[4]*. Ved å ta ein tidsserie av MR-bilete etter å ha injisert kontrastmidlet, kan ein skildre opptak og utvasking av midlet. Ved å samalikne med eit bilete teke utan kontrastmiddel, prekontrastbiletet, kan ein sjå kva område av vevet som tek opp mykje kontrastmiddel. Prinsippet bak DCE-MRI-undersøkingar av svulstar, er at kreftvev og friskt vev har ulik struktur. Karnettverket i kreftvev er kaotisk og har mykje lekkasjar, noko ein ikkje ser i friskt vev, *[5]*. Kontrastmiddelet vil leke inn i vevet i svulsten, slik at denne skil seg ut på MR-biletet. Dette kan nyttast til å lokalisere svulsten, samt å finne område i svulsten som har meir lekkasje enn andre.

Tidsserien frå DCE-MRI-målinga kan tilpassast farmakokinetiske modellar, det vil seie modellar som skildrar korleis legemiddel bevegar seg gjennom kroppen, *[6]*. Desse matematiske modellane reduserer tidsserien til nokre få modellparameterar, som kan knyttast til biologiske eigenskapar ved vevet, *[7]*. Data frå MRI-undersøkinga nytta i denne oppgåva er tilpassa Brix-modellen, ein farmakokinetisk modell for kontrastmiddel i DCE-MRI utvikla av Brix et al. (1991), *[8]*. Denne modellen gjev tre parameterar for kvart volumelement (voksel) av svulsten. Også andre modellar, til dømes RR-modellen, *[9]*, kan nyttast, men tidlegare studiar, *[10]*, tyder på at Brix-modellen er den som i størst grad kan knyttast til behandlingsutfall.

Tidlegare studiar, oppsummert av Zahra et al., *[4]*, indikerer at det er samanheng mellom DCE-MRI-bilete av svulstrn før behandling og utfallet av behandlinga. Fleire studiar, som til dømes Cooper et al., *[11]*, og Loncaster et al., *[12]*, syner at oksygenmengda i svulstane korrelerer med DCE-MRI-målingar. Låg oksygenmengde i vevet, ein tilstand kalla hypoksi, påverkar utfallet av stråleterapi negativt, *[13]*. Loncaster et al. *[12]*, knyttar også ein av parameterane frå Brix-modellen til behandlingsutfall.

Zahra et al., *[4]*, foreslår også at prediksjonar av behandlingsutfall kan nyttast til å tilpasse behandlinga som pasientane skal få. Dersom ein kan identifisere kritiske område av svulsten, foreslår dei at ein kan auke stråledosen til dette spesifikke området, slik at dette området får ein større dose utan at ein må auke dosen over heile svulsten. For pasientar som predikerast å respondere godt på behandling, foreslår dei å vurdere reduksjon eller utelukking av kjemoterapi. DCE-MRI kan altså potensielt nyttast til å skreddarsy behandlinga til kvar enkelt pasient, og såleis auke prognosane for å verte kurert for livmorhalskreft.

Datasettet som denne oppgåva byggjer på, er henta frå eit DCE-MRI-studie utført ved Det Norske Radiumhospitalet, no ein del av Oslo universitetssykehus, i perioden frå 2001 til 2004, *[14]*. Studiet omfattar 88 pasientar med livmorhalskreft i ulike stadium. Etter undersøkinga har pasientane fått behandling i form av stråleterapi, og deretter regelmessig oppfølging. 32 av pasientane fekk tilbakefall etter behandlinga, anten i form av lokalt tilbakefall eller som metastasar.

Erlend Andersen tok for seg denne DCE-MRI-undersøkinga i si masteroppgåve i 2009, *[10]*, og tilpassa data til både Brix-modellen og RR-modellen. Han nytta logistisk regresjon for å undersøkje om parameterane i modellane kan knyttast til behandlingsutfall. Analysen hans indikerer at èin av parameterane i Brix-modellen ser ut til å skilje mellom pasientar som får tilbakefall og dei som vert friske att. Parameterane frå RR-modellen gav ikkje signifikante skiljer mellom behandlingsutfall. Andersen et al., *[15]*, har seinare (2011) ved hjelp av log-rank-testar vist at persentilverdiar av DCE-MRI-målingar kan knyttast til behandlingsutfall.

Andersen et al., *[14]*, har også gjort K-means-klyngeanalysar av det same datasettet, der målet var å identifisere område av svulsten som reagerer mindre på stråleterapi, og såleis kan knyttast til lokalt tilbakefall. Her deler dei vokslane i kvar svulst inn i tre klynger etter verdiane av to parameterar frå ein farmakokinetisk modell. Dei finn at ei av klyngene er sigifikant mindre for pasientar som fekk lokalt tilbakefall enn for dei som ikkje fekk det.

Tidlegare analysar av DCE-MRI og Brix-parameterar har for det meste nytta seg av univariate statistiske metodar, der ein undersøkjer kvar variabel for seg. I denne oppgåva vil vi nytte multivariate metodar for å undersøkje desse betre kan vise om Brix-parameterane predikerer behandlingsutfall i form av progresjonsfri overleving, det vil seie om pasientane vert friske att eller ikkje. I multivariate analysar undersøkjer ein alle variablar samtidig, slik at ein også tek omsyn til samspel mellom variablar, *[16]*. Univariate analysar ser på kvar variabel for seg, noko som gjer det vanskeleg å sjå korleis dei ulike variablane varierer saman.

Denne oppgåva er bygd opp slik at teorien bak MRI generelt og DCE-MRI spesielt vert forklart i kapittel 2. Dette kapitlet tek også for seg korleis kreftsvulstar skil seg frå friskt vev, og korleis ein fastset stadiet til ein livmorhalskreftsvulst. Det avsluttast med ein innføring i farmakokinetiske modellar, og då særleg Brix-modellen.

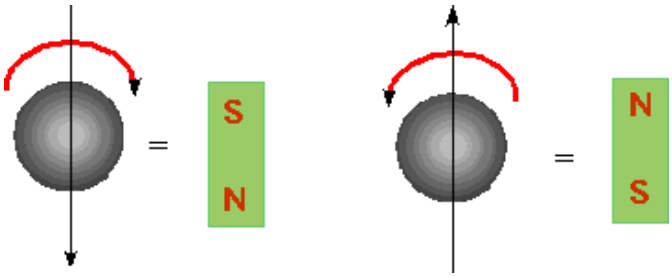
Materiale og metodar presenterast i kapittel 3, med blant anna fleire detaljar om datasettet. Det vert skildra korleis data kan representerast som deskriptiv statistikk for kvar svulst, eventuelt ved histogramfordelingar for kvar av Brix-parameterane. Dei multivariate metodane som nyttast i oppgåva vert også introduserte.

Resultata av analysane visast i form av plott og talverdiar i kapittel 4. I kapittel 5 vert resultata diskuterte, og sett i samanheng med andre studiar, før dette oppsummerast i konklusjonen i kapittel 6.

2 Teori

2.1 MRI

MRI står for *magnetic resonance imaging*, på norsk magnetresonanstomografi, og nyttar seg av dei magnetiske eigenskapane til atomkjernar, *[17]*. Kjernane er sett saman av proton og nøytron. Begge desse partiklane har spinn, det vil seie at dei roterer om sin eigen akse, *[18]*. Spinnet er ein eigenskap partiklane har, på same måte som masse og ladning, uavhengig av omgivnadane til partikkelen. Proton, nøytron og elektron har spinn på $\frac{1}{2}\hbar$, der $\hbar = h/2\pi$ og h er Planckkonstanten, $6{,}63 \times 10^{-34}$ kgm^2/s. Spinnet gjev opphav til eit magnetisk moment for kvar av partiklane, sjå *figur 1*.



Figur 1: Eit proton, nøytron eller elektron har spinn 1/2ħ. Avhengig av retninga til spinnet, gjev dette opphav til to ulike energinivå, og magnetisk moment i to ulike retningar. Henta frå: http://www.physics.carleton.ca/~watson/Physics/1000_level/1008_Modern_Physics/1008_Nuclear_Physics_app.html

Proton og nøytron er partiklar av typen fermion, som må følgje Pauli sitt eksklusjonsprinsipp når dei skal setjast saman til ein atomkjerne, *[19]*. Det vil seie at for kvart proton eller nøytron som spinn i ei retning, må det neste protonet eller nøytronet spinne i motsett retning. Kjernar med eit partal kjernepartiklar vil difor ikkje ha noko netto magnetisk moment, medan kjernar med oddetal kjernepartiklar vil ha upara partiklar, og difor magnetisk moment. Dette kallast det magnetiske dipolmomentet μ_B (eining: J/T) til kjernen.

Dei magnetiske dipolmomenta vil verte påverka av eit ytre magnetfelt, og det er dette ein utnyttar i MRI. I prinsippet kan ein bruke alle kjernar som har dipolmoment, men det er vanlegast å sjå på hydrogenkjernar (^{1}H), altså proton, *[20]*. Det kjem av at hydrogen finst i store mengder i kroppen. Kring 60% av kroppen er vatn (H_2O), *[17]*, og det er i tillegg mykje hydrogen i feittvev. Ein annan eigenskap som gjer hydrogen ekstra eigna, er at det reagerer kraftig på eksterne magnetfelt. Kor sterkt ein atomkjerne reagerer på magnetfelt er gitt ved det gyromagnetiske forholdet γ. For ^{1}H er $\gamma = 42{,}6$ Mhz/T, *[17]*. For samanlikning er det gyromagnetiske forholdet for nokre vanlege atomkjernar gitt i *tabell 1*.

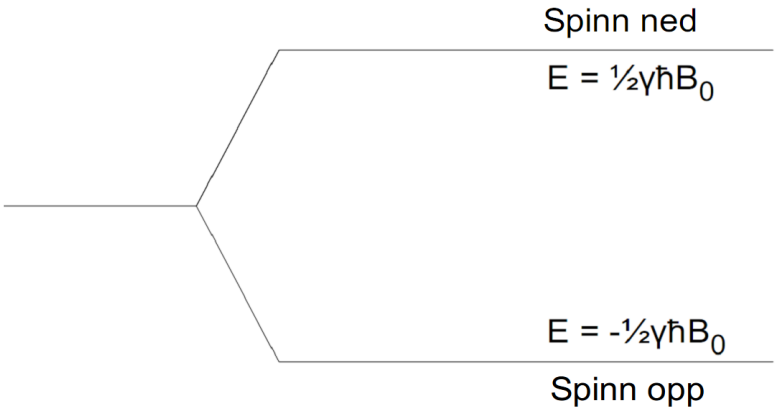
Tabell 1: Det gyromagnetiske forholdet (MHz/T) for nokre atomkjernar. Henta frå Hornak, [17].

Kjerne	Gyromagnetisk forhold γ (MHz/T)
^{1}H	42,58
^{2}H	6,54
^{31}P	17,25
^{23}Na	11,27
^{14}N	3,08
^{13}C	10,71
^{19}F	40,08

MRI-undersøkinga startar med at vevet ein vil undersøkje blir utsett for eit kraftig magnetfelt $\mathbf{B_0}$, typisk kring 1-2 T, *[3]*. Protona vil vekselverke med magnetfeltet, slik at dei magnetiske dipolmomenta til kvart proton rettar seg etter $\mathbf{B_0}$. Det kan dei gjere på to måtar, anten ved å la $\boldsymbol{\mu_B}$ og $\mathbf{B_0}$ peike i same retning (parallelt, også kalla "spinn opp") eller ved å la dei peike i motsett retning (antiparallelt, "spinn ned"). Desse to tilstandane har kvar sitt energinivå, synt i *figur 2*. Energien er gitt ved

$$E = \gamma \hbar I B_0$$

der $I = \pm 1/2$ er spinnkvantetalet, γ er det gyromagnetiske forholdet, $\hbar = h/2\pi$ og h er Planckkonstanten og B_0 er den magnetiske flukstettleiken, *[3]*.



Figur 2: Det ytre magnetfeltet B_0 fordeler protona på to ulike energinivå, der protona med spinn retta same veg som det ytre feltet har mindre energi enn dei som er retta motsett veg.

Protona som har spinn parallelt med det ytre magnetfeltet har mindre energi enn dei som er antiparallelle. Rett etter at B_0 er påført vil det vere like mange proton i kvar av dei to tilstandane. Det er såleis ikkje noko netto magnetisering til stades. Energiutveksling med omgivnadane vil føre til at nokre proton fell frå høgt til lågt energinivå. Det vil snart oppstå ei termisk jamvekt, der protona fordeler seg på dei to energinivåa etter Boltzmannfordelinga, gitt ved

$$\frac{N_{parallell}}{N_{antiparallell}} = e^{\frac{\Delta E}{k_B T}} = e^{\frac{\gamma \hbar B}{k_B T}}$$

der $N_{parallell}$ er talet på parallelle spinn, $N_{antiparallell}$ er talet på antiparallelle spinn, γ er det gyromagnetiske forholdet, k_B er Boltzmann-konstanten, $1{,}38\times10^{-23}$ J/K, og T er temperaturen målt i Kelvin, *[20]*.

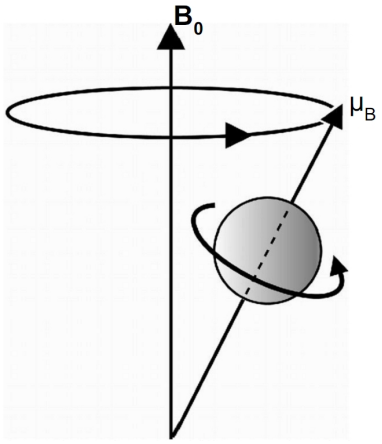
Det vil såleis vere flest proton i den lågaste energitilstanden. Sidan det er fleire proton parallelt med det eksterne magnetfeltet enn det er proton som er antiparallelle med magnetfeltet, vil vi få ei netto magnetisering $\mathbf{M_0}$ av vevet, der magnetiseringa er summen av alle dipolmomenta i vevet,

$$\boldsymbol{M}_0 = \sum \boldsymbol{\mu}_B$$

I tillegg til å endre orientering, vil kvart enkelt proton presere kring retninga til det ytre magnetfeltet, som vist *figur 3*. Frekvensen til denne presesjonsrørsla kallast Larmorfrekvensen $\boldsymbol{\omega_0}$, *[3]*, og er avhengig av styrken B_0 til magnetfeltet,

$$\omega_0 = \gamma B_0$$

der γ er det gyromagnetiske forholdet til kjernen. For eit proton i eit magnetfelt på 1 T, er ω_0 = 42,6 MHz. Sidan protona preserer uavhengig av kvarandre (ute av fase), vil nettomagnetiseringsvektoren $\mathbf{M_0}$ ikkje presere, men ha konstant storleik og retning.



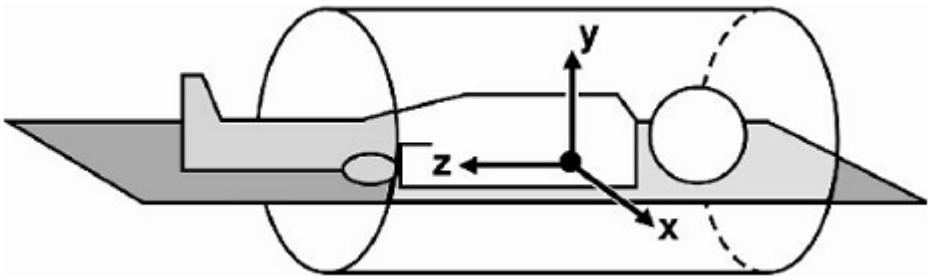
Figur 3: Ein partikkel med spinn $\boldsymbol{\mu_B}$ som blir utsett for eit magnetfelt $\boldsymbol{B_0}$, vil presere. Henta frå Hashemi og Bradley, [21].

Sjølve målingane i MR-undersøkinga gjerast med spolar plassert kring pasienten, *[21]*. Nokre spolar set opp eksterne magnetfelt, medan andre er mottakarspolar som tek opp signalet frå pasienten. For at det skal induserast straum i mottaktarspolane, fortel Faraday si lov at spolen må utsetjast for eit varierande magnetfelt, der

$$\varepsilon = -N \frac{d \Phi_B}{dt}$$

ε er den induserte spenninga, N er talet på viklingar i spolen og Φ_B er den magnetiske fluksen gjennom spolen, *[22]*.

Det er vanleg å definere eit aksekors som synt i *figur 4*, der *z*-aksen går langs kroppen til pasienten, og spolane kring pasienten ligg i *xy*-planet, *[21]*. For at det skal dannast eit signal, er det altså ikkje nok med det statiske magnetfeltet $\boldsymbol{B_0}$, som gir ein statisk $\boldsymbol{M_0}$, det trengst også eit magnetfelt som varierer med tida. $\boldsymbol{B_0}$ ligg i *z*-retning, og det same gjer difor $\boldsymbol{M_0}$. Det varierande magnetfeltet må gi ein magnetiseringskomponent i *xy*-planet, slik at den magnetiske fluksen gjennom *xy*-planet varierer, og gjev eit signal i mottakarspolen.



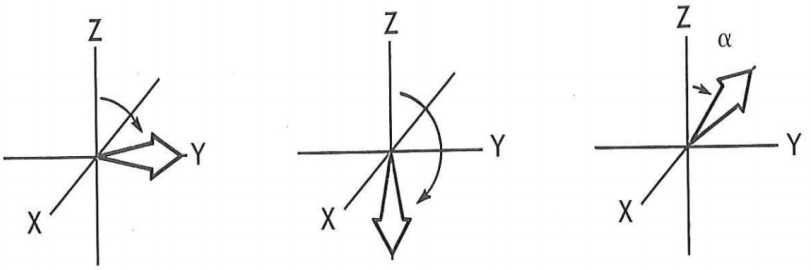
Figur 4: Person i MR-maskin. z-aksen går langs kroppen, medan mottakarspolen ligg i xy-planet. Henta frå Hashemi og Bradley, [21].

Ein måte å gjere dette på, er å sende ein puls av elektromagnetisk stråling gjennom vevet. Sidan elektromagnetisk stråling, som namnet seier, blant anna består av eit magnetfelt, vil denne pulsen kunne vekselverke med dipolmomenta til protona. For at det skal oppstå resonans mellom protona og det nye magnetfeltet, må frekvensen til den elektromagnetiske strålinga vere lik Larmor-frekvensen ω_0, *[20]*. Strålinga vil då kunne eksitere proton frå det låge til det høge energinivået, sjå *figur 2*. Energien til fotona som då vert absorbert, er gitt ved

$$E = \hbar \omega_0$$

Elektromagnetisk stråling med denne energien er radiobølgjer, difor vert strålingspulsen gjerne referert til som rf-pulsen (radiofrekvens). Ikkje alle proton vil absorbere energi. Proton på det høge energinivået kan også gå ned til det låge ved å sende ut stråling. Men sidan det i utgangspunktet var flest proton på lågt energinivå, vil det vere netto absorbsjon av energi i vevet. Dersom rf-pulsen varar lenge nok, kan vi igjen få like mange proton på begge energinivå. Då vert magnetiseringa M_z i *z*-retning null.

Energiabsorpsjonen frå rf-pulsen vil også føre til synkronisering av presesjonsrørsla til protona, *[21]*. Før pulsen var presesjonsrørslene ute av fase, slik at nettomagnetiseringa ikkje hadde noko komponent i *xy*-planet. Rf-pulsen vil gi magnetisering også i dette planet, slik at vi får ein komponent M_{xy}. Kor stor del av ***M*** som ligg i *xy*-planet, kjem an på styrken og varigheita til rf-pulsen. Ein flippvinkel på 90º, det vil seie at M_z forsvinn heilt og all magnetisering ligg i *xy*-planet, er vanleg, *[20]*. Dette er synt til venstre i *figur 5*. Resten av likningane i dette kapitlet tek utgangspunkt i at ein nyttar 90º-pulsar.

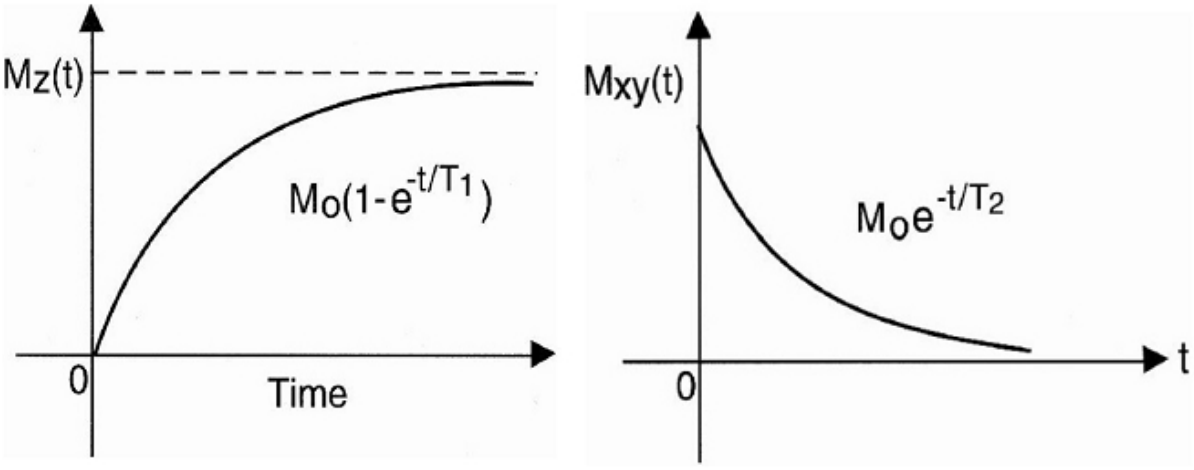


Figur 5: Venstre: 90º-flippvinkel. Midt: 180º-flippvinkel. Høgre: α º-flippvinkel. Henta frå Bushong, [3].

Straks rf-pulsen slåast av, vil protona byrje å presere i utakt att, og komponenten M_{xy} forsvinn. Utviklinga til M_{xy} er gitt ved

$$M_{xy}(t) = M_0 e^{-t/T_2} \qquad (1)$$

der tidskonstanten T_2 vert kalla spinn-spinn-relaksasjonstida. Denne utviklinga skuldast vekselverknader mellom enkeltatom og -molekyl, som gjev lokale inhomogenitetar i magnetfeltet. Grafar for M_{xy} og M_z er synt i *figur 6*.

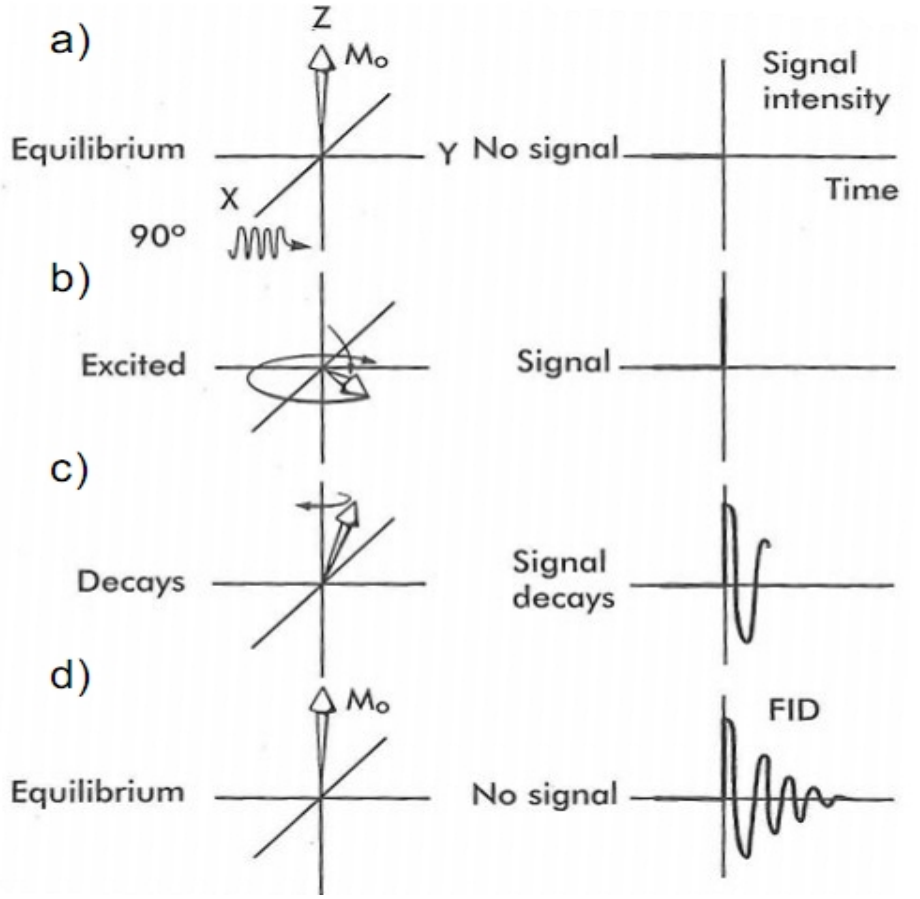


Figur 6: Utviklinga til magnetiseringskomponentane M_z og M_{xy} etter at rf-pulsen er slått av. Henta frå Hashemi og Bradley, [21].

Dei ytre magnetfelta kan også ha inhomogenitetar, som gjer at M_{xy} forsvinn endå raskare. Denne faktiske relaksasjonstida vert kalla T_2*, og erstattar då T_2 i likning (1) .
I tillegg vil energiutveksling med omgivnadane igjen føre til termisk jamvekt mellom dei to moglege energinivåa til protona, slik at magnetiseringa får ein *z*-komponent att, gitt ved

$$M_z(t) = M_0(1-e^{-t/T_1}) \qquad (2)$$

der T_1 er spinn-gitter-relaksasjonstida. Dette namnet kjem av at relaksasjonen skuldast kontakt mellom sjølve protona (spinna) og omgivnadane (gitteret). Signalet danna i mottakarspolen når magnetiseringa fell tilbake til utganspunktet den hadde før rf-pulsen, kallast FID - *free induction decay*, sjå *figur 7*.

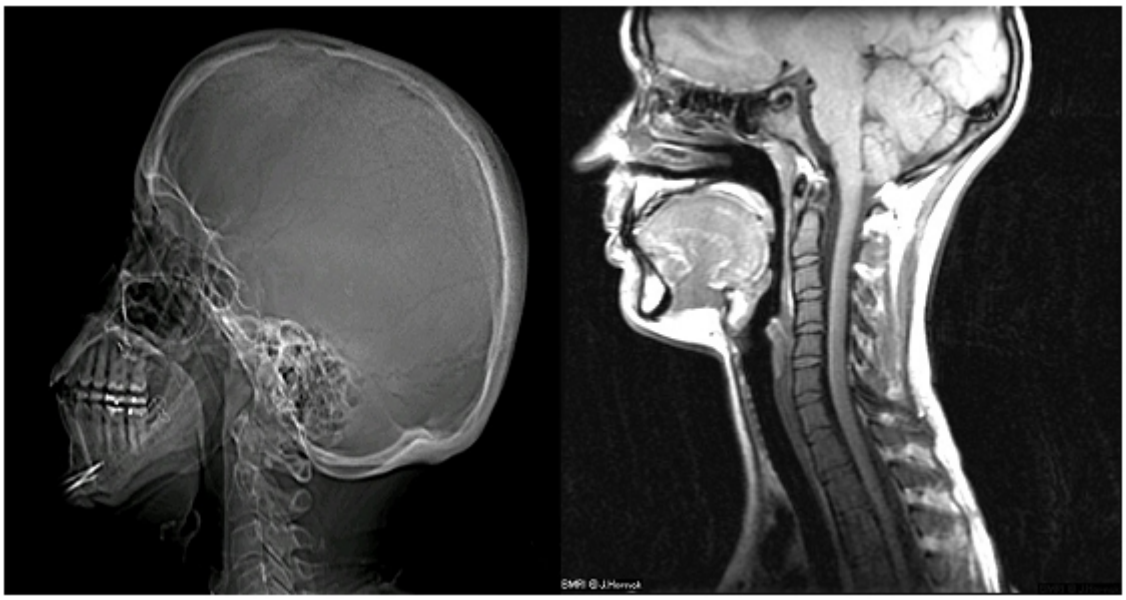


Figur 7: a) Ved jamvekt vert ein 90º rf-puls sendt inn. b) Denne flippar magnetiseringa ned til xy-planet, og det vert danna eit signal. c) Etter kvart som magnetiseringa fell tilbake til z-aksen, vil signalet minke. d) Til slutt er systemet tilbake i jamvekt. Signalet som vert danna, kallast FID, "free induction decay". Modifisert etter figur av Bushong, [3].

Ulike typar vev har ulike relaksasjonstider. Ved å finne relaksasjonstidene ut i frå mottakar-signala, kan ein difor finne ut kva type vev som finnast i det aktuelle området. Sidan vevet i kreftsvulstar skil seg frå friskt vev, jamfør *kap 2.3*, kan ein nytte MRI-bilete til å lokalisere svulstar.

For å danna komplette MR-bilete treng ein også romleg informasjon, ein må vite kvar i kroppen signala kjem frå. Dette oppnår ein ved å nytte gradientfelt, det vil seie magnetfelt som varierer i *x*-, *y*- eller *z*-retning, *[21]*. Med desse vil berre eit lite område av kroppen bli eksitert om gongen, slik at ein heile tida kan finne ut kvar signalet kjem frå. Signala vil då gjelde dette snittet av kroppen. For å få bilete av heile den aktuelle kroppsdelen med den oppløysinga ein ønskjer, lyt ein kombinere rf-pulsar og gradientfelt riktig. Slike samansetningar kallast pulssekvensar. Det finst to hovudtypar pulssekvensar: spinn-ekko-sekvensar og gradient-ekko-sekvensar. Innanfor desse igjen, finst det mange ulike variantar, men vi går ikkje nærmare inn på det i denne oppgåva. For ei innføring i pulssekvensar, sjå til dømes Hashemi og Bradley, *[21]*.

Også røntgenbilete skil mellom ulike typar vev. Røntgenbilete skil mellom hardt vev, som bein, og mjukt vev, medan MR-bilete skil mellom ulike typar mjukt vev, *[3]*. Difor er røntgen best eigna til å undersøkje til dømes beinbrot, medan MRI kan finne kreftsvulstar. *Figur 8* syner eit røntgenbilete og eit MR-bilete av hovud og nakke. Røntgenbiletet syner tenner og bein, medan MR-biletet syner det mjuke vevet. Legg særleg merke til tunga og strukturane i hjernen, som ikkje synast på røntgenbiletet.

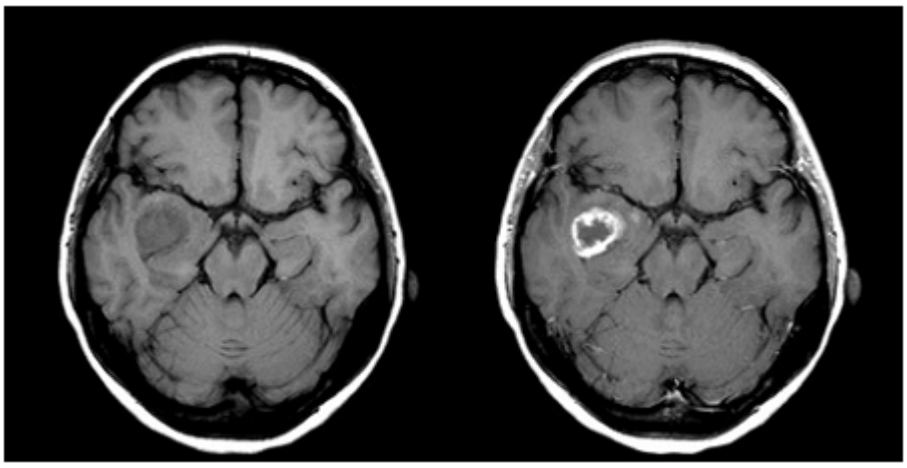


Figur 8: Røntgen-bilete (venstre) og MR-bilete (høgre). Røntgenbiletet syner først og fremst bein og tenner, medan det mjuke vevet er meir tydeleg på MR-biletet.
Kjelder: Røntgen: Aaron G. Filler, MD, PhD. MRI: Hornak, [17].

2.1.1 DCE-MRI

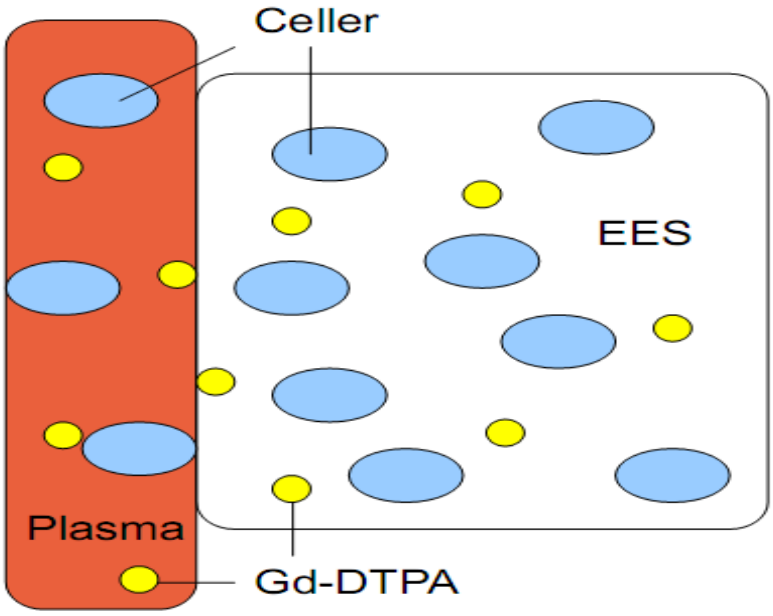
Ved dynamisk kontrastforsterka (DCE, *Dynamic Contrast Enhanced*) MRI føl ein konsentrasjonen av eit kontrastmiddel i vevet over tid, *[4]*. Dei vanlegaste kontrastmidla baserer seg på gadolinium (Gd^{3+}), *[23]*. Dette metallet er paramagnetisk, har åtte frie elektron og lang elektronspinn-relaksasjonstid. I ione-form er Gd giftig, så det må bindast som eit ligand, til dømes i Gd-DTPA, for å kunne brukast. Gd-DTPA er kontrastmidlet i undersøkinga nytta i denne oppgåva.

Ei DCE-MRI-undersøking startar med ein injeksjon av kontrastmiddel, *[4]*. Dette vert frakta med blodet til den delen av kroppen som skal undersøkjast. Formålet med kontrastmidlet er å endre T_1- og/eller T_2-relaksasjonstida ved spinninteraksjon mellom elektron i kontrastmidlet og proton i vevet, *[23]*. Dette vil auke kontrasten mellom vev som tek opp mykje kontrastmiddel og vev som tek opp lite. Ettersom kreftvev har dårlegare karnettverk enn friskt vev, sjå meir i *kap 2.3*, vil svulsten skilje seg ut på bileta. Eit døme på MR-bilete med og utan kontrastmiddel er vist i *figur 9*. Her ser ein at kontrastmidlet får svulsten til å tre tydeleg fram.



Figur 9: MR-bilete av ein pasient med hjernesvulst. Biletet til venstre er teke utan kontrastmiddel, medan biletet til høgre er teke etter injeksjon av kontrastmiddel. Kontrastmiddlet fører til at svulsten kjem tydelegare fram. Henta frå Bjørnerud, [23].

Når kontrastmidlet kjem til vevet, vil det byrje å leke ut gjennom kapillærveggane og samle seg i området mellom cellene, kalla det ekstracellulære ekstravaskulære rommet EES, *[23]*. Gd-DTPA er for stort til å trenge gjennom cellemembranen, og vil difor ikkje gå inn i sjølve cellene, jamfør *figur 10*. Etter kvart som konsentrasjonen av kontrastmiddel vert høg i EES, vil det leke tilbake att til blodet, og vaskast ut av kroppen. Opptaks- og utvaskingsratane er avhengige av blodgjennomstrøyminga, permeabiliteten til kapillærveggane og eigenskapar ved sjølve kontrastmidlet, til dømes storleiken. Utveksling av kontrastmiddel mellom blodet og EES kan skildrast med ulike farmakokinetiske modellar, *[24]*, sjå *kapittel 2.4*.



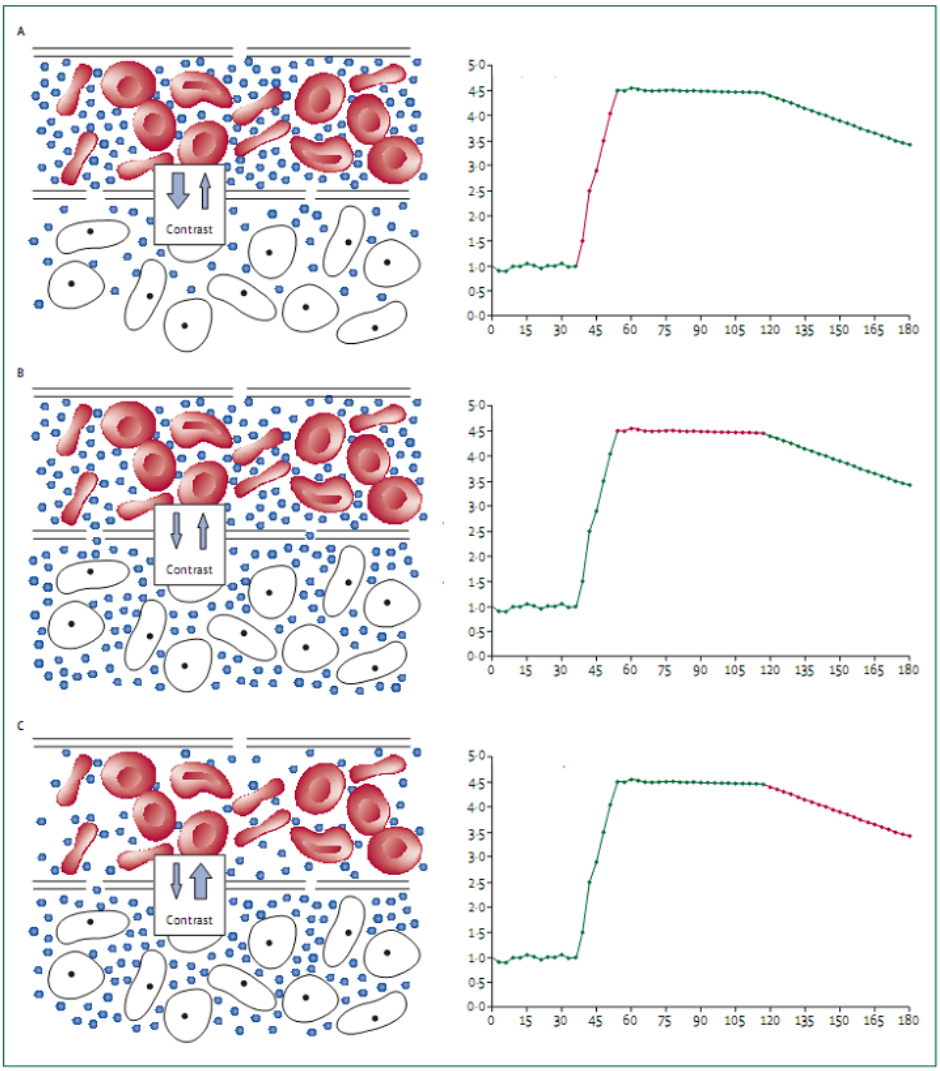
Figur 10: Kontrastmidlet Gd-DTPA (gult) kan befinne seg i blodplasmaet (raudt) og i det ekstracellulære ekstravaskulære rommet EES (kvitt), men ikkje i cellene (blå). Blodplasmaet er den delen av blodet som ikkje er celler. EES er den delen av vevet som ikkje er celler.

Før undersøkinga tek ein eit prekontrastbilete, som brukast som referanse for dei andre bileta, *[4]*. Etter injeksjonen takast ein serie bilete med gitte tidsintervall mellom, som vil skildre opptaket og utvaskinga av kontrastmidlet. Ved å samanlikne bileta med prekontrastbiletet, kan ein sjå korleis signalintensiteten endrar seg. Ein kan då definere relativ signalauke, *RSI*, som endringa i signalintensitet delt på signalintensiteten i prekontrastbiletet,

$$RSI(t_n) = \frac{S(t_n) - S(t_0)}{S(t_0)}$$

der $S(t_n)$ er signalintensiteten ved tida t_n, og $S(t_0)$ er signalintensiteten i prekontrastbiletet. *Figur 11* syner korleis den relative signalauken utviklar seg etter at kontrastmidlet er injisert. Det er først ein opptaksfase, med netto transport av kontrastmiddel frå blodet til EES. Deretter er det ein platåfase, der kontrastmiddelkonsentrasjonen er lik i blodet og i EES, før kontrastmidlet går tilbake til blodet og forsvinn ut av kroppen i utvaskingsfasen.

Under nokre føresetnader (spolert gradientekkosekvens med kort repetisjonstid), vil *RSI* vere proporsjonal med produktet av kontrastmiddelkonsentrasjonen og T_1-tida til vevet, *[25]*. Dersom ein går ut i frå at T_1 er tilnærma konstant i vevet, vil *RSI* vere proporsjonal med konsentrasjonen åleine. Sidan konsentrasjonen er avhengig av eigenskapane til karnettverket i og rundt vokselen, vil den, og såleis *RSI*, variere over svulsten. Ved å finne *RSI* for kvar voksel, får ein eit kart som viser korleis karnettverket i området fungerer. Sidan karnettverket i kreftsvulstar skil seg frå det i friskt vev, kan dette nyttast til å lokalisere svulstar.



Figur 11: Tre fasar for kontrastmidlet i vevet.
A) Opptaksfasen. Netto transport av kontrastmiddel frå blodet (nede) til EES (oppe). B) Platåfase. Lik konsentrasjon av kontrastmiddel i EES og blodet. C) Utvaskingsfasen. Netto transport av kontrastmiddel frå EES til blodet. Grafane har relativ signalauke (RSI) langs y-aksane og tid langs x-aksane. Figur henta frå Zahra et al., [4].

Ein fordel med DCE-MRI framfor vanleg MRI, er at ein ikkje berre kan finne svulsten, men også sjå variasjonar internt i svulsten. Dette kan potensielt nyttast til å skreddarsy behandlinga av svulsten, til dømes ved å gi kraftigare stråledose til enkelte delar av svulsten, *[4]*.

2.2 Stadiar i livmorhalskreft

FIGO, *The International Federation of Gynecology and Obstetrics*, har laga ei klassifisering av livmorhalskreftsvulstar, som fortel kor alvorleg tilfellet er, *[1]*. Denne inndelinga, vist i *tabell 2*, baserer seg på storleiken til svulsten, og på kvar svulsten har spreidd seg. Utfallet av behandlinga er svært avhengig av stadiet før behandling. Til dømes er 89% av pasientar med stadie I i live 5 år etter behandling, men berre 18% av pasientane med stadie IV, *[1]*. I tillegg til dei fem hovudstadia i tabellen, finst det underkategoriar av kvart stadium.

Tabell 2: Dei ulike hovudstadia av livmorhalskreft. 0 er det mildaste tilfellet og IV er det mest alvorlege. Klassifiseringa er definert av FIGO, The International Federation of Gynecology and Obstetrics. Kjelde: Oncolex, [1].

Stadie	Skildring
0	Svulsten er ikkje spreidd til omliggande vev.
I	Svulsten er berre i livmorhalsen.
II	Svulsten er utanfor livmorhalsen, men når ikkje bekkenveggen eller ytre 1/3 av skjeden.
III	Svulsten når bekkenveggen og/eller den ytre 1/3 av skjeden.
IV	Svulsten når utanfor bekkenet, eller infiltrerer endetarm eller blære.

2.3 Kreftsvulstar

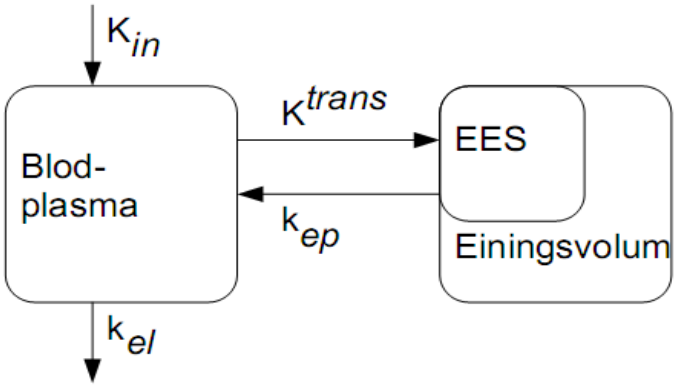
Kreftsvulstar skil seg frå friskt vev ved å ha eit svært kaotisk karnettverk, grunna ein prosess kalla angiogenese, *[5]*. Som alle andre celler, forbrenn kreftceller oksygen. Oksygenet vert frakta til cellene gjennom karnettverket. For at cellene skal overleve, kan dei difor ikkje ligge lengre frå næraste blodkar enn diffusjonslengda til oksygenet. Det vil seie at ein svulst i utgangspunktet ikkje kan ha ein radius på meir enn 100-200 µm. For å kunne vekse vidare, må svulsten danne nye blodkar. Dette kallast angiogenese. Desse nye blodkara har mange forgreiningar og varierande diameter, noko som gjer blodgjennomstrøyminga meir ujamn enn normalt, i tillegg til at dei er meir utsette for lekkasje. I DCE-MRI-undersøkingar vil område med kreftvev ha større lekkasje av kontrastmiddel frå blodbana, og vil difor synast på bileta som område med kraftigare signal. Eit døme på dette er synt i *figur 9*, der svulsten trer mykje tydelegare fram i biletet som er teke med kontrastmiddel enn i biletet utan kontrastmiddel.

Dårlegare blodgjennomstrømning i kreftvevet fører også til at det er mindre oksygen i kreftvev enn i friskt vev. Denne mangelen på oksygen i blodet kallast hypoksi, *[13]*. Vev med hypoksi reagerer mindre på strålebehandling enn oksygenrikt vev. Det er difor mogleg at identifisering av hypoksiske område av ein svulst kan nyttast til å predikere utfallet av behandlinga, sjå til dømes Cooper et al., *[11]*, og Loncaster et al., *[12]*.

2.4 Farmakokinetiske modellar

Farmakokinetikk skildrar, ved hjelp av matematiske modellar, korleis legemiddel bevegar seg gjennom kroppen, *[6]*. Ved å dele svulsten opp i fleire delar, tilsvarande vokslane i MRI-undersøkinga, kan ein få informasjon om korleis svulsten ser ut inni, ikkje berre kor stor den er og kvar den ligg. Resultata frå DCE-MRI-undersøkingar nyttast til å estimerere parameterane i modellen som vert nytta, *[9]*.

Ein svulst består av blodplasma, den delen av blodet som ikkje er celler, *[26]*, celler, og området mellom cellene (EES), *[4]*. For å skildre fordelinga av kontrastmiddel i kvar del av svulsten, kan ein difor tenkje seg å nytte ein treromsmodell, det vil seie ein modell der kontrastmidlet kan opphalde seg på tre ulike stader, cellene, blodet og området mellom cellene, samt overførast mellom desse. Sidan Gd-DTPA ikkje kan trenge inn i celler, er det tilstrekkeleg med ein toromsmodell, der blodplasma er eitt rom, og EES og cellene er det andre, *[8]*. Gd-DTPA kan overførast mellom blodplasma og EES. *Figur 12* syner denne toromsmodellen. Einingsvolmet består av cellene og EES, området mellom cellene. Ratekonstanten K^{trans} gjeld overføring frå blodet til EES, medan k_{ep} er ratekonstanten for transport den andre vegen.



Figur 12: Toromsmodell for distribusjon av kontrastmiddel. Dei to romma er blodplasma (til venstre) og einingsvolumet (til høgre). Einingsvolumet består av cellene og området mellom dei, det ekstracellulære ekstravaskulære rommet EES. K^{trans} er overføringsraten av kontrastmiddel frå blodplasma til EES, medan k_{ep} er overføringsraten frå EES til blodplasma. K_{in} er infusjonsraten for kontrastmidlet, medan k_{el} er utvaskingsraten av kontrastmiddel frå blodplasma.

Vidare går ein ut i frå at konsentrasjonen av kontrastmiddel i både EES og plasma er homogen, og at *fast exchange*-kravet er oppfylt, det vil seie at alle vassmolekyl i vevet har lik tilgang til kontrastmidlet, *[23]*. Ein tenkjer seg at kontrastmiddel berre kan overførast mellom plasma og EES, med andre ord at det ikkje kan gå frå EES i eit delvolum til EES i eit anna, *[8]*. Konsentrasjonen i EES vil då variere på grunn av utveksling med blodplasma, medan blodplasmaet er eit reservoar der konsentrasjonen varierer grunna transport gjennom karnettverket.

Samanhengen mellom dei to ratekonstantane K^{trans} og k_{ep} er gitt ved

$$K^{trans} = k_{kep} v_e \qquad (3)$$

der v_e er volumfraksjonen av EES i einingsvolumet, det vil seie kor stor del av einingsvolumet som ikkje er celler. Ei oversikt over dei ulike storleikane nytta i farmakokinetiske modellar i denne oppgåva er gitt i *tabell 3*.

Tabell 3: Storleikar nytta i farmakokinetiske modellar for kontrastmiddel i EES og blodplasma. Kjelder: Tofts et al., [27], og Brix et al., [8].

Namn	**Eining**	**Skildring**
K^{trans}	min^{-1}	Overføringskonstant for kontrastmiddel frå blodplasma til EES.
k_{ep}	min^{-1}	Ratekonstant for overføring av kontrastmiddel frå EES til blodplasma.
C_t	mM	Konsentrasjon av kontrastmiddel i vevet (einingsvolumet).
C_e	mM	Konsentrasjon av kontrastmiddel i EES.
C_p	mM	Konsentrasjon av kontrastmiddel i blodplasma.
k_{el}	min^{-1}	Ratekonstant for eliminasjon av kontrastmiddel frå blodplasma.
K_{in}	mol min^{-1}	Infusjonsrate for kontrastmiddel.
V_p	ml	Totalt blodplasmavolum.
V_e	ml	Totalt EES-volum.
V_t	ml	Totalt einingsvolum.
v_e	-	Volumfraksjon av EES i einingsvolumet. $v_e = \frac{V_e}{V_t}$
t_{in}	min	Infusjonstid for kontrastmidlet.

Sidan kontrastmidlet berre kan vere i EES og ikkje i cellene, vil samanhengen mellom konsentrasjonen C_t av kontrastmiddel i einingsvolumet og konsentrasjonen C_e av kontrastmiddel i EES vere gitt ved

$$C_t = C_e v_e$$

Utviklinga av konsentrasjonen i einingsvolumet i tida etter injeksjonen av kontrastmiddel er gitt av Tofts et al., *[27]*, som

$$\frac{dC_t}{dt} = K^{trans}(C_p - C_e)$$

der C_p er konsentrasjonen av kontrastmiddel i blodplasma.

Ved å utnytte at $K^{trans} = v_e k_{ep}$ og at $C_t = v_e C_e$, kan dette omformast til

$$\frac{dC_t}{dt} = K^{trans} C_p - k_{ep} C_t \qquad (4)$$

Dersom ein set konsentrasjonen C_t ved tida $t = 0$ til å vere null, det vil seie at det ikkje er kontrastmiddel i vevet ved $t = 0$, har likning (4) løysinga

$$C_t(t) = K^{trans} \int_0^t C_p(t') e^{-k_{ep}(t-t')} dt'$$

2.4.1 Brix-modellen

Brix et al. (1991), *[8]*, har utvikla ein toromsmodell for DCE-MRI med Gd-DTPA som kontrastmiddel. Denne modellen er i utgangspunktet laga for lang infusjonstid av kontrastmiddel, 4 min i studien som artikkelen frå 1991 byggjer på. Undersøkinga i denne oppgåva er utført med bolusinjeksjonar, det vil seie at alt kontrastmidlet sprøytast inn på ein gong. Modellen må difor tilpassast slike raske injeksjonar.

Endringa i konsentrasjon av kontrastmiddel i dei to romma, blodet og EES, sjå *figur 12*, kan skildrast ved differensiallikningane

$$\frac{dC_p}{dt} = \frac{K_{in}}{V_p} - (K^{trans} + k_{el}) C_p + k_{ep} \frac{V_e}{V_p} C_e \qquad (5)$$

$$\frac{dC_e}{dt} = K^{trans} \frac{V_p}{V_e} C_p - k_{ep} C_e \qquad (6)$$

Forklaring av dei ulike storleikane er gitt i *tabell 3*.

Ein går ut i frå at $V_p >> V_t$, det vil seie at den delen av einingsvolumet som kontrastmidlet oppheld seg i, er svært liten samalikna med heile blodvolumet. C_p vil difor ikkje endrast særleg av overføringar av kontrastmiddel til og frå EES. Ein kan difor sjå vekk frå ledda $K^{trans} C_p$ og $k_{ep} \frac{V_e}{V_p} C_p$ i likning (5).

Likninga kan difor skrivast som

$$\frac{dC_p}{dt} = \frac{K_{in}}{V_p} - k_{el} C_p$$

Dersom ein også føreset at permeabiliteten er lik for transport i begge retningar, vil samanhengen mellom K^{trans} og k_{ep} vere gitt ved

$$K^{trans} V_p = k_{ep} V_e$$

Ved å nytte dette, samt samanhengen $C_t = v_e C_e$, kan likning (6) omformast til

$$\frac{dC_t}{dt} = v_e k_{ep} C_p - k_{ep} C_t$$

Desse likningane løysast under føresetnad om at begge startkonsentrasjonane er null, $C_t(0) = 0$ og $C_p(0) = 0$, og får

$$C_p(t) = \frac{K_{in}}{V_p k_{el}} (e^{-k_{el} t_{in}} - 1) e^{-k_{el} t} \qquad (7)$$

og

$$C_t(t) = \frac{K_{in} v_e}{V_p} \frac{k_{ep}}{k_{ep} - k_{el}} \left(\frac{e^{k_{el} t_{in}} - 1}{k_{el}} e^{-k_{el} t} - \frac{e^{k_{ep} t_{in}} - 1}{k_{ep}} e^{-k_{ep} t} \right) \qquad (8)$$

der t_{in} er infusjonstida, det vil seie den tida det tek å injisere kontrastmidlet, og t er tid etter at injeksjonen er ferdig. For raske injeksjonar, som er nytta i vår undersøking, vil, t_{in} vere svært kort, og vi kan nytte tilnærmingane

$$e^{k_{ep} t_{in}} \approx 1 + k_{ep} t_{in} \quad \text{og} \quad e^{k_{el} t_{in}} \approx 1 + k_{el} t_{in}$$

Likingane (7) og (8) reduserast då til

$$C_p(t) = \frac{K_{in} t_{in}}{V_p} e^{-k_{el} t} \qquad (9)$$

og

$$C_t(t) = \frac{K_{in} v_e t_{in}}{V_p} \frac{k_{ep}}{k_{ep} - k_{el}} (e^{-k_{el} t} - e^{-k_{ep} t}) \qquad (10)$$

For gradientekkosekvensar med kort repetisjonstid, kan den relative signalauken skildrast som

$$RSI(t) = \frac{C_t(t)}{\tau_1} T_{1,0}$$

der $1/\tau_1$ er spinn-gitter-relaksasjonsraten til kontrastmidlet og $T_{1,0}$ er T_1-tida for vevet utan kontrastmiddel, *[25]*.

Set utrykket for $C_t(t)$ inn som konsentrasjonen i vevet og får

$$RSI(t) = \frac{T_{1,0}}{\tau_1}\frac{K_{in}v_e t_{in}}{V_p}\cdot\frac{k_{ep}}{k_{ep}-k_{el}}(e^{-k_{el}t}-e^{-k_{ep}t})$$

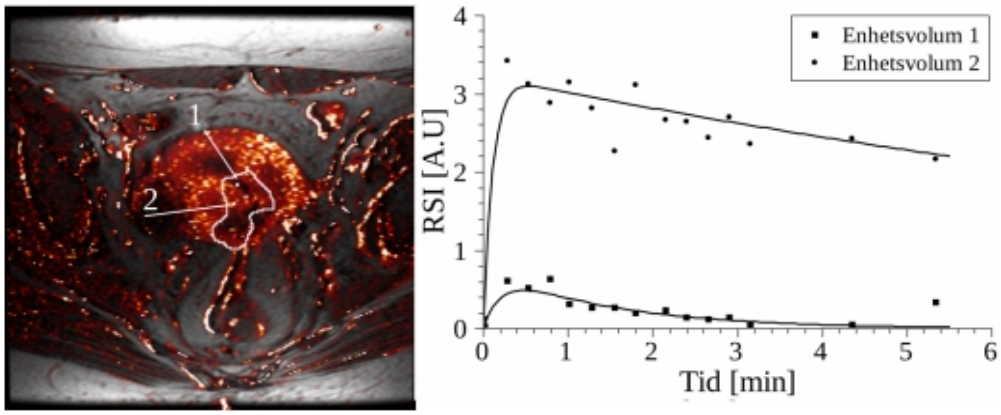
eller

$$RSI(t) = \frac{A\,k_{ep}}{k_{el}-k_{ep}}(e^{-k_{ep}t}-e^{-k_{el}t}) \qquad (11a)$$

der

$$A = T_{1,0}-\tau_1\frac{K_{in}v_e t_{in}}{V_p} \qquad (11b)$$

A kallast amplituden. Denne innheld informasjon om T_1-relaksasjonstida til vevet ($T_{1,0}$), T_1-relaksasjonstida til kontrastmidlet (τ_1), om konsentrasjonen av kontrastmiddel i blodplasma-reservoaret ($\frac{K_{in}t_{in}}{V_p}$) og om andelen celler (v_e).



Figur 13: RSI for to ulike einingsvolum er tilpassa Brix-modellen. Verdiar av A, k_{ep} og k_{el} for kvart einingsvolum er rekna ut på grunnlag av sjølve vokselen samt fire nabovokslar, jamfør figur 16. MR-biletet er sett saman av prekontrastbiletet og RSI-verdiane for DCE-MRI-serien. Dess lysare fargen er, dess større er RSI. Svulsten er innteikna med kvitt omriss. Henta frå Erlend Andersen, [10].

I DCE-MRI-undersøkinga måler ein den relative signalauken. Denne tilpassast så Brix-modellen ved å finne A, k_{ep} og k_{el} for kvar voksel. *Figur 13* syner resultatet av ei slik tilpassing for to ulike einingsvolum. Tilpassinga blir gjort på grunnlag av målingane frå den aktuelle vokselen, samt fire nabovokslar. Meir om dette i *avsnitt 3.2*.

Dei tre parameterane skildrar altså dei biologiske eigenskapane til svulsten. A gjev relaksasjonstider, kontrastmiddelkonsentrasjon og celletettleik. k_{ep} syner permeabiliteten til vevet, sidan den er overføringsraten av kontrastmiddel frå EES til blodplasma. Blodgjennomstrøyminga, gitt ved k_{el}, er eliminasjonsraten av kontrastmiddel frå blodet.

Det finst også andre modellar som kan brukast til å skildre opptaket av kontrastmiddel i svulsten. RR-modellen (*Reference Region*), utvikla av Yankeelov et al. (2005), *[9]*, samanliknar vevet ein undersøkjer med eit referansevev. Tidlegare analysar, *[10]*, indikerer at Brix-modellen truleg er best eigna for å knyttast til behandlingsutfall. Denne oppgåva begrensar seg difor til Brix-modellen.

3 Materiale og metodar

3.1 Programvare

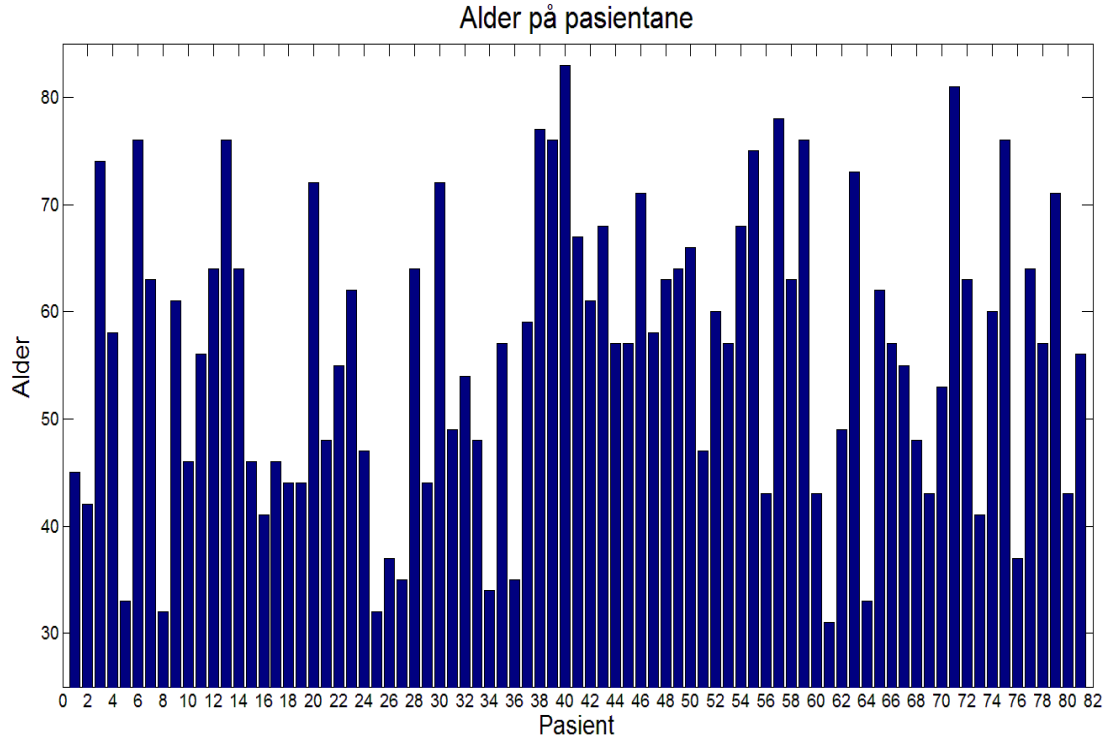
Matlab (versjon R2011b, Mathworks, Natick, Massachusetts, USA) med tilleggspakken Statistics Toolbox (versjon 7.6) nyttast til å lese inn og sortere data, samt til deskriptiv statistikk (meir om dette i kapitla *3.2* og *3.3*). Nokre av skripta som er nytta finnast som vedlegg til denne oppgåva. Til den multivariate analysen nyttast PLS_Toolbox (versjon 6.2.1, Eigenvector Research, Wenatchee, Washington, USA) saman med Matlab, samt programmet The Unscrambler X (versjon 10.1, Camo, Oslo, Noreg).

3.2 Datasettet

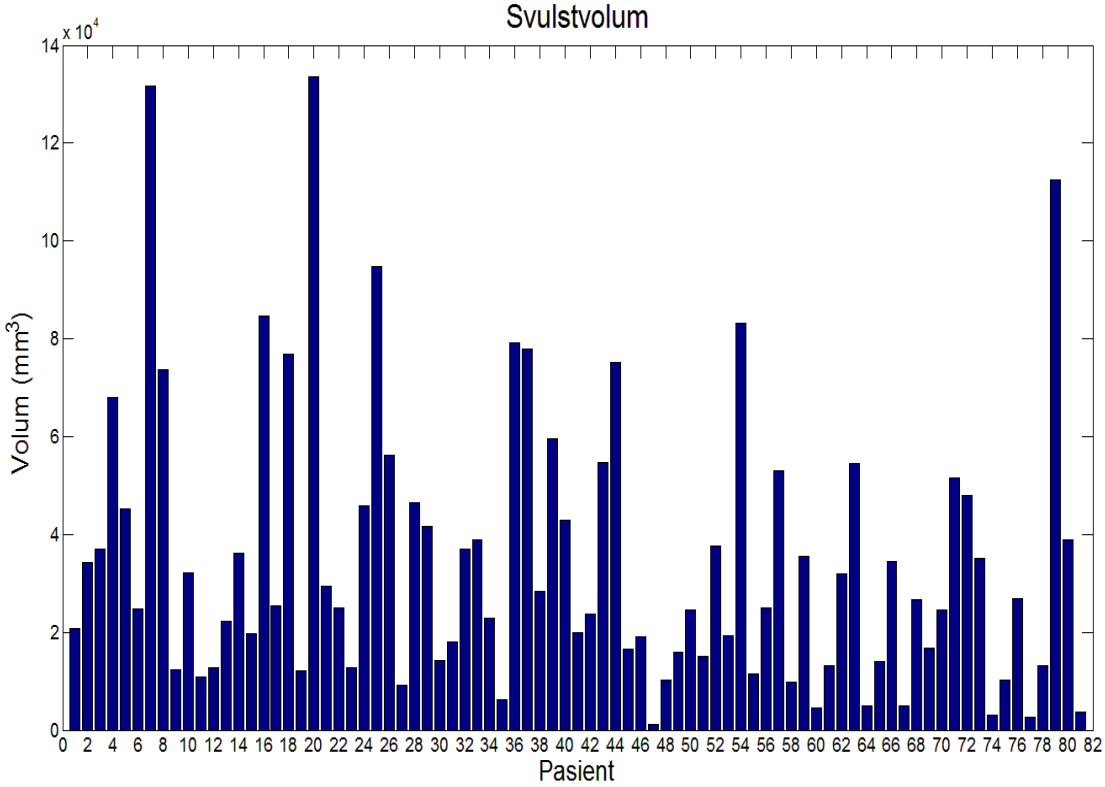
Datagrunnlaget for analysen er DCE-MRI-bilete frå 81 pasientar med livmorhalskreft. Pasientane var undersøkte i tida 2001 til 2004 ved Det Norske Radiumhospitalet (no ein del av Oslo universitetssykehus), *[14]*. Den opprinnelege undersøkinga var av 88 pasientar, men sju av dei har blitt ekskluderte frå studiet grunna dårleg biletkvalitet, *[10]*.
Gjennomsnittsalderen for pasientane var 56 år, og dei hadde ulike stadiar av kreftsjukdom då undersøkinga vart gjennomført, som synt i *tabell 4*. Ei oversikt over alder og svulstvolum for kvar av pasientane er gitt i *figur 14* og *figur 15*. Etter bileta vart teke, har alle pasientane har fått behandling i form av stråleterapi for sjukdommen, og deretter jamnleg oppfølging.

Tabell 4: Fordeling av stadie (I-IV) før behandling. Det er totalt 81 pasientar.

Stadie	Tal pasientar	Andel pasientar
I	2	2,5%
II	44	54,3%
III	29	35,8%
IV	6	7,4%

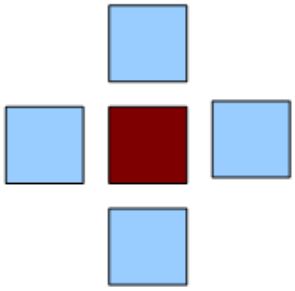


Figur 14: Alderen til kvar av dei 81 pasientane på undersøkingstidspunktet. Gjennomsnittsalderen er 56 år. Laga med Matlab.



Figur 15: Svulstvolumet (mm^3) for kvar av dei 81 pasientane. Gjennomsnittsvolumet er 34 913 mm^3. Laga med Matlab.

Undersøkinga omfattar eit prekontrastbilete, samt ein aksial T_2-vekta serie og ein dynamisk kontrastforsterka (DCE) serie. Til DCE-serien er kontrastmidlet Gd-DTPA nytta. Etter injeksjonen tok ein bilete kvart 15. sekund i 3 minutt, og så to bilete med 1 minutts mellomrom, til saman 14 bilete. MR-bileta er delte inn i volumelement, vokslar, med storleik 0,78 mm x 0,78 mm x 5 mm. Tjukkelsen på kvart snitt er altså 5 mm.



Figur 16: Ein voksel (raud) med fire naboar (blå).

Den T_2-vekta serien nyttast til å definere svulstvolumet, det vil seie å bestemme kvar grensa mellom svulsten og det friske vevet går, sjå *figur 13*. Data frå DCE-serien er så tilpassa Brix-modellen, gitt i likning (11b), ved å berekne parameterane A, k_{ep} og k_{el}, sjå *tabell 5*, for kvar voksel, *[10]*. Eit døme på slik tilpassing er synt i *figur 13*. Verdiane er berekna på grunnlag av denne vokselen og dei fire næraste nabovokslane, som vist i *figur 16*. Dette er gjort for å minske støymengda i data, *[14]* .

Tabell 5: Parameterar for kvar voksel, rekna ut i frå gjennomsnittet av denne vokselen og dei fire nabovokslane, som vist i figur 16.

Parameter	**Forklaring**
x	Vokselen sin x-koordinat.
y	Vokselen sin y-koordinat.
z	Vokselen sin z-koordinat.
A	Konsentrasjon av kontrastmiddel i vokselen.
k_{ep}	Overføringsrate av kontrastmiddel til vokselen.
k_{el}	Utvaskingsrate av kontrastmiddel frå vokselen.
A_stddev	Standardavvik for A over dei fem vokslane.
k_{ep}_stddev	Standardavvik for k_{ep} over dei fem vokslane.
k_{el}_stddev	Standardavvik for k_{el} over dei fem vokslane.
chisqr	Kjikvadratverdi for tilpassing til Brix-modellen.
fit_status	Mål for resultat av iterasjonen for tilpassing til Brix-modellen.Verdiane 1-4 indikerer at iterasjonen har konvergert, medan 5 syner at ein har nytta maksimalt tal på iterasjonar utan å oppna konvergens

I tillegg til desse modellparameterane, innheld datasettet koordinatar (*x*,*y*,*z*) for kvar voksel, samt kjikvadratverdi for tilpassinga til Brix-modellen og verdien *fit_model* som indikerer resultatet av iterasjonen for tilpassinga til Brix-modellen for denne vokselgruppa. *fit_status* = 1 - 4 vil seie at iterasjonen har konvergert innanfor ønska toleranse, medan *fit_status* = 5 betyr at ein har nådd maksimalt tal på iterasjonar utan å kome innanfor ønska toleranse.

Utfallet av behandlinga er gitt i datasettet ved storleikane progression_free_survival (*pfs*, progresjonsfri overleving) og locoregional_control (*lc*, lokal kontroll). Progresjonsfri overleving syner om pasienten er frisk (*pfs* = 0) eller har fått tilbakefall (*pfs* = 1). Tilbakefallet kan vere lokalt, det vil seie at sjølve svulsten veks, eller i form av metastasar, det vil seie spreiing til andre delar av kroppen. Lokal kontroll, *lc*, er ein underkategori av progresjonsfri overleving, og fortel om tilbakefallet er lokalt (*lc* = 1) eller ikkje (*lc* = 0), *[28]*.Talet på pasientar med dei ulike utfalla er gitt i *tabell 6*. Tabellen syner at mange av pasientane, 23 av 32, som har fått tilbakefall, ikkje har fått lokalt tilbakefall, med andre ord at metastasar er den vanlegaste forma for tilbakefall blant desse pasientane.

Tabell 6: Talet på pasientar for kvart utfall. Pfs står for progresjonsfri overleving, lc står for lokal kontroll. Pasientar med pfs = 0 er blitt friske att, og pasientar med pfs = 1 har fått tilbakefall. Dersom pasientane med tilbakefall har lc = 0, har dei fått tilbakefall i form av metastasar, medan lc = 1 syner at dei har fått lokalt tilbakefall. Det er 81 pasientar totalt.

	pfs = 0	**pfs = 1, lc = 0**	**pfs = 1, lc = 1**
Tal pasientar	49	23	9

Datasettet er lagra i ei tekstfil. Fila inneheld informasjon om pasientane, samt Brix-parameterane. Data for kvar pasient avsluttast med linja #EOP, medan sjølve fila endar med #EOF. Dei første linjene for kvar pasient, det vil seie headeren, inneheld namn, figo-status, fødselsår, progresjonsfri overleving, lokal kontroll, lymfeknuteinfiltrasjon, undersøkingsdato og talet på fridomsgrader i kjikvadrattesten for tilpassinga, sjå *tabell 7*. Lymfeknuteinfiltrasjon vil seie at at det er kreftceller i lymfeknutane. Progresjonsfri overleving, lokal kontroll og lymfeknuteinfiltrasjon er binære variablar.

Tabell 7: Informasjon gitt for kvar pasient.

Parameter	**Forklaring**
name	MM-nummeret til pasienten. Dette nummeret fortel kor tid pasienten vart med i studiet.
FIGO	Viser kva stadie svulsten hadde då bileta vart teke. Fem hovudnivå (0 – IV). Jamfør *tabell 2*.
birthyear	Fødselsdatoen til pasienten.
study_date	Undersøkingsdato.
progression_free_survival	1 dersom pasienten er frisk, 0 dersom tilbakefall.
locoregional_control	1 for lokalt tilbakefall, 0 for metastasar.
Lymphnode_infiltration	2 dersom det er funne kreftceller i lymfeknutar, 1 elles.
Degrees_of_freedom	Fridomsgraden for kjikvadrattest av tilpassing.

Eit døme på header for ein pasient er synt nedanfor:

#name MM001
#FIGO 2b
#birthyear[DDMMYYYY] 27021956
#progression_free_survival 0
#locoregional_control 0
#Lymphnode_infiltration[1=True] 1
#study_date[DDMYYYY] 27022001
#Degrees_of_freedom 12
 # X Y Z A kep kel A_stddev kep_stddev kel_stddev chisqr fit_status

Etter headeren følgjer Brix-parameterane for pasienten. Desse er ordna i èi linje per voksel, der kvar linje inneheld tal for dei 11 parameterane:
X Y Z A kep kel A_stddev kep_stddev kel_stddev chisqr fit_status

Eit Matlab-skript, sjå vedlegget *kapittel 7.1*, nyttast til å lese inn data frå tekstfila. I tillegg bereknast pasienten sin alder då undersøkinga fann stad, og volumet av svulsten. Volumet finnast ved å telje kor mange vokslar svulsten består av, og multiplisere dette med volumet 3,042 mm^3 av kvar voksel.

Vokslar med unormalt høge verdiar av èin eller fleire av parameterane A, k_{ep} og k_{el} fjernast før den vidare analysen, slik at variablane får desse grensene:

$$0 \leqslant A \leqslant 10 \quad , \quad 0 \leqslant k_{ep} \leqslant 12 \quad , \quad 0 \leqslant k_{el} \leqslant 0{,}5$$

Dette fordi dei høge verdiane kjem av at modellen bryt saman i enkelte vokslar, og gir urealistiske verdiar, *[28]*. Ein kan også sjå dette av storleiken *fit_status*. Mange av vokslane med for høge verdiar av A, k_{ep} og/eller k_{el} har *fit_status* = 5, det vil seie at modelltilpassinga ikkje har konvergert innan maksimalt tal på iterasjonar.

3.3 Deskriptiv statistikk

Analysane er utførte med deskriptive statistiske parameterar berekna frå A, k_{ep} og k_{el}, sjå *tabell 8*. Dette for å enklare kunne samanlikne pasientar med ulik svulststorleik, og for å kunne tolke resultata vi får.

Dei statistiske parameterane reknast ut ved hjelp av eit skript laga i Matlab, sjå *kapittel 7.1* i vedlegget. Pasientinformasjonen samt dei statistiske parameterane lagrast i ei matrise og eksporterast til Unscrambler, som nyttast til fleire av dei multivariate analysane. Meir om programvaren i *kapittel 3.1*.

Tabell 8: Deskriptiv statistikk nytta på A, k_{ep} og k_{el}.

Statistisk parameter	**Forklaring**
mean	Gjennomsnittsverdi
median	Medianverdi (dvs verdien i midten når data er sortert etter stigande/synkande verdi)
mode	Vanlegaste verdi
min	Minste verdi
max	Største verdi
std	Standardavvik
skew	Skeivskap. Fortel kor symmetrisk data ligg kring gjennomsnittsverdien.
kurt	Kurtose. Fortel kor spiss fordelinga av verdiane er.
10, 20, 25, 30, 40, 50, 60, 70, 75, 80, 90	Persentilar.
75-25	Differansen mellom 75%-persentilen og 25%-persentilen.
90-10	Differansen mellom 90%-persentilen og 10%-persentilen.

Gjennomsnitt, median og vanlegaste verdi er mål for senter av data, medan standardavviket og persentilane skildrar spreiinga i data.

Skeivskap definerast som

$$skew = \frac{E(x-\mu)^3}{\sigma^3}$$

[29], der x er variabelen, μ er gjennomsnittet av x, σ er standardavviket til x og $E(x-\mu)^3$ er forventingsverdien til $(x-\mu)^3$. Denne storleiken fortel kor skeivt eit histogram over data er. Eit normalfordelt histogram vil ha skeivskap null. Skeivskap større enn null vil seie at histogrammet hellar mot høgre, medan histogram med skeivskap mindre enn null hellar mot venstre.

Kurtose definerast i Matlab som

$$kurt = \frac{E(x-\mu)^4}{\sigma^4}$$

[29], og fortel om forma til histogrammet over data. Dersom kurtosen er 3, er histogrammet forma som normalfordelingskurva. Større verdiar vil seie at histogrammet er spissare, medan mindre verdiar vil seie at det er mindre spist, *[30]*.

3.4 Histogram

Fordelinga av Brix-parameterane over vokslane kan også skildrast ved hjelp av histogram. Det vil seie at ein deler data inn i *N* intervall. Breidda av kvart intervall bør veljast slik at histogrammet syner variasjonen i data best mogleg, utan at det er for mange grupper. I *figur 17* ser vi tre ulike histogram over dei same observasjonane. I det første, *figur 17a*, som har svært korte intervall, ser det ut som det er meir variasjon i observasjonane enn det som faktisk er tilfelle. Histogrammet med få intervall, *figur 17c*, mistar derimot noko av variasjonen. Det beste kompromisset vil vere histogrammet synt i *figur 17b*. Her får ein fram fordelinga utan å bli for mykje eller for lite detaljert. Det finst ei rekkje reglar for å finne den optimale breidda. Scott, *[31]*, foreslår å velje breidda

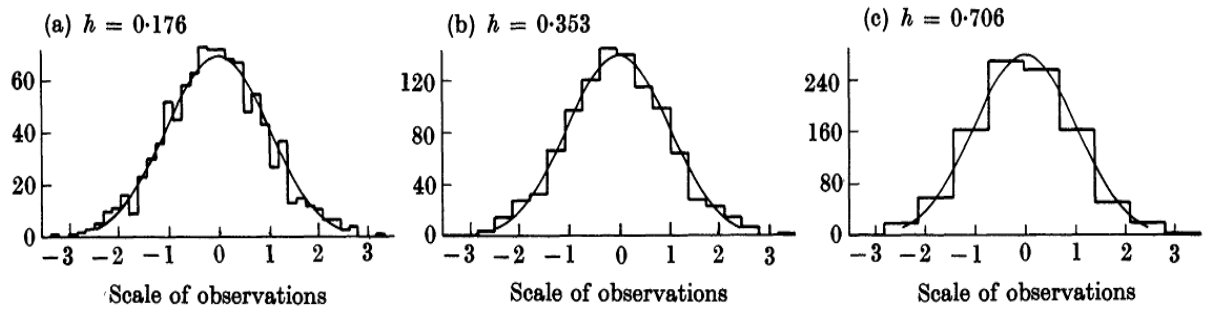
$$h = 3{,}49\, s\, n^{-\frac{1}{3}}$$

der h er breidda, s er eit estimat av standardavviket og n er talet på observasjonar.

Eit meir robust alternativ til denne regelen vart foreslått av Freedman og Diaconis, *[32]*, som seier

$$h = 2\, IQR\, n^{-\frac{1}{3}}$$

der *IQR* er *inter-quartile-range*, det vil seie avstanden mellom 75%-persentilgrensa og 25%-persentilgrensa. Dette målet er meir robust, sidan *IQR* er mindre sensitiv for ekstreme observasjonar enn det standardavviket er. Freedman- og Diakonis-regelen vert difor nytta for histogram i denne oppgåva. Desse reglane går i ut i frå at alle intervalla er like lange, men ein kan også velje å ha ulike storleikar på intervalla.



Figur 17: Histogram med ulik intervallbreidde. Observasjonane ligg mellom -3 og 3, og følgjer fordelinga gitt ved den heiltrukne linja. Intervallbreiddene er a) 0,176, b) 0,353 og c) 0,706. Henta frå Scott, [31].

Eit histogram kan også normaliserast. Då deler ein talet på observasjonar i kvar gruppe på totalt tal observasjonar. Arealet under histogrammet vil då vere 1. Dette er ein fordel dersom ein skal samanlike histogram med ulikt tal observasjonar. Ettersom talet på observasjonar i denne oppgåva er avhengig av storleiken på kvar svulst, vert histogramma normaliserte.

3.5 Statistiske metodar

Datasettet innheld mange variablar, noko som gjer det eigna til multivariat analyse. Metodane nytta i samband med denne oppgåva, er prinsipalkomponentanalyse (PCA), diskriminant analyse, SIMCA, SVM, regresjon, klyngeanalyse og *partial least sqares* (PLS). Desse vert presenterte i dei neste kapitla. Sjå til dømes Johnson og Wichern, *[16]*, eller Esbensen, *[33]*, for utfyllande informasjon om multivariat analyse.

Data kan delast inn i forklaringsvariablar (X) og responsvariablar (Y), der vi ønskjer å forklare responsvariablane ved hjelp av forklaringsvariablane. Her består Y av variabelen som indikerer behandlingsutfall, det vil seie progresjonsfri overleving (*pfs*). X er dei statistiske parameterane til A, k_{ep} og k_{el}, samt alder, volum og stadie, til saman 68 forklaringsvariablar. Dersom histogramverdiar nyttast, vert desse X. I matrisene X og Y er det èi rad per pasient og èi kolonne per variabel, som vist nedanfor. Variablane volum og stadie kan også nyttast som responsvariablar.

$$X = \begin{bmatrix} . & Alder & Volum & Stadie & A_{mean} & A_{median} & \dots \\ pasient\ 1 & . & . & . & . & . & \dots \\ pasient\ 2 & . & . & . & . & . & \dots \\ \dots & . & . & . & . & . & \dots \end{bmatrix}$$

3.5.1 Prinsipalkomponentanalyse (PCA)

Målet med prinsipalkomponentanalyse er å erstatte forklaringsvariablane med (færre) nye ortogonale variablar kalla prinsipalkomponentar, *[16]*. Prinsipalkomponentane (PC) er lineærkombinasjonar av dei opprinnelege forklaringsvariablane. Dersom det er korrelasjon mellom forklaringsvariablane, vil variasjonen i data kunne skildrast med færre prinsipalkomponentar enn opprinnelege variablar. PCA ser berre på X-variablar, ikkje på responsvariablar. Ein kan likevel sjå etter grupperingar av variablar observasjonar i resultata av prinsipalkomponentanalysen, og undersøkje om desse samsvarar med dei ulike utfalla.

Før analysen vert variablane sentrerte og standardiserte,

$$Z_i = \frac{X_i - \bar{X}_i}{\sigma_i} \qquad (12)$$

der Z_i er den standardiserte verdien, X_i er den originale verdien, $\overline{X_i}$ er gjennomsnittsverdien for variabelen X_i, og σ_i er standardavviket til X_i. Alle Z_i vil ha forventningsverdi 0 og standardavvik 1, noko som gjer at ein enklare kan samanlikne dei ulike variablane, uavhengig av talverdi og varians.

Det finst ulike algoritmar for å finne prinsipalkomponentane. Den enklaste, som brukast på mindre datasett som ikkje manglar data, *[33]*, baserer seg på singulærverdidekomposisjon av kovariansmatrisa. Ved hjelp av singulærverdidekomposisjon, SVD, kan kovariansmatrisa Σ uttrykkast som

$$\Sigma = \sum_{k=1}^{p} \lambda_k \boldsymbol{v_k} \boldsymbol{v_k}^t$$

der λ_k er eigenverdiane til matrisa, og $\boldsymbol{v_k}$ er dei tilhøyrande eigenvektorane.

Prinsipalkomponentane (skårar, *scores*) T_k vert då

$$T_k = \boldsymbol{v_k}^t X$$

der X er datamatrisa. Merk at det ikkje vil vere kovarians mellom dei ulike prinsipalkomponantane,

$$Cov(T_i, T_k) = 0$$

Alle prinsipalkomponentane vil såleis vere ortogonale. Variansen til kvar prinsipalkomponent vil vere den tilsvarande eigenverdien til kovariansmatrisa.

$$Var(T_k) = \lambda_k$$

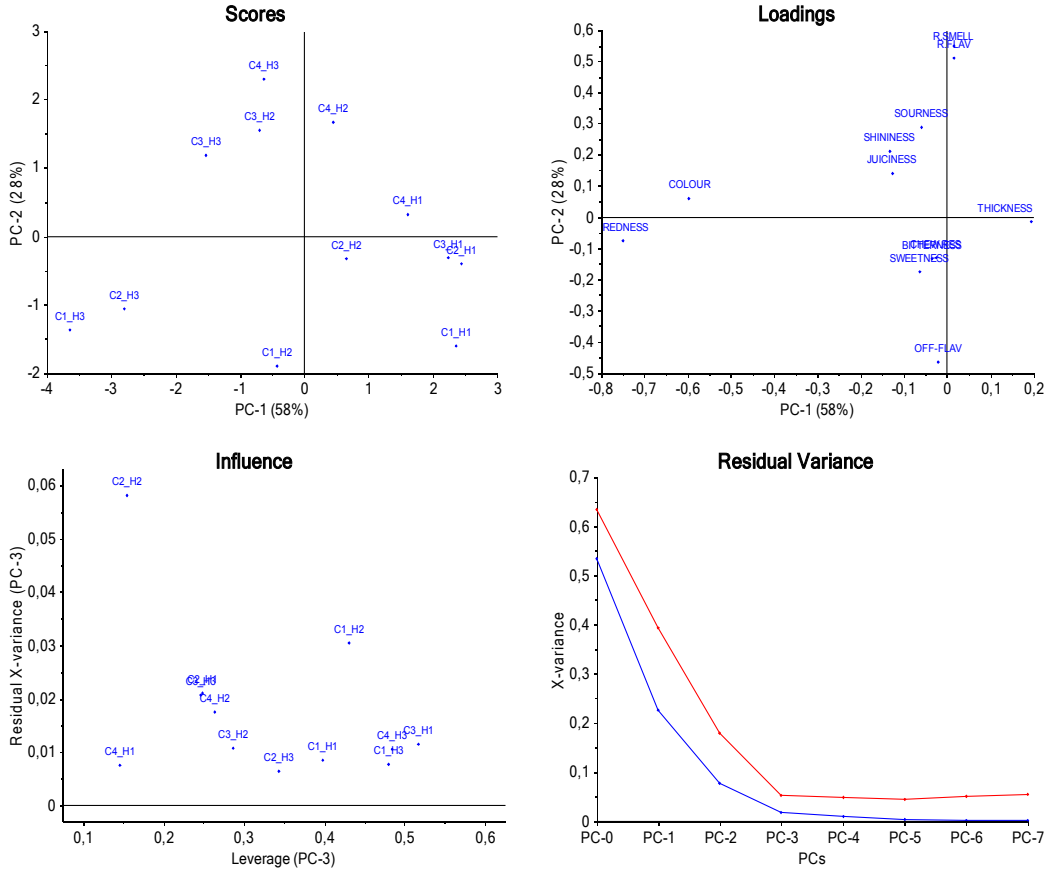
Denne algoritmen finn alle prinsipalkomponentar i same utrekning, og er den mest nøyaktige. Den kan derimot ikkje brukast på datasett som manglar verdiar. Om ein har store datasett og berre er ute etter dei første komponantane, kan algoritmar som finn ein og ein komponent passe, til dømes NIPALS, *[33]*.

NIPALS (*Non-linear Iterative Partial Least Squares*) er den andre algoritmen tilgjengeleg i Unscrambler. Med denne metoden itererer ein seg fram til prinsipalkomponentane, og stoppar når ein har funne så mange komponentar som ein ønskjer. NIPALS fungerer også på datamatriser som manglar data, *[33]*.

Resultatet av PCA er eit sett med skårar og ladningar. Skårar er koordinatar for kvart objekt, i vårt tilfelle pasientar, i det nye prinsipalkomponent-koordinatsystemet. Ladningane syner korleis kvar av dei opprinnelege variablane bidreg til dei nye prinsipalkomponentane. I tillegg finn ein forklart varians for kvar komponent, det vil seie kor stor del av variasjonen i data som forklarast av denne komponenten. Dette kan uttrykkast i likninga

$$X = TP^t + E$$

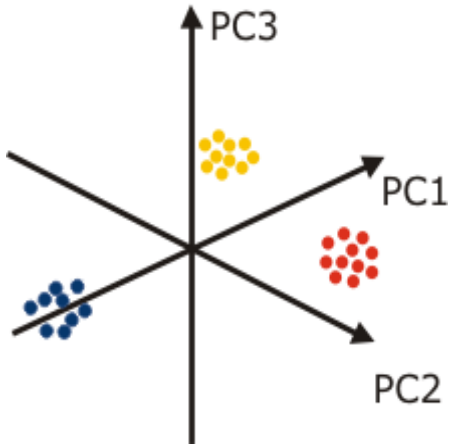
der X er datasettet, T er skårmatrisa, P er ladningsmatrisa og E er residualmatrisa.



Figur 18: Resultat av PCA i Unscrambler. Øvst til venstre: Skår-plott med PC-1 og PC-2. Øvst til høgre: Ladningsplott med PC-1 og PC-2. Nederst til venstre: Momentplott for høgste tilrådde prinispalkomponent, her PC-3. Nederst til høgre: Plott av residualvarians. Den blå kurva gjeld kalibrering, medan den raude gjeld validering.
Laga med Unscrambler (Tutorial B).

Den enklaste måten å undersøkje resultata på, er å plotte dei. Ein kan då sjå etter grupperingar av observasjonar og/eller variablar, og finne ut kva variablar som har stor eller liten påverknad. Ei grundig innføring i tolking av PCA-resultat er gitt i Esbensen, *[33]*.

Skårplott syner objekta og nyttar prinispalkomponentane som aksar, som synt øvst til venstre i *figur 18*. Objekt som ligg nærme kvarandre i desse plotta, er like i høve til dei prinispalkomponentane dei er plotta for. Ein kan såleis bruke skårplott til å sjå etter objekt-grupperingar, som ikkje kom tydeleg fram før prinispalkomponentanalysen. *Figur 19* syner eit døme på svært tydelege grupperingar av objekt. Skårplott kan også avsløre objekt som skil seg mykje frå dei andre. Desse verdiane bør undersøkjast for å sjå om det har skjedd ein feil, eller om det er andre årsaker til avviket.



Figur 19: Døme på tredimensjonalt skårplott med dei tre første prinsipalkomponentane. Objekta deler seg inn i tre grupper, der objekta i same gruppe vil vere like med omsyn til desse prinsipalkomponentane, medan det vil vere større skilnad mellom objekt i ulike grupper. Figuren er henta frå hjelpefunksjonen til Unscrambler.

Ladningsplotta liknar på skårplotta, men syner variablane, ikkje objekta. Også her kan ein sjå etter grupperingar, for å finne variablar som korrelerer. Eit slikt plott er synt øvst til høgre i *figur 18*. Plott av scorar og ladningar bør undersøkjast saman. Eit objekt i skårplottet vil ha høg verdi av variablar som ligg på same plass i ladningsplottet, og låg verdi av variablar som ligg på motsett side i ladningsplottet. Dette vil syne kva variablar som er viktige for kvart enkelt objekt, eller for grupper av objekt.

Plottet nederst til venstre i *figur 18* er eit momentplott. Dette syner kor mykje kvart enkelt objekt påverkar den aktuelle prinsipalkomponenten. Dersom eit objekt har høgt moment, kan det vere at dette objektet påverkar modellen mykje i høve til dei andre objekta, og ein bør vurdere om objektet skal utelukkast frå modellen.

Sjølve tilpassinga av modellen kallast kalibrering. For å vurdere kvaliteten på modellen, nyttar ein validering. Det vil seie at ein prøver ut modellen på eit anna datasett enn det den vart laga med. Dersom ein ikkje har eit anna datasett, kan ein nytte delar av det opprinnelege settet til validering. Full kryssvalidering er ein slik type validering. Då fjernast èin og èin observasjon frå datasettet, og dei som er att verte nytta som valideringssett. Her har vi eit datasett som består av 81 pasientar. Til kryssvalideringa har vi difor 81 ulike valideringssett med èin pasient i kvart. Målet med validering er å undersøkje om modellen berre passar til det aktuelle datasettet, eller om den også kan brukast til andre sett.

Også valideringa kan visualiserast i plott, som synt nederst til høgre i *figur 18*. Her ser vi residualvarians for kalibreringa (blå) og for valideringa (raud). Valideringskurva syner naturleg nok større residualvarians enn kalibreringskurva, men skilnaden er ikkje stor. Det indikerer at valideringa for denne modellen er god, det vil seie at modellen representerer denne type data godt. Dersom modellen skal brukast vidare, er det viktig med god validering.

I denne oppgåva nyttast SVD-algoritmen og full kryssvalidering til PCA-analyse.

3.5.2 Klassifisering

Klassisifisering, det vil seie å dele observasjonar inn i grupper, kan vere anten overvaka (*supervised*), det vil seie at du på førehand veit kva grupper som finst, eller ikkje-overvaka (*unsupervised*), der ein ikkje har førehandskunnskap om gruppene. Diskriminant analyse (DA), SIMCA og SVM er døme på overvaka klassifisering, medan klyngeanalyse er ikkje-overvaka, *[16]*.

3.5.3 Diskriminant analyse

Ved diskriminant analyse søkjer ein å dele observasjonar inn i førehandsdefinerte grupper. Diskriminant analyse baserer seg på Bayes regel for betinga sannsyn, *[16]*. Dersom ein har i grupper, kalla π_i, og vil fordele k observasjonar på desse gruppene, er sannsynet for at observasjonen $\boldsymbol{x}_k$ høyrer til gruppa π_i gitt ved

$$P(\pi_i|\boldsymbol{x}_k) = \frac{P(\boldsymbol{x}_k|\pi_i)\,P(\pi_i)}{P(\boldsymbol{x}_k)}$$

der $P(\boldsymbol{x}_k|\pi_i)$ er sannsynet for å observere $\boldsymbol{x}_k$ dersom ein undersøkjer gruppe $\boldsymbol{\pi}_i$, $P(\pi_i)$ er sannsynet for at ein tilfeldig observasjon høyrer til gruppe π_i, medan $P(\boldsymbol{x}_k)$ er sannsynet for observasjonen $\boldsymbol{x}_k$. Analysen går ut på å plassere alle observasjonar i den klassen som maksimerer sannsynet $P(\pi_i|\boldsymbol{x}_k)$.

I Unscrambler kan ein velje mellom tre ulike typar diskriminant analyse: lineær, kvadratisk og Mahalanobis. I lineær diskriminant analyse (LDA) går ein ut i frå at alle gruppene har same kovariansmatrise, medan i kvadratisk diskriminant analyse (QDA) estimerast kovariansmatrisa for kvar gruppe separat, *[16]*. Dersom ein trur at det er same variasjon internt i alle gruppene, kan ein difor nytte LDA, elles vil QDA passe betre.

Det finst ulike måtar i rekne ut avstanden mellom ein observasjon og sentrum av ei gruppe, men dei vanlegaste er euklidsk avstand og Mahalanobis-avstand. Den euklidske avstanden mellom to observasjonar $\boldsymbol{x}_1$ og $\boldsymbol{x}_2$ av p variablar, er definert som

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(x_{11}-x_{21})^2+(x_{12}-x_{22})^2+\ldots+(x_{1p}-x_{2p})^2} = \sqrt{(\boldsymbol{x}_1-\boldsymbol{x}_2)^t(\boldsymbol{x}_1-\boldsymbol{x}_2)}$$

Dette er det ein vanlegvis tenkjer på som avstand, det vil seie lengda av den rette linja mellom dei to punkta.

Alternativt kan ein nytte Mahalanobis-avstanden, definert som

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1-\boldsymbol{x}_2)^t\,\Sigma^{-1}(\boldsymbol{x}_1-\boldsymbol{x}_2)}$$

der Σ er kovariansmatrisa mellom $\boldsymbol{x}_1$ og $\boldsymbol{x}_2$. Dette avstandsmålet tek med andre ord omsyn til spreiinga til objekta.

Sannsyna $P(\pi_i)$ for kvar gruppe kan anten setjast likt for alle gruppene, det vil seie at ein går ut i frå at det er like stort sannsyn for å vere i kvar klasse, eller ein kan estimere $P(\pi_i)$ ut i frå datasettet. Det siste er lurt dersom ein trur at det ikkje er like stort sannsyn for kvar klasse, men ikkje kjenner tala for desse sannsyna.

Resultatet av analysen er definisjonar av grensene mellom gruppene. Desse kan seinare nyttast til å klassifisere nye objekt. Dersom ein til dømes deler bilete av kreftsvulstar inn i to grupper etter om pasienten vert frisk eller ikkje, ønskjer ein å kunne seie om ein ny pasient vil bli frisk ved å nytte modellen på det nye biletet.

Tabell 9: Døme på forvirringsmatrise frå klassifiering med to klassar. Talet på riktig klassifiserte objekt er SP+SN.

		Faktisk verdi	
		Positiv	**Negativ**
Predikert verdi	**Positiv**	Sann Positiv (SP)	Falsk Positiv (FP)
	Negativ	Falsk Negativ (FN)	Sann Negativ (SN)

Resultatet av diskriminant analyse, og andre former for klassifisering, kan oppsummerast i form av ei forvirringsmatrise, synt i *tabell 9*. Her er (SP+SN) objekt klassifisert riktig, medan (FP+FN) objekt har hamna i feil klasse. Den samla nøyaktigheita vert såleis

$$Nøyaktigheit = \frac{SP + SN}{SP + FP + FN + SN}$$

Det finst også andre mål for resultatet av ei klassifisering, til dømes sensitiviteten,

$$Sensitivitet = \frac{SP}{SP + FN}$$

som syner kor mange som vart korrekt klassifiserte som positive, i forhold til kor mange som burde ha blitt det. På same måte fortel spesifisiteten kor stor del av dei som faktisk er negative, som blir klassifiserte som det,

$$Spesifisitet = \frac{SN}{SN + FP}$$

Sjå Olson og Delen, *[34]*, for fleire mål.

I diskriminant analyse kan det ikkje vere fleire variablar enn det er observasjonar i kvar gruppe. Metoden eignar seg såleis best til analysar der ein har mange observasjonar og få variablar, *[16]*.

3.5.4 SIMCA

SIMCA (*Soft Independent Modeling of Class Analogy*) er ein metode som nyttar prinsipalkomponentanalyse for å klassifisere objekt, *[33]*. I denne metoden lagast èin PCA-modell for kvar klasse, det vil seie at kvar klasse modellerast individuelt. Nye observasjonar projiserast inn i alle modellane, og blir klassifisert til den klassen dei har minst avstand til.

PCA-modellane for kvar klasse kan undersøkjast i plott av Q og T^2, *[35]*. Q er kvadratsummen av residuala til modellen,

$$Q_i = \boldsymbol{e}_i \boldsymbol{e}_i^t$$

der $\mathbf{e}_i$ er rad *i* i residualmatrisa E. Q måler såleis avvik frå modellen.

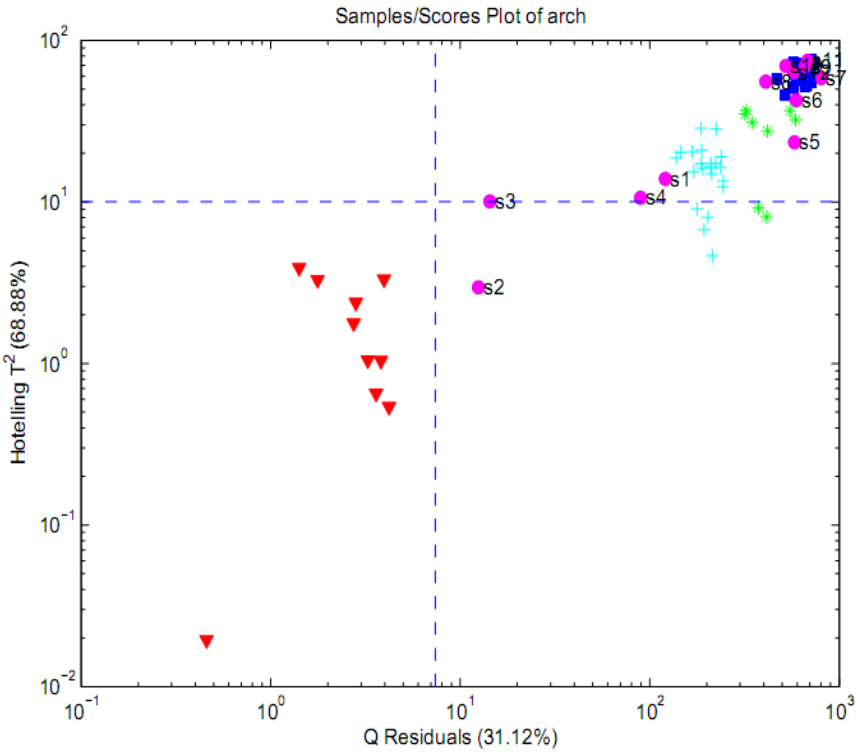
T^2 definerast som

$$T_i^2 = t_i \lambda^{-1} t_i^t$$

der $\mathbf{t}_i$ er rad *i* i skårmatrisa T. T^2 måler difor variasjonen innanfor modellen.
Figur 20 syner døme på eit plott av Q mot T^2.

Ved eit gitt signifikansnivå, til dømes 5%, kan ein finne grenseverdiar for Q og T^2. Ein observasjon som passar inn i PCA-modellen, har Q- og T^2-verdiar innanfor desse grensene. I *figur 20* ligg dei raude punkta innanfor grensene, medan dei andre ikkje gjer det. Det syner at PCA-modellen høver godt til dei raude punkta, som er den klassen den vart laga for, og at objekta i dei andre klassane ikkje ligg innanfor denne modellen.

Dersom eit objekt ligg innanfor signifikansnivåa til fleire av PCA-modellane, kan dette objektet klassifiserast til å høyre til begge klassane. Og dersom ein objekt ligg utanfor signifikansgrensene for alle PCA-modellane, vert denne ikkje plassert i noko klasse. SIMCA gjev såleis ikkje like eintydige resultat som diskriminant analyse og klyngeanalyse, sjå *kapittel 3.5.6*, der kvart objekt berre klassifiserast til ein klasse. Dette kjem av at PCA-modellane for kvar klasse er laga uavhengig av kvarandre, og difor kan overlappe.

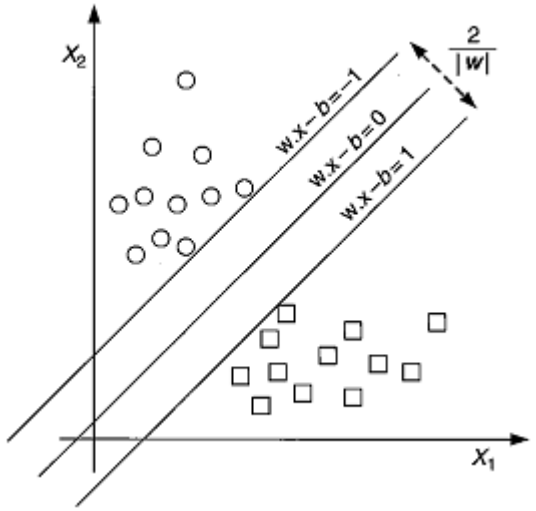


Figur 20: Døme på plott av T^2 mot Q for ein PCA-modell laga frå klassen representert med dei raude punkta. Dei andre punkta syner dei andre klassane i datasettet. Dei stipla linjene er signifikansgrensene for Q og T^2. Dei raude punkta ligg innanfor signifikansgrensene, medan dei andre punkta ligg utanfor. Det vil seie at dei raude passar inn i modellen, medan resten ikkje gjer det.
Henta frå "Chemometrics Tutorial for PLS_Toolbox and Solo", [35].

3.5.5 Støttevektormaskiner (SVM)

Støttevektormaskiner (*Support Vector Machines*, SVM) er nok ein metode som kan nyttast til klassifisering av objekt. Denne teknikken utnyttar det at nokre objekt påverkar grensa mellom klassane meir enn andre, i tillegg til at den kan nytte ikkje-lineære transformasjonar for å få eit betre skilje mellom klassar.

Metoden går ut på å finne det beste hyperplanet som skil mellom gruppene, det vil seie det hyperplanet som maksimerer marginen kring grensa, som synt i *figur 21*.



Figur 21: To grupper av data representert med ulike symbol (ring og firkant). Linja i midten, markert med $\boldsymbol{w}\cdot\boldsymbol{x}-b=0$*, er grensa mellom dei to klassane. Marginen kring grensa er gitt ved linjene* $\boldsymbol{w}\cdot\boldsymbol{x}-b=+-1$*. Henta frå Olson og Delen, [34].*

Støttevektorane er dei observasjonane som ligg nærast grensa mellom gruppene, det vil seie dei observasjonane som ligg på marginane. Dersom desse observasjonane fjernast, vil grensa mellom gruppene flytte seg, medan dette ikkje vil skje om vi fjernar andre observasjonar. Støttevektorane er difor dei observasjonane som er viktigast for å skilje mellom klassane.

Det optimale skiljet mellom klassane vil gi så store marginar som mogleg, og samtidig ha få støttevektorar. Då er avstanden mellom klassane så stor som mogleg, noko som hindrar at nye observasjonar plasserast i feil gruppe, i tillegg til at modellen kan vere meir robust, sidan dei fleste observasjonane kan fjernast utan at dette endrar grensa.

For å uttrykke dette matematisk, søkjer ein altså eit hyperplan på forma

$$\boldsymbol{w}\cdot\boldsymbol{x}-b=0$$

der $\boldsymbol{x}$ er eit punkt, $\boldsymbol{w}$ er ein vektor vinkelrett på hyperplanet og b er ein parameter som styrer storleiken på margen, *[34]*. Dette planet skal skilje klassane av objekt frå kvarandre.

Marginane kring denne grensa er parallelle hyperplan gitt ved

$$\boldsymbol{w}\cdot\boldsymbol{x}-b=+1$$

og

$$\boldsymbol{w}\cdot\boldsymbol{x}-b=-1$$

Desse tre plana, sjølve grensa og dei to marginane, er synte i *figur 21*.

Dersom objekta er lineært separable, vil vi plassere grensa slik at ingen objekt hamnar mellom marginane. Dette kan vi utrykke som

$$\boldsymbol{w}\cdot\boldsymbol{x_i}-b\geq+1$$

eller

$$\boldsymbol{w}\cdot\boldsymbol{x_i}-b\leq-1$$

for alle objekta $\boldsymbol{x_i}$. Dette kan også skrivast som

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq+1$$

der $y_i = +1$ for den eine klassen og $y_i = -1$ for den andre.

Kravet om å ha så breie marginar som mogleg, vil seie at ein må maksimere avstanden $2/\|\boldsymbol{w}\|$ mellom dei, synt i *figur 21*, noko som tilsvarar å minimere $\|\boldsymbol{w}\|$.

Desse to krava til saman tilsvarar å minimere $\frac{1}{2}\|\boldsymbol{w}\|^2$ med betingelsen $y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)-1\geq 0$ for alle i. Dette problemet kan løysast ved hjelp av Lagrange-multiplikatorar, som vist hjå til dømes Ivanciuc, *[36]*.

Ofte vil det ikkje vere mogleg å skilje perfekt mellom klassane. For å ta omsyn til dette, introduserer vi eit slingringsmonn i form av variablane $\xi_i \geq 0$, slik at

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq+1-\xi_i$$

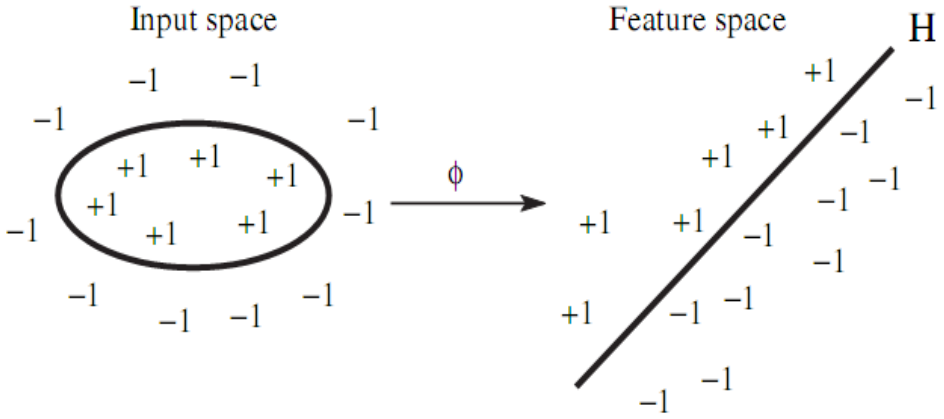
Korrekt klassifiserte objekt vil ha $\xi_i = 0$. I tillegg til krava vi hadde tidlegare, vil ein no ha så få $\xi_i \neq 0$ som mogleg. Funksjonen som skal minimerast vert no

$$\frac{1}{2}\|\boldsymbol{w}\|^2+C\Sigma\xi_i$$

det $C > 0$ er ein kostparameter, som indikerer kor straffbart det skal vere å feilklassifisere. Tilleggsbetingelsen seier no at

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x_i}-b)\geq 1-\xi_i$$

for alle i.



Figur 22: Objekta markerte med +1 høyrer til èi gruppe, medan objekta markerte med -1 høyrer til ei anna. Det er ikkje mogleg å skilje dei to gruppene ved hjelp av ei rett linje. Ved å transformere observasjonane til eit feature-rom, kan dei enklare skiljast. Henta frå Ivanciuc, [36].

Ikkje alle klassar er lineært separable. Eit døme på dette er vist til venstre i *figur 22*, der det ikkje er mogleg å skilje dei to klassane ved hjelp av ei rett linje. For å løyse dette, kan ein transformere objekta til eit høgaredimensjonalt rom, kalla *feature*-rommet H, der dei kan vere lineært separable. Ein kan såleis nytte lineære klassifiseringsmetodar på ikkje-lineære samanhengar. Denne transformasjonen skjer ved funksjonen Φ, som også definerer kernelfunksjonen $K(\mathbf{x}_1,\mathbf{x}_2)$ gitt ved

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Phi(\boldsymbol{x}_1)^t \Phi(\boldsymbol{x}_2)$$

Sidan det er berre dette skalarproduktet, ikkje Φ åleine, som opptrer i funksjonen som skal minimerast, treng vi berre å kjenne kernelfunksjonen, ikkje Φ.

Det blir nytta mange ulike kernelfunksjonar, men desse er nokre av dei vanlegaste:

Lineær:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^t \boldsymbol{x}_2$$

Polynomisk:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\gamma \boldsymbol{x}_1^t \boldsymbol{x}_2 + r)^d \quad , \quad \gamma > 0$$

Radial basis-funksjon (RBF):

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\gamma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2) \quad , \quad \gamma > 0$$

Sigmoid:

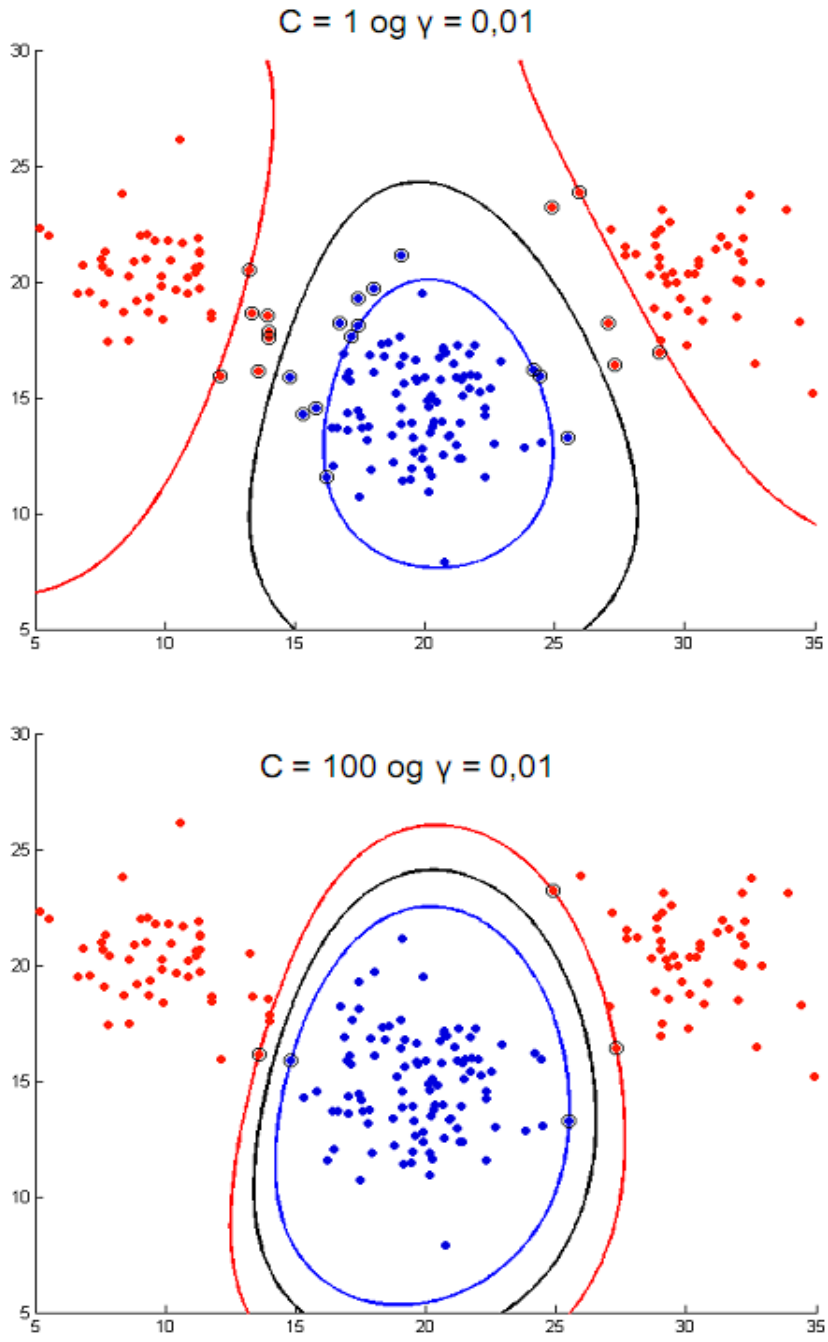
$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh(\gamma \boldsymbol{x}_1^t \boldsymbol{x}_2 + r)$$

γ, r og d er parameterar som må veljast for å finne den optimale grensa mellom gruppene. Hsu, Chang og Lin, *[37]*, tilrår å starte med RBF-kernelen.

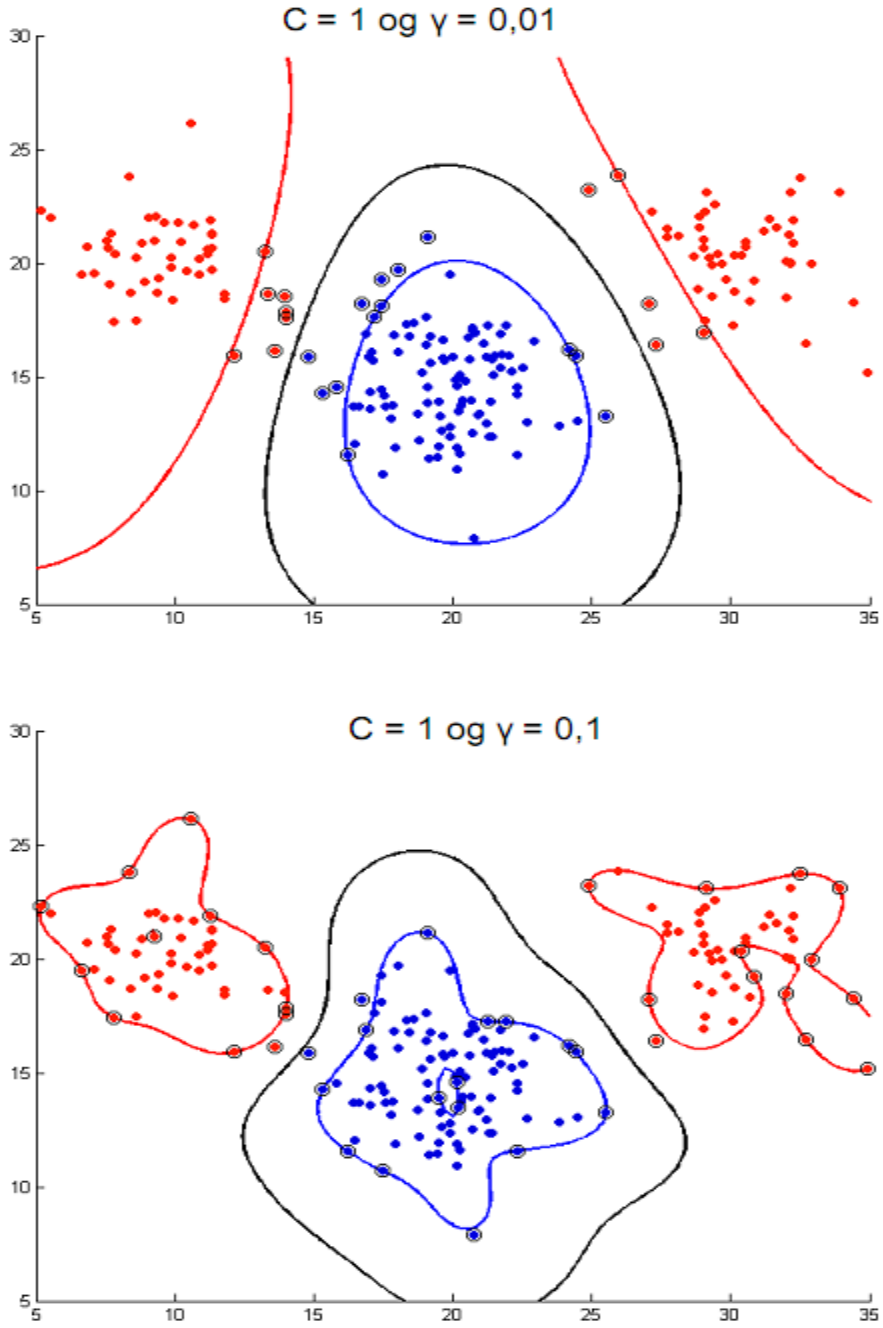
Både kostparameteren C og kernelparameterane, γ for RBF-kernelen, må spesifiserast for å løyse problemet. Sidan ein sjeldan på førehand veit kva verdiar av desse som gjev best resultat, tilrår Hsu, Chang og Lin, *[37]*, eit gridsøk over ulike (C,γ)-kombinasjonar kombinert med kryssvalidering. Ein deler då objekta inn i v grupper, og testar modellen med det gitte (C,γ)-paret på dei (v -1) andre gruppene. Etter å ha testa alle kombinasjonar, vel ein dei (C,γ) som gjev størst nøyaktigheit, det vil seie dei som fører til at flest mogleg objekt blir klassifiserte riktig.

Effekten av å variere C og γ er synt i *figurane 23 og 24*. Låg verdi for C gjev breie margar, medan ein stor C gjev smale margar. Breiare margar kan føre til fleire støttevektorar slik som i *figur 23*, sidan det no er lågare straff for å feilklassifisere eit objekt. γ kontollerer forma på grensa, dess større verdia av γ, dess meir komplisert kan grensa vere.

For ei meir omfattande innføring i SVM, sjå til dømes Ivanciuc, *[36]*, og Hsu, Chang og Lin, *[37]* .



Figur 23: Resultat frå SVM-analyse. Den svarte linja er grensa mellom klassane, medan dei raude og blå linjene er marginane. Blå punkt høyrer til èin klasse, dei raude til den andre klassen. Punkt som er markerte med ein sirkel er støttevektorar. I begge modellane er det nytta γ = 0.01. I det øvste biletet er C = 1, medan vi har C = 100 i den nederste biletet. Henta frå Eigenvector, [38].

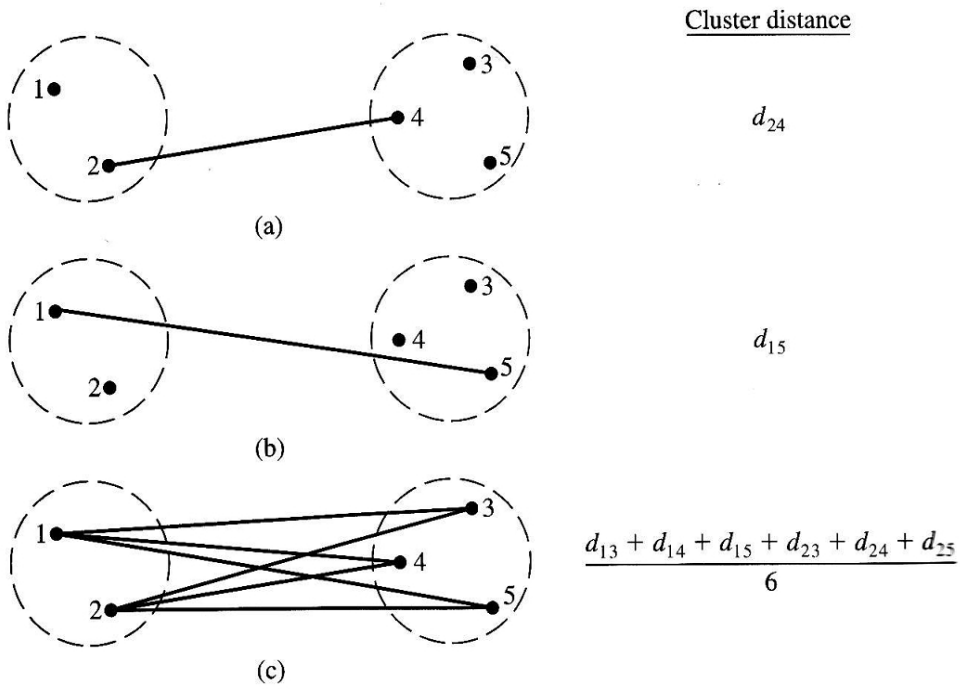


Figur 24: Resultat frå SVM-analyse. Den svarte linja er grensa mellom klassane, medan dei raude og blå linjene er marginane. Blå punkt høyrer til èin klasse, dei raude til den andre klassen. Punkt som er markerte med ein sirkel er støttevektorar. I begge modellane er det nytta C = 1. I det øvste biletet er γ = 0.01, medan vi har γ = 0.1 i den nederste biletet. Henta frå Eigenvector, [38].

3.5.6 Klyngeanalyse

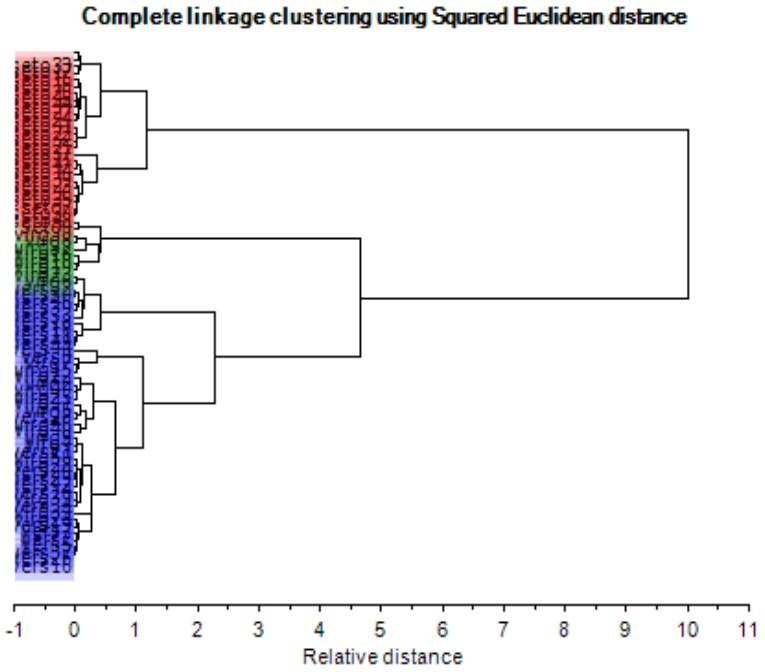
På same måte som for diskriminant analyse, vil ein med klyngeanalyse dele observasjonar i grupper, men her kjenner ein ikkje gruppene på førehand. Ein treng framleis ein mål på kor nærme to objekt ligg kvarandre, til dømes euklidsk avstand, sjå *avsnitt 3.5.3*.

Avstand mellom kvar klynge kan også bereknast på fleire måtar, som synt i *figur 25*. Enkeltkopling vil seie at ein definerer avstanden mellom to klynger som avstanden mellom dei to objekta i kvar sin klasse som ligg nærast kvarandre, medan komplett kopling er avstanden mellom dei to objekta som er lengst frå kvarandre. Gjennomsnittskopling er gjennomsnittet av avstanden mellom alle objekta i kvar klynge.



Figur 25: Tre ulike metodar for å rekne ut avstanden mellom to klynger. a) Enkeltkopling. b) Komplett kopling. c) Gjennomsnittskopling. Henta frå Johnson og Wichern, [16].

Objekta kan grupperast anten ved å starte opp med alle objekt i same gruppe, og så splitte opp gruppa heilt til kvart objekt ligg i kvar si gruppe, eller omvendt, det vil seie å starte med ei gruppe per objekt. Resultatet kan visualiserast i eit dendrogram, synt i *figur 26*. Ut i frå dendrogrammet, kan ein avgjere kor mange grupper ein synest det er naturleg å dele data inn i. På figuren har analysen resultert i tre ulike grupper, fordi det mykje mindre skilnad mellom objekta innanfor desse gruppene enn det er mellom gruppene.



Figur 26: Døme på dendrogram. Objekta ligg i venstre i figuren, og lengda på linjene indikerer kor langt det er mellom kvart objekt. Her har klyngeanalysen delt objekta inn i tre hovudgrupper, synt med tre ulike fargar på figuren. Figuren er henta frå hjelpefunksjonen i Unscrambler.

Ein mykje brukt algoritme for klyngeanalyse, særleg for store datasett, *[33]*, er K-means-klynger. Denne startar med å dele objekta tilfeldig inn i *K* grupper (klynger). Ein finn sentrum, det vil seie gjennomsnittsverdien, for kvar gruppe, og reknar ut avstanden frå kvart objekt til sentrum av alle gruppene. Objektet flyttast så til den gruppa som ligg nærast. Deretter finn ein sentrum av kvar av dei nye gruppene, og undersøkjer på nytt om nokre observasjonar no skal flyttast. Dette held fram til ingen observasjonar skal flyttast. Ein har då delt objekta inn i *K* grupper.

Det finst ein tilsvarande algoritme, K-medians-klynger, som nyttar medianverdi i staden for gjennomsnittsverdi til å finne sentrum av gruppene. Denne metoden er mindre sensitiv for avvikarar, det vil seie observasjonar som er veldig ulike dei andre, men er også tregare enn K-means-klynger.

Sjå til dømes Johnson og Wichern, *[16]*, for meir utfyllande skildringar av ulike typar klyngeanalyse.

3.5.7 Lineær regresjon

Lineær regresjon går ut på å skrive responsvariabelen som ein vekta sum av forklaringsvariablane,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

der y er responsvariabelen, x_1, x_2, ..., x_n er forklaringsvariablane, β_0, β_1, ..., β_n er regresjonskoeffisientane og ε er residualet, *[39]*. Dette kan også skrivast på matrise form, som

$$y = X\boldsymbol{\beta} + \epsilon$$

der X er matrisa med forklaringsvariablane, medan $\boldsymbol{\beta}$ er vektoren med regresjonskoeffisientane.

Regresjonskoeffisientane skal bestemmast slik at dei minimerer kvadratsummen

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n))^2$$

der y_i er den observerte verdien. Kvadratsummen uttrykker altså skilnaden mellom faktisk og estimert verdi.

Vektoren $\boldsymbol{b}$ kan då uttrykkast som

$$\boldsymbol{b} = (X^t X)^{-1} X^t y$$

Dette krev at det ikkje er korrelasjon mellom dei ulike forklaringsvariablane. Dersom ein vil nytte lineær regresjon, bør variablane difor vere uavhengige av kvarandre. Ei svært grundig innføring i lineær regresjon finnast hjå Montgomery, Peck og Vining, *[39]*.

3.5.8 Partial Least Squares (PLS)

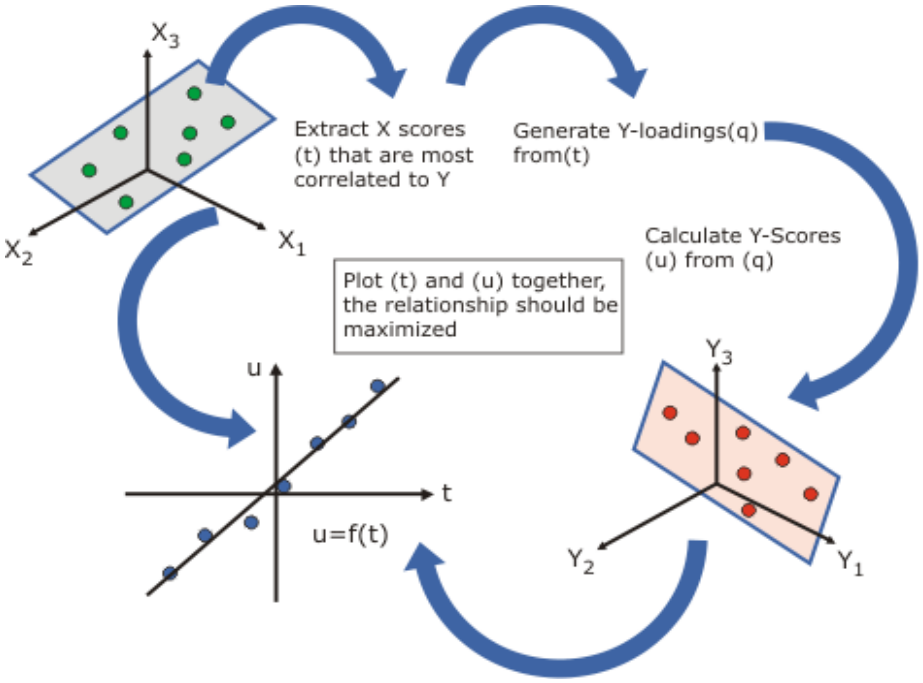
Partial Least Squares (PLS) er ein annan regresjonsmetode som baserer seg på prinsipalkomponentanalyse (PCA). Målet er å gjere kovariansen mellom X og Y så stor som mogleg. I PLS utfører ein prinsipalkomponentanalysar på både X og Y, og får

$$X = \sum T \cdot P^t + E$$

$$Y = \sum U \cdot Q^t + F$$

der T og U er skårmatrisene, P og Q er ladningsmatrisene og E og F er residualmatrisene, *[40]*.

Det spesielle med PLS er at ein ikkje utfører PCA på X og Y kvar for seg, men på begge samtidig. Det gjer ein ved å la $\boldsymbol{u_1}$, det første estimatet av den første skårvektoren i Y, vere utgangspunkt for $\boldsymbol{t_1}$, det første estimatet av den første skårvektoren i X. Den $\boldsymbol{t_1}$ ein då finn, nyttast igjen som utgangspunkt til neste estimat av $\boldsymbol{u_1}$. Slik byttar ein fram og tilbake, heilt til løysinga konvergerer. Ein illustrasjon av denne prosessen er gitt i *figur 27*. Ei meir omfattande forklaring finnast til dømes i Haenlein og Kaplan, *[40]*.



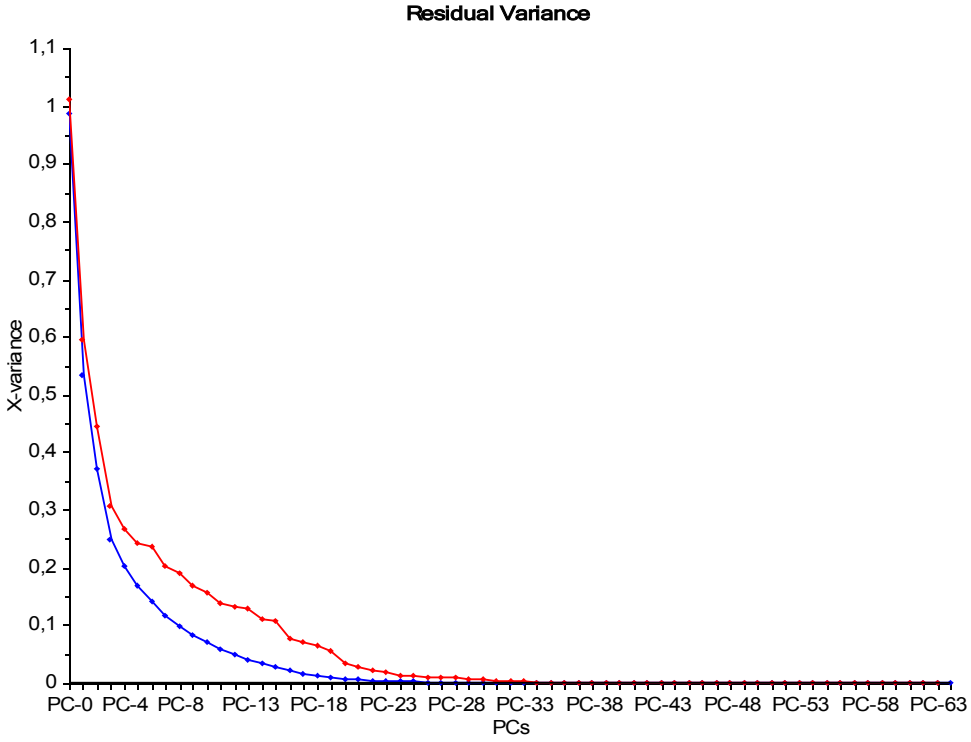
Figur 27: For å finne den beste modellen av data med PLS, nyttar ein prinsipalkomponentanalyse og itererer seg mellom forklaringsvariablane X og responsvariablane Y fram til løysinga konvergerer. Figuren er henta frå hjelpefunksjonen i Unscrambler X.

Resultata av PLS kan undersøkjast ved hjelp av plott, på same måte som ved PCA. Les Esbensen, *[33]*, for gode råd om tolking av PLS-resultat.

4 Resultat

4.1 Prinsipalkomponentanalyse

Prinsipalkomponentanalyse av dei deskriptive statistiske parameterane for A, k_{ep} og k_{el}, samt parameterane alder, volum og figo-stadie, vart utført som skildra i kapittel *3.5.1*. Variablane $k_{el\ min}$ og $k_{el\ mode}$ er utelukka frå analysen, då desse er like for alle svulstane, og såleis ikkje har varians. Det er difor 64 forklaringsvariablar, og 64 er høgast moglege tal på prinsipalkomponentar (PC).



Figur 28: Residualvarians for PCA-modell av alder, volum, stadie og deskriptive statistiske parameterar for A, k_{ep} og k_{el}. Maksimalt tal på prinsipalkomponentar er 64. Den blå kurva syner residualvarians for kalibreringa, medan den raude kjem frå full kryssvalidering. Laga med Unscrambler.

Figur 28 syner residualvariansplottet for modellen, det vil seie kor mykje varians som er att for kvar ny prinsopalkomponent ein legg til. Den raude kurva, som syner residualvarians frå full kryssvalidering, følgjer godt den blå kurva som syner residualvarians for sjølve modellen. Det indikerer at modellen endrar seg lite dersom èin observasjon fjernast.
Residualvarianskurva fell raskt, noko som syner at ein stor del av variasjonen i data kan forklarast med få prinsipalkomponentar. Den første prinsipalkomponenten forklarer heile 46% av variansen, og det trengst åtte komponentar for å forklare over 90 %, som vist i *tabell 10*.

Tabell 10: Forklart varians for dei ti første komponentane i PCA-modellen av alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el}.

Komponent	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8	PC-9	PC10
Forklart varians	46%	17%	12%	5%	3%	3%	3%	2%	2%	1%
Samla forklart varians	46%	63%	75%	80%	83%	86%	89%	91%	93%	94%

4.1.1 Ladningar

Figur 29 - 31 syner skår- og ladningsplott for dei 12 første prinsipalkomponentane. I skårplotta er objekta, det vil seie kvar pasient, merka med raudt og blått. Dei raude punkta representerer pasientar med tilbakefall (*pfs* = 1), medan dei blå er pasientane som vart friske att (*pfs* = 0). Ingen av prinsipalkomponentane ser ut til å dele objekta inn i desse to kategoriane.

PC-1, den første prinsipalkomponenten, ser ut til å styrast mest av persentilverdiane, då desse gjev størst utslag i ladningsplottet, synt øvst til venstre i *figur 29*. Persentilane for A og k_{el} har postitive verdiar i PC-1, medan persentilane for k_{ep} har negative verdiar. Det vil seie at objekt med høg verdi av PC-1 vil ha høg verdi for A- og k_{el}-persentilane, medan dei vil ha låge k_{ep}-persentilar. Denne komponenten skildrar difor svulstar med generelt høge verdiar av A og k_{el}, men låge verdiar av k_{ep}. Dette vil vere svulstar der mesteparten av svulsten tek opp mykje kontrastmiddel (A) og vaskar det raskt ut at (k_{el}), men som tek opp kontrastmidlet sakte (k_{ep}). Rask utvasking indikerer stor blodgjennomstrøyming, medan låg opptaksrate tyder på vev med låg permeabilitet, det vil seie lite lekkasje.

Objekt med høge verdiar av PC-2, sjå ladningsplottet øvst til venstre i *figur 29*, kan assosierast med høge verdiar av dei låge persentilane til A og k_{ep}. I tillegg vil dei ha låge maksimalverdiar for både A, k_{ep} og k_{el}. Også verdiane for skeivskap, A_{skew}, $k_{ep\ skew}$ og $k_{el\ skew}$ vil vere låge, samt at volumet av svulsten ikkje vil vere så stort. Om ein undersøkjer svulstvolumet til pasientane med høg verdi av PC-2 i *figur 15*, har dei fleste eit mindre volum enn gjennomsnittet, men det gjeld ikkje alle. Det kan skuldast at denne komponenten berre forklarer 17% av variasjonen i data, slik at det store utslaget i PC-2 kan balanserast av andre prinsipalkomponentar. PC-2 skildrar såleis små svulstar der store delar av svulsten tek opp mykje kontrastmiddel raskt (A og k_{ep}), og vaskar det raskt ut (k_{el}).

For PC-3 vil høge verdiar av komponenten tilsvare høge verdiar av k_{el}-persentilane, særleg dei øvste persentilane. Ladningsplottet til PC-3 er synt nederst til høgre i *figur 29*. Objekta med høg verdi for PC-3 vil også ha låge verdiar av A-persentilane, samt $k_{el\ kurt}$ og $k_{el\ skew}$, det vil seie skeivskap og kurtose for k_{el}. Desse svulstane har høg utvaskingsrate k_{el}, og såleis stor blodgjennomstrøyming, samt lågt kontrastmiddelopptak, A, over heile svulsten.

Objekt med høg verdi av PC-4, sjå øvst til høgre i *figur 29*, har låge verdiar for persentilavstandane *75%-25%* og *90%-10%* for A, i tillegg til lågt standardavvik A_{std}. Dei vil ha høge verdiar for låge persentilar av A og k_{el}, samt $k_{el\ skew}$, $k_{el\ kurt}$ og $k_{el\ max}$. Desse svulstane har såleis lite spreiing i kor mykje kontrastmiddel som blir teke opp, A, det er stort sett likt over heile svulsten. Pasient 54, som har den lågaste verdien i PC-4, er også den pasienten med størst persentilavstandar *75%-25%* og *90%-10%*. Ein kunne tenkje at denne prinsipalkomponenten skil mellom homogene (lite variasjon) og heterogene (stor variasjon) svulstar med omsyn på A, men ei nærare undersøking syner at dette ikkje er tilfelle. Det kan igjen skuldast at denne komponenten forklarer kun 5% av variansen.

Høge verdiar PC-5 indikerer høg alder og høge verdiar av $k_{ep\ skew}$ og $k_{ep\ kurt}$, samt låge verdiar av A_{skew} og A_{kurt}. Pasientar med stor PC-5 vil difor generelt vere eldre enn dei andre, noko som i stor grad bekreftast av aldersfordelinga i *figur 14*.

Svulstar med høg verdi av PC-6 har stort volum og høg verdi av $k_{el\ max}$ og $k_{ep\ mode}$. Dei vil ha låge verdiar av $A_{skew,}$ $A_{kurt,}$ $k_{ep\ skew}$ og $k_{ep\ kurt}$. Med andre ord er dette store svulstar der nokre vokslar har ekstra stor utvaskingsrate k_{el}. Dei fleste vokslane har stor opptaksrate k_{ep}.

Det er få variablar som har negativ ladning i PC-7, dei fleste ligg kring null. $k_{ep\ min}$ og A_{min} har derimot større positiv ladnding enn dei andre variablane. Svulstar med stor PC-7 kan difor tolkast som svulstar der ingen av vokslane har svært liten opptaksrate (k_{ep}) eller liten konsentrasjon av kontrastmidlet (A).

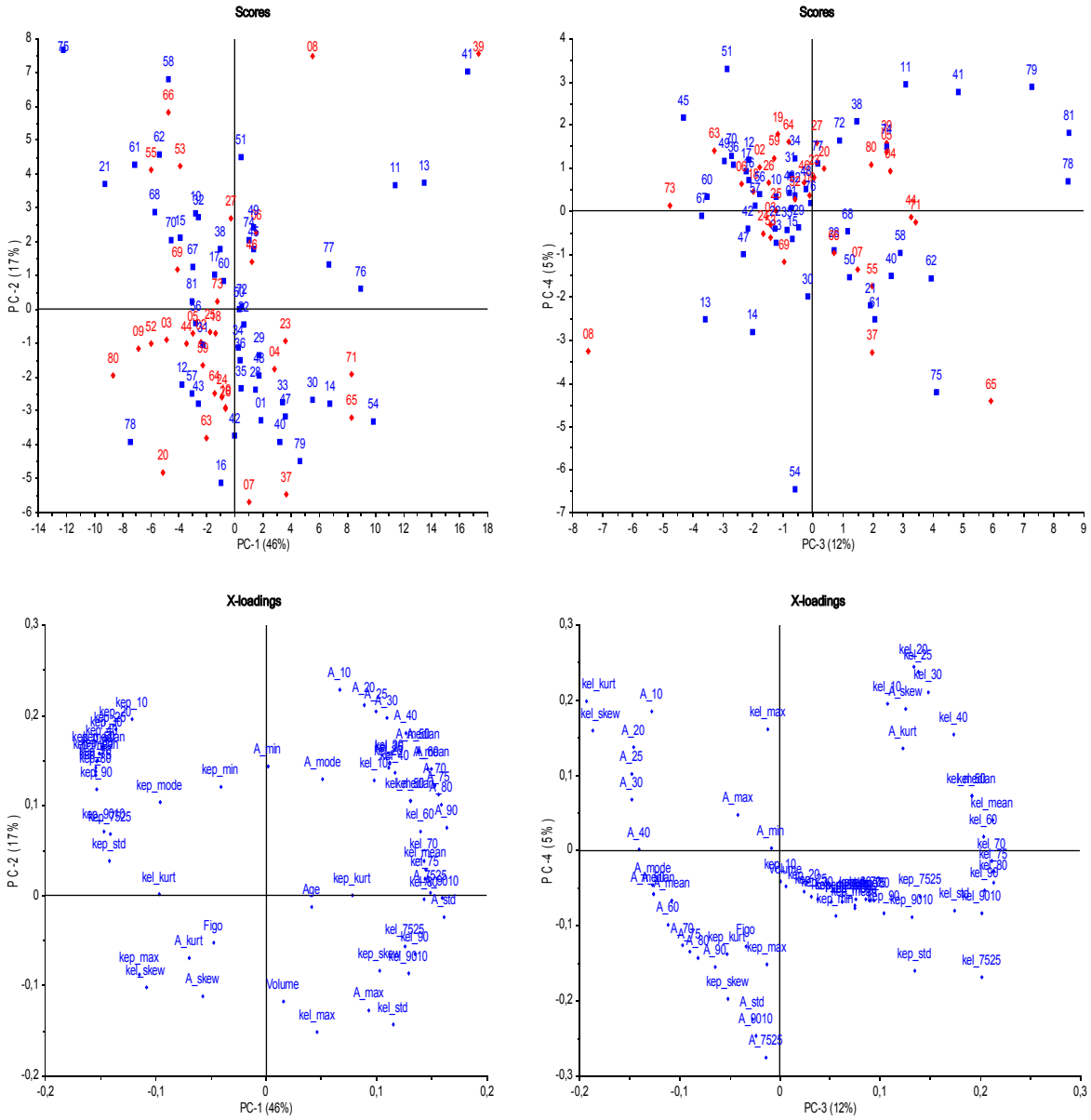
PC-8 har høge positive ladningar for stadie (*figo*), A_{max}, A_{mode} og $k_{ep\ min}$ og noko negative verdiar av $k_{ep\ mode}$, $k_{ep\ kurt}$ og $k_{el\ kurt}$. Svulstane med høg verdi her vil vere dei i utgangspunktet alvorlegaste svulstane, det vil seie dei som hadde høgast stadie før behandling. Ein del vokslar vil ha høge kontrastmiddelkonsentrasjonar, A, og det er få vokslar med låg opptaksrate, k_{ep}.

Pasientar med høg verdi av PC-9 ser ut til å vere eldre enn dei andre, ettersom variabelen alder har stor positiv verdi i ladningsplottet. Ei undersøking av aldersfordelinga i *figur 14* bekreftar dette. Dei vil også ha låg A_{mode} og $k_{ep\ mode}$, det vil seie at dei fleste vokslane har låge verdiar for kontrastmiddelkonsentrasjon og opptaksrate.

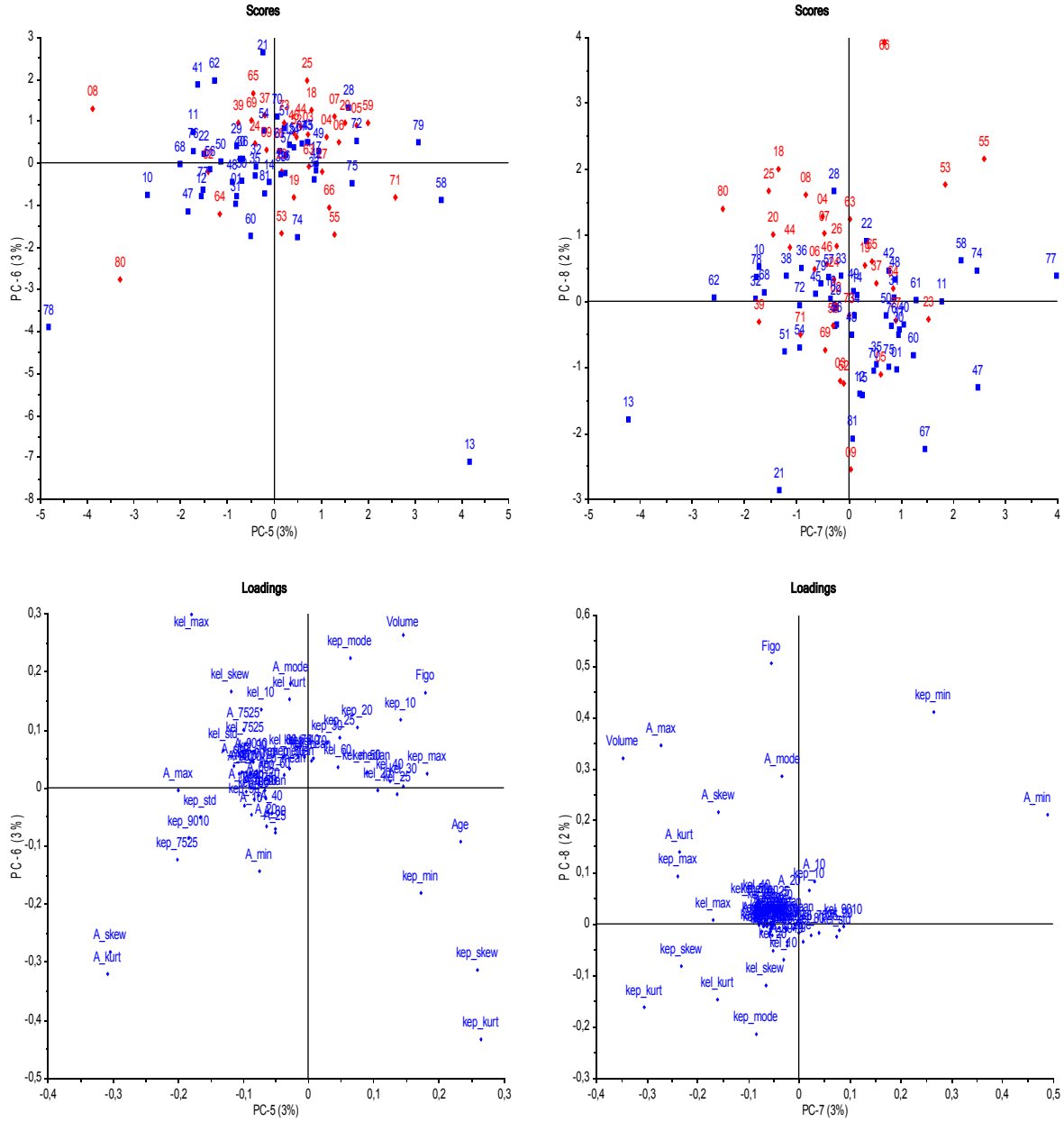
PC-10 styrast først og fremst av variablane volum, stadie og alder, med positiv verdi for volum og negativ verdi for stadie og alder. Høg verdi av PC-10 indikerer altså yngre pasientar med store, men mindre alvorlege, svulstar.

I PC-11 er $k_{el\ max}$ variabelen med størst positiv verdi, medan stadie og $k_{el\ 10}$ har størst negativ verdi. Det vil seie at svulstar med stor PC-11 har svært høg utvaskingsrate i ein eller fleire vokslar, og få vokslar med låg utvaskingsrate. Stadiet til svulsten før behandling var lågt.

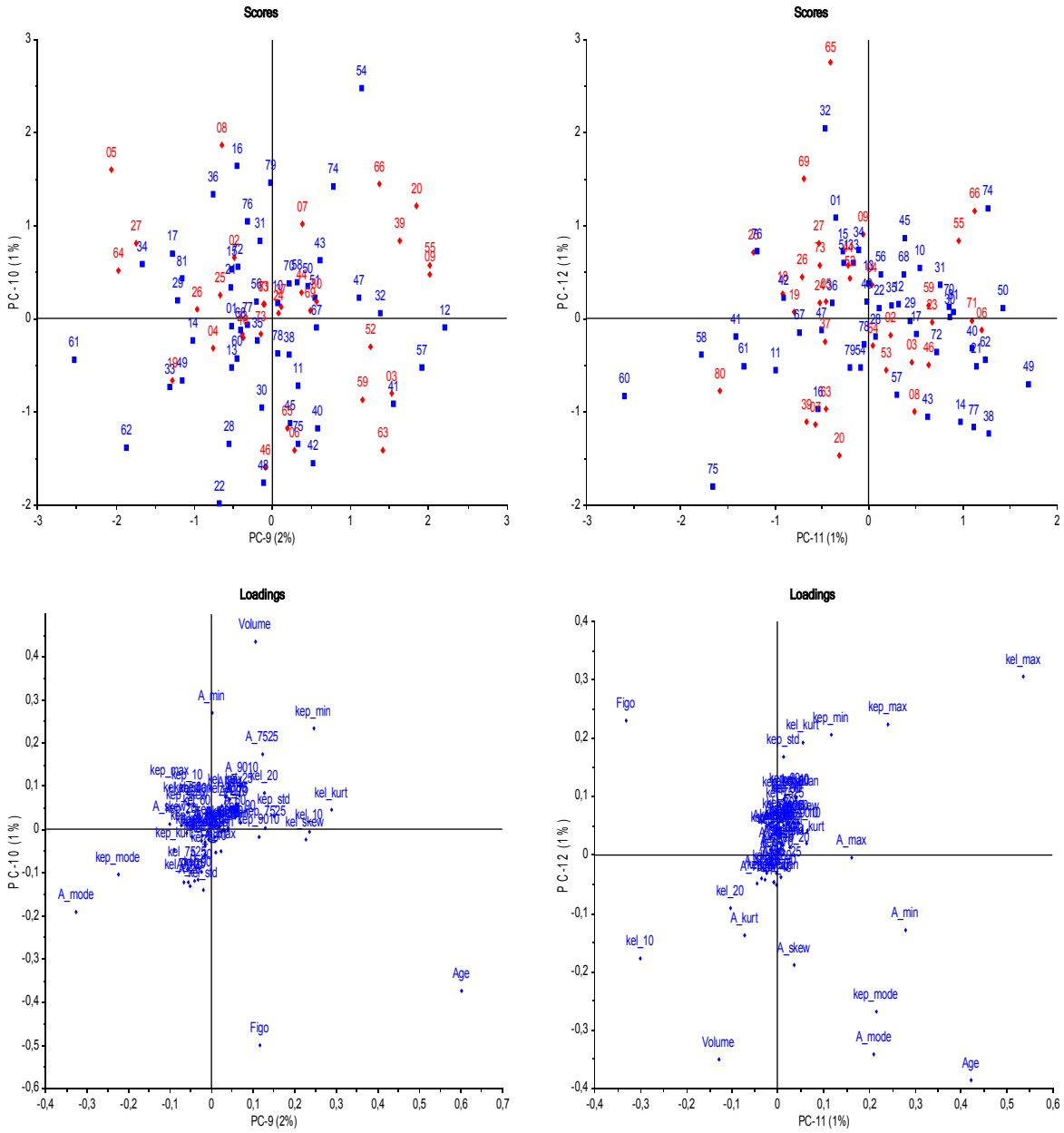
PC-12 har ingen variablar som merkar seg ut med stor postitiv verdi, men $k_{el\ max}$ er noko større enn dei andre. Variablane volum, alder og A_{mode} har størst negative verdiar. Pasientar med stor PC-12 vil difor vere unge og ha små svulstar. Dei fleste vokslane i svulsten tek opp lite kontrastmiddel, og ein eller fleire vokslar har høg utvaskingsrate.



Figur 29: Skårar (øvst) og ladningar (nederst) for PC-1 og PC-2 (til venstre), og PC-3 og PC-4 (til høgre). PCA-modell av alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el}. Dei blå punkta i skårplotta er pasientar med pfs = 0, medan dei raude er pasientar med pfs = 1. Tala er pasientnummera, frå 1 til 81. Laga med Unscrambler.

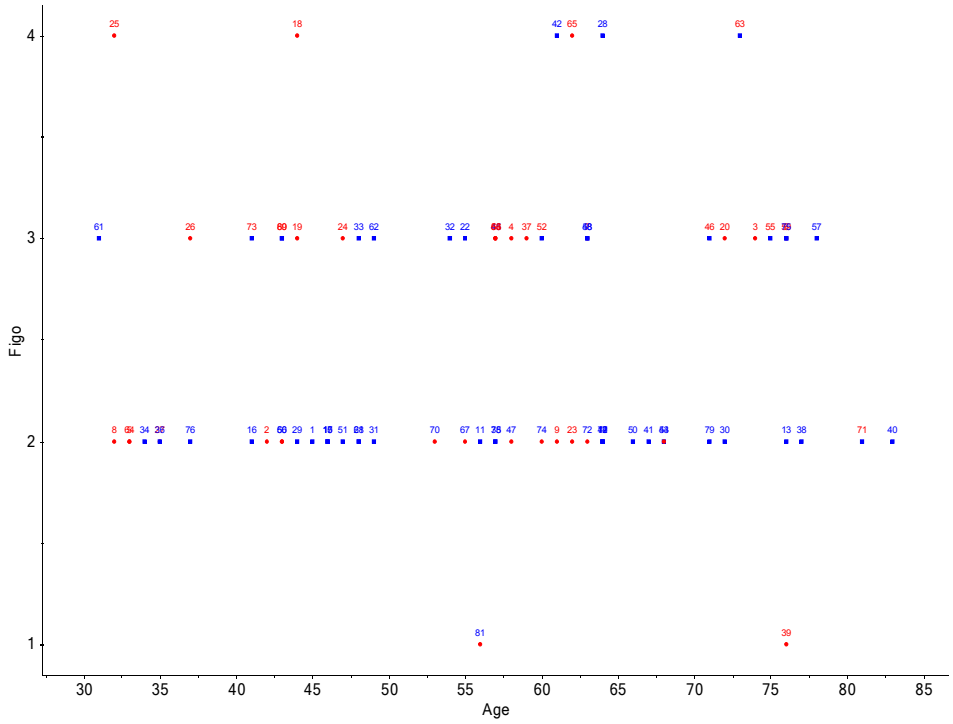


Figur 30: Skårar (øvst) og ladningar (nederst) for PC-5 og PC-6 (til venstre), og PC-7 og PC-8 (til høgre). PCA-modell av alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el}. Dei blå punkta i skårplotta er pasientar med pfs = 0, medan dei raude er pasientar med pfs = 1. Tala er pasientnummera, frå 1 til 81. Laga med Unscrambler.

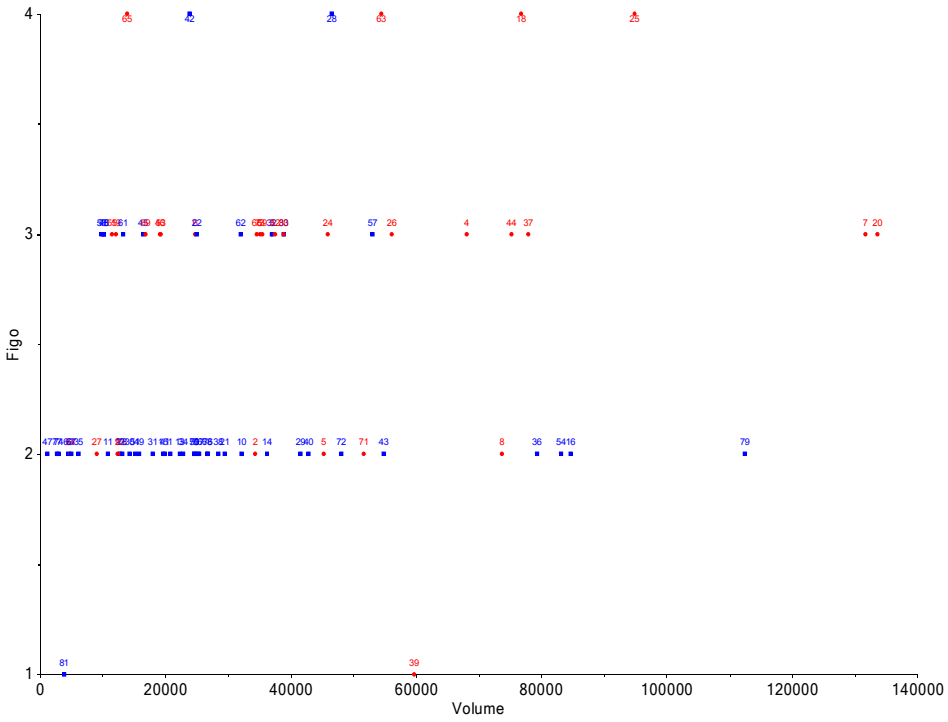


Figur 31: Skårar (øvst) og ladningar (nederst) for PC-9 og PC-10 (til venstre), og PC-11 og PC-12 (til høgre). PCA-modell av alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el}. Dei blå punkta i skårplotta er pasientar med pfs = 0, medan dei raude er pasientar med pfs = 1. Tala er pasientnummera, frå 1 til 81. Laga med Unscrambler.

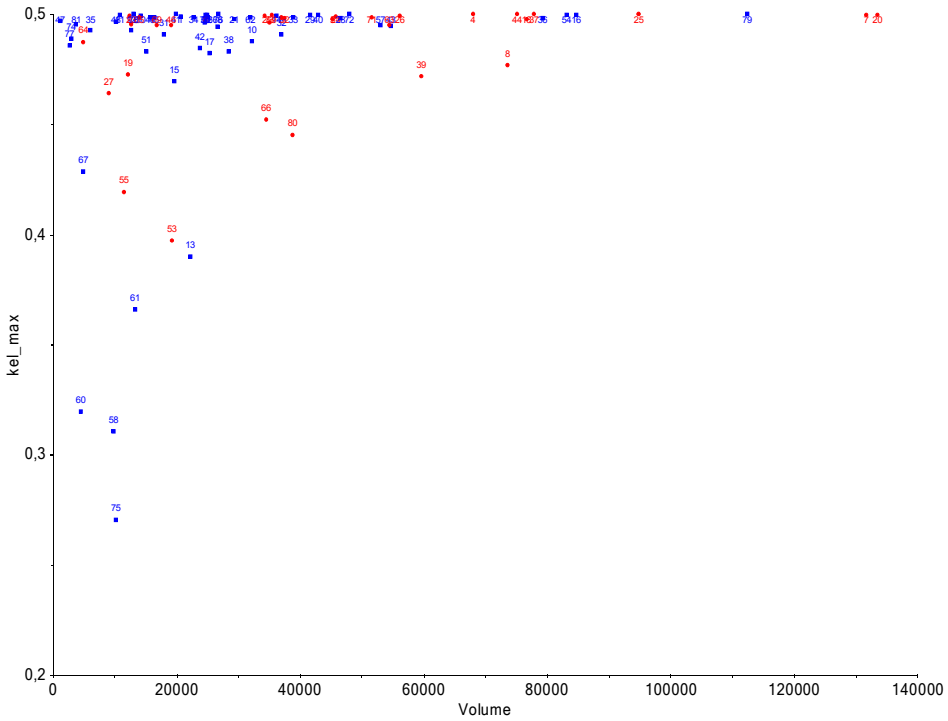
Ladningsplotta undersøkjast også for å sjå etter variablar som korrelerer. Ettersom alder og stadie har motsett forteikn i fleire ladningaplott, kan det tyde på at yngre pasientar som regel har høgare stadie og omvendt. Spreiingsplott av alle pasientane med alder og stadie langs aksane, vist i *figur 32*, syner derimot ikkje denne tendensen. Volum og stadie kan sjå ut til å korrelere, det vil seie at stort volum også indikerer høgt stadie og omvendt. Eit spreiingsplott av stadie mot volum er synt i *figur 33*. Denne syner at eit fleirtal av svulstane med lite volum også har lågt stadie, men at tendensen ikkje er veldig sterk. Volum ser i tillegg ut til å samanfalle med maksimalverdien av k_{el} over svulsten ($k_{el\ max}$) i ladningsplotta. Spreiingsplottet i *figur 34* syner at svulstar med lite volum oftare har låg verdi av $k_{el\ max}$ enn det dei store svulstane har. Det vil seie at kontrastmiddelet vert vaska saktare ut av små svulstar, med andre ord har dei små svulstane svakare blodgjennomstrøyming.



Figur 32: Stadie (figo) plotta mot alder. Raude punkt svarar til pasientar som får tilbakefall, medan blå punkt svarar til pasientar som vert friske att. Laga med Unscrambler.



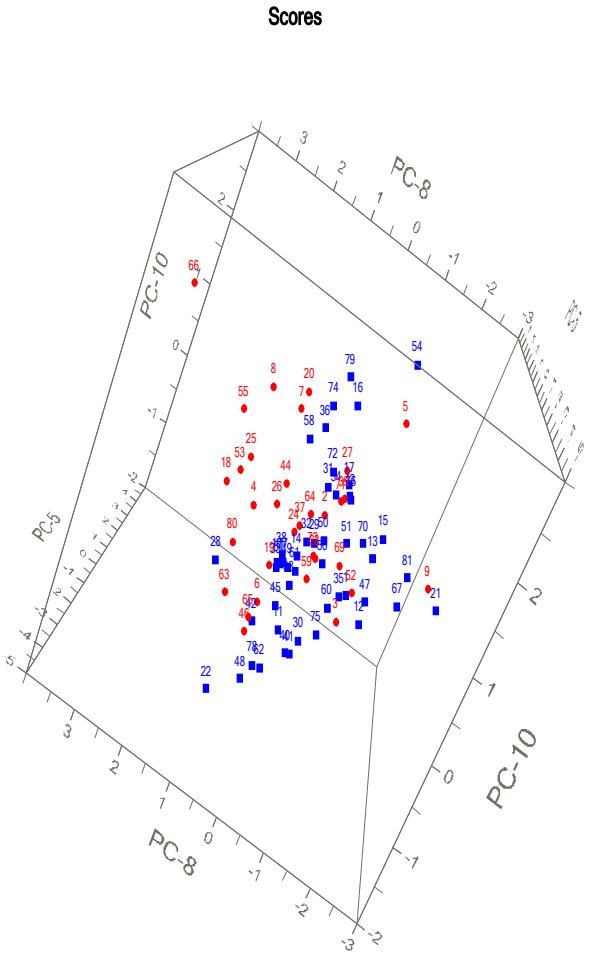
Figur 33: Stadie (figo, med verdiane 1-4) plotta mot svulstvolum (mm³). Dei raude punkta svarar til pasientar med tilbakefall (pfs = 1), medan dei blå svarar til pasientar som vert friske att (pfs = 0). Laga med Unscrambler.



Figur 34: Maksimalverdi av utvaskingsraten k_{el} plotta mot svulstvolum (mm³). Dei raude punkta syner pasientar med tilbakefall, medan dei blå er pasientar som vart friske att. Laga med Unscrambler.

4.1.2 Skårar

Skårplotta, synt i *figur 29 – 31*, undersøkjast for grupperingar av objekt som samsvarar dei to hovudgruppene i behandlingsutfall, det vil seie dei som vart friske (*pfs* = 0) og dei som fekk tilbakefall (*pfs* = 1). Denne modellen syner ingen slike tydelege grupper. Den einaste antydninga til gruppeinndeling kan finnast i det tredimensjonale skårplottet for komponentane PC-5, PC-8 og PC-10, vist i *figur 35*. Undersøking av ladningsplott for PC-8, sjå *figur 30*, syner at høge verdiar av PC-8 blant anna er knytta til stadiet svulsten hadde før behandlinga. Skårplottet indikerer, ikkje uventa, at pasientar med lågt stadie oftare vert friske enn pasientar med høgt stadie. Fleire tredimensjonale skårplott er synte i vedlegget, *avsnitt 7.3*.



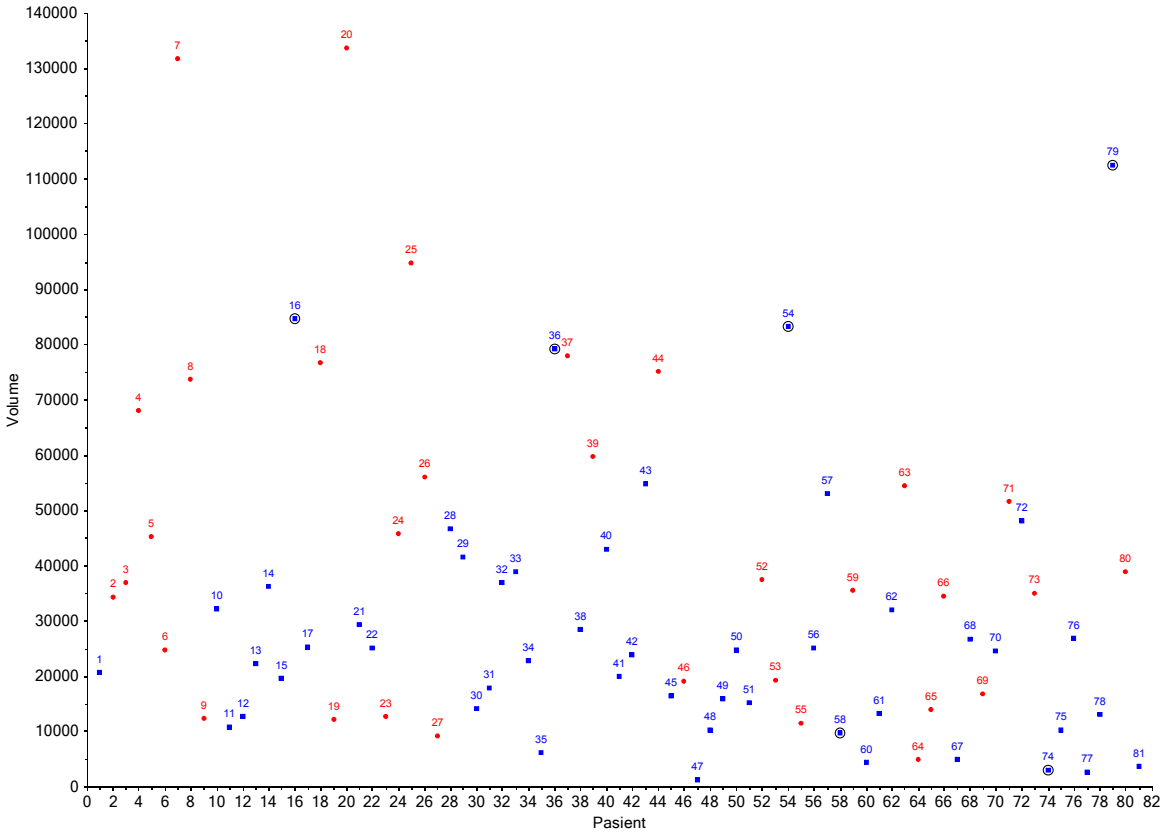
Figur 35: Skårplott for komponentane PC-5, PC-8 og PC-10 for PCA-modellen av alder, stadie, volum og statistiske parameterar av A, k_{ep} og k_{el}. Blå punkt svarar til pasientar som vart friske att (pfs = 0), medan raude punkt svarar til pasientar som fekk tilbakefall (pfs = 1). Pasientane som vart friske att ser ut til å stort sett ha negative skårar i PC-8.
Laga med Unscrambler.

Ladningsplottet i *figur 30* tyder også på at pasientane med negative verdiar av PC-8 tek opp lite kontrastmiddel i mesteparten av svulsten, i tillegg til at opptaksraten k_{ep} er stor i mange vokslar. Det kan altså vere at pasientar med svulstar som tek opp raskt men lite kontrastmiddel, har større sjanse for å bli friske.

I 3D-plottet vert også fleire pasientar med *pfs* = 1 (raud) plasserte saman med dei friske pasientane (*pfs* = 0). Pasientane 3, 9, 52 og 69 undersøkjast for å sjå om dei skil seg frå dei andre pasientane med tilbakefall. Desse har noko låge verdiar av *A*- og k_{el}- parameterane, særleg dei høge persentilane, medan dei har høge verdiar av k_{ep}. Dette ser altså ut til å vere svulstar som tek opp kontrastmiddelet raskt.

Det er også ei klynge av pasientar med *pfs* = 0 øvst i plottet, blant dei raude punkta. Blant desse pasientane, 16, 36, 54, 58, 74 og 79, har fire langt større volum enn dei andre svulstane som vart helbreda. Dette kjem fram i *figur 36*, der dei seks pasientane er markerte.

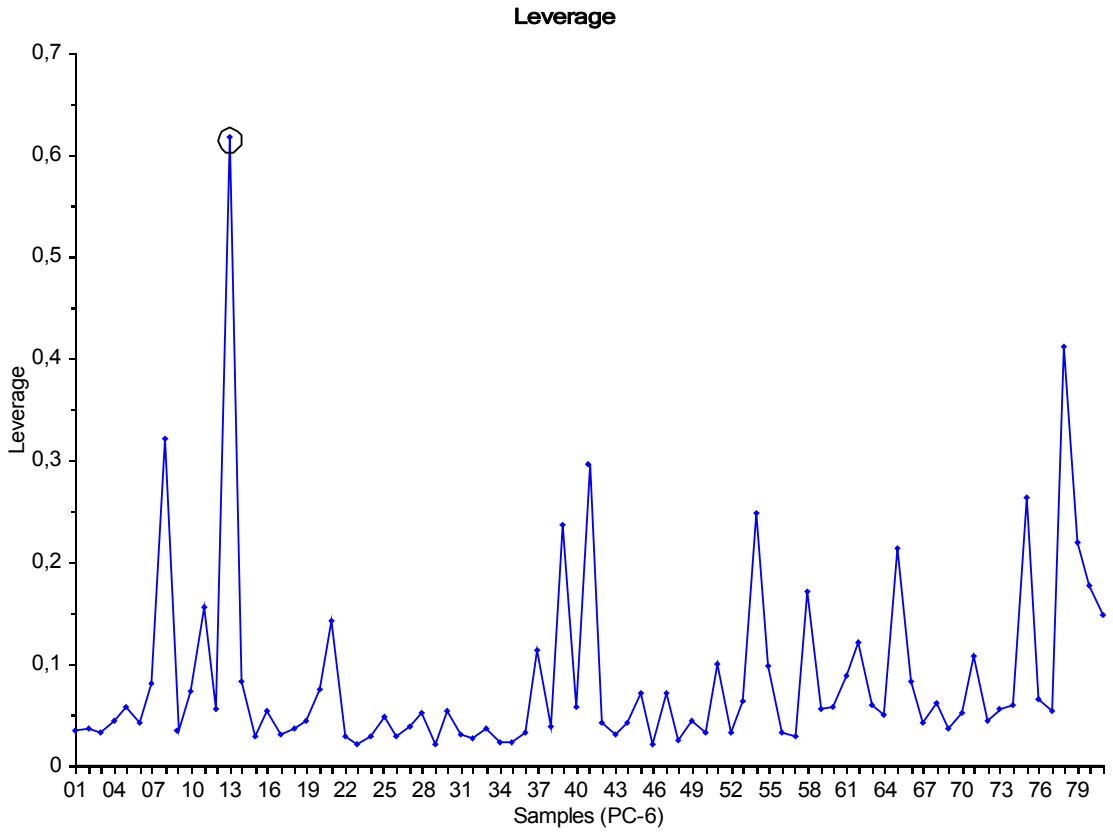
Pasient 28, som også er ein pasient som vert frisk, men er plassert saman med pasientane med tilbakefall i skårplottet, syner seg å ha noko stort svulstvolum (46 588 mm^3) og høgt stadie (IV). Denne pasienten vert altså frisk att, trass ein stor og alvorleg svulst.



Figur 36: Svulstvolum (mm^3) for kvar pasient. Dei raude punkta syner pasientar som får tilbakefall, medan dei blå punkta syner pasientar som vert friske att. Dei markerte punkta er pasientar som vert friske att, og som ligg saman med pasientar med tilbakefall i skårplottet av PC-5, PC-8 og PC-10, figur 35. Laga med Unscrambler.

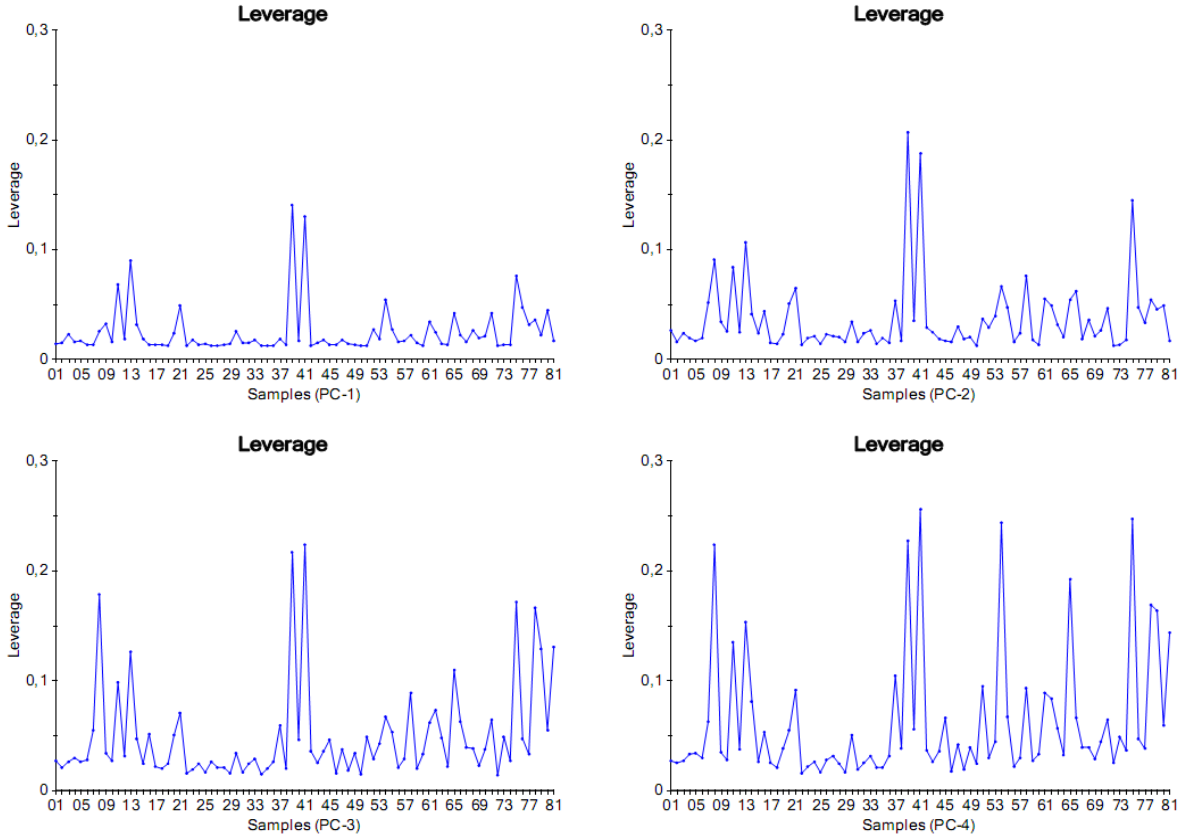
Skårplottet for PC-6, *figur 30*, syner at pasient nr 13 skil seg tydeleg frå dei andre. Dersom vi ser på momentplottet for denne prinsipalkomponenten, *figur 37*, har denne pasienten svært høgt moment i denne komponenten. Dette tyder på at pasient 13 er forskjellig frå dei andre, og at PC-6 i stor grad brukast til å forklare akkurat denne svulsten.

Nærare undersøking av verdiane for pasient 13 syner at denne er blant dei eldste, 76 år, og har eit svulstvolum som er noko mindre enn gjennomsnittet (22 280 mm^3, medan gjennomsnittet er 34 913 mm^3). Verdiane for opptaksraten k_{ep} er svært små, det vil seie at alle persentilane til k_{ep} er låge, samt at stadardavviket er lite. Det kan tyde på at alle vokslane har låg opptaksrate, noko som også kan sjåast av dei høge verdiane for skeivskap og kurtose. k_{ep}-fordelinga til denne svulsten er såleis veldig konsentrert om dei låge verdiane, noko som tydar på låg permeabilitet i vevet. Det vil seie at lite kontrastmiddel trenger gjennom vevet, og at der såleis er lite lekkasje. Svulsten har i tillegg høge verdiar for persentilane til amplituden *A*.



Figur 37: Momentplott for PC-6 i PCA-modellen av alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Pasient nr 13 (markert med sirkel) utmerkar seg med særleg høgt moment. Laga med Unscrambler.

Undersøking av momentplott, sjå *figur 38*, syner at det ikkje er store momentproblem i modellen. Med unntak av èin pasient, pasient nr 13, har ingen av pasientane moment over 0,5 før ein kjem opp i høge komponentar (PC-16). Pasient nr 13 skil seg ut ved å ha mykje høgare moment enn dei andre.



Figur 38: Momentplott for dei fire første komponentane i PCA-modellen av alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el}. Plotta syner at ingen av objekta har særleg høgt moment i desse komponentane. Laga med Unscrambler.

Denne PCA-modellen forklarer altså datagrunnlaget godt, men kan i liten grad nyttast til å skilje mellom pasientar som blir friske og pasientar som får tilbakefall.

Vi har også laga desse PCA-modellane:

- Alder, volum, stadie og statistiske parameterar for *A*.
- Alder, volum, stadie og statistiske parameterar for k_{ep}.
- Alder, volum, stadie og statistiske parameterar for k_{el}.
- Alder, volum og stadie.

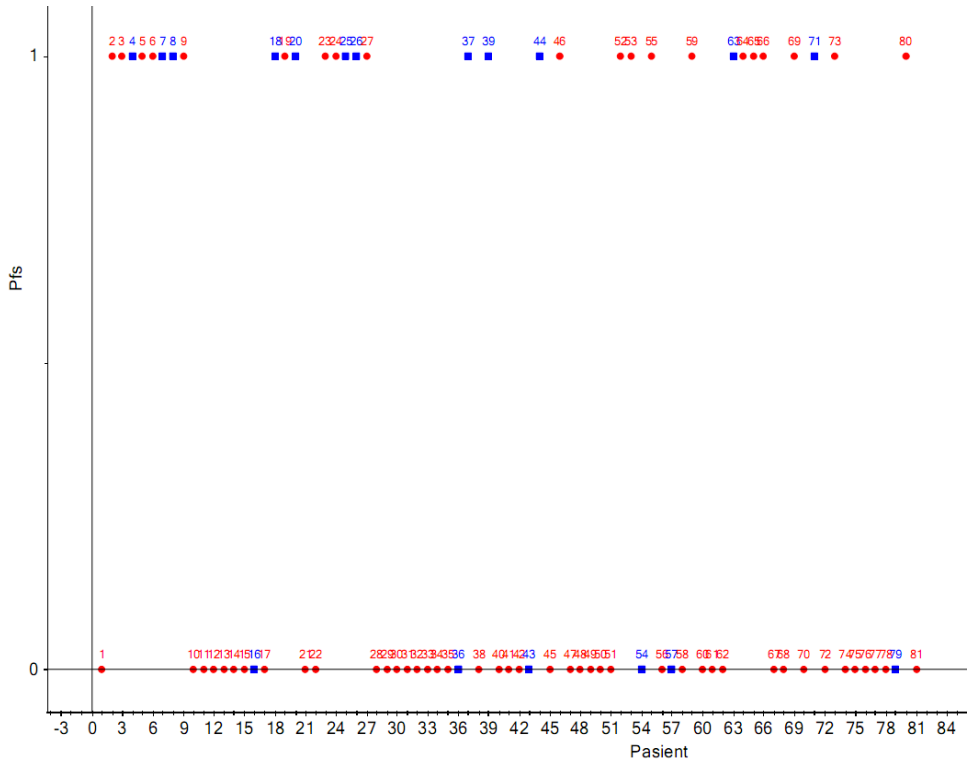
Ingen av desse modellane lukkast betre i å skilje pasientane med og utan tilbakefall. Residualvariansplott, samt nokre plott av skårar og ladningar for desse modellane finnast i vedlegget, *avsnitt 7.3*.

4.2 Klyngeanalyse

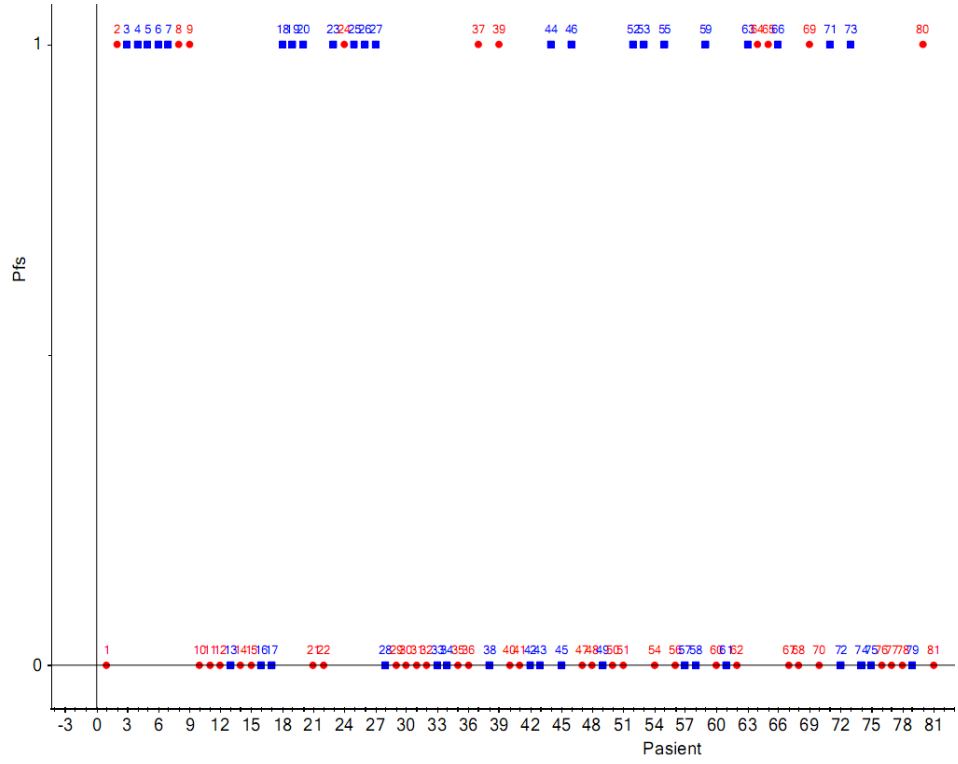
To ulike K-means-klyngeanalysar er gjennomførte, èin med dei statistiske parameterane pluss alder, volum og stadie som forklaringsvariablar, og èin med skårar frå utvalde prinsipalkomponentar frå PCA-modellen som forklaringsvariablar. Prinsipalkomponentane PC-5, PC-8 og PC-10 nyttast fordi desse skil best mellom dei to ulike utfalla, *pfs* = 0 og *pfs* = 1, i skårplotta for PCA-modellen. Dette er synt i *figur 35*.

Som synt i *figurane 39* og *40*, gjev ingen av dei to klyngemodellane gruppeinndeling som svarar til behandlingsutfall. Begge modellane deler pasientane inn i to grupper, men inndelinga samsvarar ikkje med utfall, og ser heller ikkje ut til å skildre volum, stadie, alder eller andre parameterar.

Algoritmen K-medians vart også forsøkt, sidan denne skal vere meir robust mot avvikarar, det vil seie objekt som skil seg veldig frå dei andre. Heller ikkje dette gav grupper som samsvarar med utfall. Plott av desse modellane er synte i vedlegget, *kapittel 7.3*.



Figur 39: K-means-klynger basert på alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Pasientnummer langs x-aksen og progresjonsfri overleving langs y-aksen. Dei ulike fargane (raud og blå) syner dei to klyngene. Laga med Unscrambler.



Figur 40: K-means-klynger basert på tre prinispalkomponentar (5, 8 og 10) frå PCA-modellen. Pasientnummer langs x-aksen og progresjonsfri overleving langs y-aksen. Dei ulike fargane (raud og blå) syner dei to klyngene. Laga med Unscrambler.

4.3 Diskriminant analyse

Skårmatrisene for prinsipalkomponentane PC-5, PC-8 og PC-10 nyttast som forklaringsvariablar i diskriminant analyse med progresjonsfri overleving som respons. Sidan variablane er PCA-skårar, nyttar vi ikkje noko vekting. Analysane er gjennomførte med tre ulike algoritmar: lineær, kvadratisk og Mahalanobis. For alle tre prøvde vi både å gå ut i frå likt sannsyn for begge grupper, samt å rekne ut sannsyna frå datasettet. Unscrambler har ikkje validering for diskriminant analyse, så *tabellane 11 – 16* er kalibreringsresultat.

Tabell 11: Lineær diskriminant analyse (kalibrering) der ein går ut i frå likt sannsyn for begge utfall.

Nøyaktigheit: 69% Sensitivitet: 73% Spesifisitet: 63%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	36	12
	1	13	20

Tabell 12: Lineær diskriminant analyse (kalibrering) der sannsynet for kvart utfall er kalkulert ut i frå data.

Nøyaktigheit: 69% Sensitivitet: 90% Spesifisitet: 38%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	44	20
	1	5	12

Tabell 13: Kvadratisk diskriminant analyse (kalibrering) der ein går ut i frå likt sannsyn for begge utfall.

Nøyaktigheit: 69% Sensitivitet: 76% Spesifisitet: 59%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	37	13
	1	12	19

Tabell 14: Kvadratisk diskriminant analyse (kalibrering) der sannsynet for kvart utfall er kalkulert ut i frå data.

Nøyaktigheit: 70% Sensitivitet: 92% Spesifisitet: 38%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	45	20
	1	4	12

Tabell 15: Mahalanobis diskriminant analyse (kalibrering) der ein går ut i frå likt sannsyn for begge utfall.

Nøyaktigheit: 67% Sensitivitet: 69% Spesifisitet: 63%		**Faktisk verdi**	
		0	1
Predikert verdi	**0**	34	12
	1	15	20

Tabell 16: Mahalanobis diskriminant analyse (kalibrering) der sannsynet for kvart utfall er kalkulert ut i frå data.

Nøyaktigheit: 67% Sensitivitet: 69% Spesifisitet: 63%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	34	12
	1	15	20

Det ser ut til at modellane har lettare for å plassere pasientane som vert friske i riktig gruppe, men feilgrupperer oftare pasientane med tilbakefall. Den mest lovande metoden, QDA med sannsyn kalkulert ut i frå data, nyttast i tillegg på alle dei ti første prinsipalkomponentane. Resultatet av dette er gitt i *tabell 17*. Dette gjev ei langt betre klassifisering, særleg for pasientane med tilbakefall. Dette betyr at informasjonen som skil pasientane med tilbakefall frå dei som vert friske kan ligge ikkje i PC-5, PC-8 eller PC-10, men i andre av dei ti første prinsipalkomponentane.

Tabell 17: Kvadratisk diskriminant analyse (kalibrering) der sannsynet for kvart utfall er kalkulert ut i frå data. Dei ti første komponentane frå PCA-modellen, samt alder, stadie og volum, er nytta som forklaringsvariablar.

Nøyaktigheit: 84% Sensitivitet: 80% Spesifisitet: 91%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	39	3
	1	10	29

For å få validerte LDA-resultat, nyttar vi eit Matlab-skript for LDA laga av Ulf Geir Indahl, *[41]*. Dette skriptet nyttar full kryssvalidering, slik at vi får forvirringsmatriser for både kalibrering og validering. I tillegg bereknast ein p-verdi for kvar klasse, som syner om klassifiseringa frå modellen er signifikant i samanlikning med å tilfeldig plassere svulstane i dei to klassane. P-verdien finnast ved å stokke tilfeldig om på rekkefølgja av pasientane i responsvektoren, og tilpasse modellen til desse. Dette gjerast eit høgt tal gonger, her 10 000. Resultata frå desse tilpassingane vert så samanlikna med resultata frå den ekte modeltilpassinga, og p-verdien reknast ut.

Dette skriptet nyttar ein variabelseleksjon, der variablane som trengst for å forklare ein viss prosent av variansen i datasettet veljast ut. LDA utførast så med desse utvalde variablane som forklaringsvariablar. Denne reduksjonen i talet på variablar, gjer at vi kan nytte LDA på alder, stadie, volum og statistiske parameterar direkte, i staden for å nytte prinsipalkomponentar. Ei oversikt over andel forklart varians for kvar variabel er gitt i *tabell 24*.

Tabell 18: LDA (kalibrering) der variablane som trengst for å forklare 99% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 77% Sensitivitet: 76% Spesifisitet: 78%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	37	7
	1	12	25

Tabell 19: LDA (validering) der variablane som trengst for å forklare 99% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 62% Sensitivitet: 63% Spesifisitet: 59%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	31	13
	1	18	19

Resultatet av klassifiseringa som nytta dei variablane som trengst for å forklare 99% av variasjonen i datasettet, gav forvirringsmatrisa som er synt i *tabell 18*. Nøyaktigheita er 77%, og er til skilnad frå fleire av dei andre diskriminantanalysane om lag like god for begge utfalla. Forvirringsmatrisa etter full kryssvalidering er gitt i *tabell 19*. Denne syner eit langt svakare resultat, der fleire pasientar er feilklassifiserte. P-verdiane vart berekna til å vere 0,316 for pasientane som vart friske og 0,072 for pasientane med tilbakefall. Det vil seie at om ein vel eit signifikansnivå på 5%, er prediksjonane frå modellen ikkje signifikant annleis enn tipping for nokre av klassane.

For å teste LDA med endå færre variablar, køyrer vi LDA-skriptet med dei variablane som krevjast for å forklare 90% av variansen i datasettet. Forvirringmatrisa er synt i *tabell 20*, medan *tabell 21* syner forvirringsmatrisa etter full kryssvalidering. Reduksjonen av talet på variablar har ikkje hatt stor innverknad på kalibreringsresultatet, det er om lag heilt likt det vi fekk då vi nytta variablar som forklarte 99% av variansen. Resultatet frå kryssvalideringa er derimot betre, noko som kan tyde på at klassifiseringa er mindre avhengig av enkeltsvulstar no som den byggjer på færre variablar. Også p-verdiane er lågare, 0,011 for begge gruppene. Det vil seie at LDA-klassifisering med desse variablane signifikant klassifiserer begge utfall.

Tabell 20: LDA (kalibrering) der variablane som trengst for å forklare 90% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 77% Sensitivitet: 78% Spesifisitet: 75%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	38	8
	1	11	24

Tabell 21: LDA (validering) der variablane som trengst for å forklare 90% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 67% Sensitivitet: 67% Spesifisitet: 65%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	33	11
	1	16	21

Det at LDA-modellen basert på få av variablane gjev betre resultat enn den som baserer seg på fleire, kan indikere at mange av variablane i datasettet ikkje kan knyttast mot utfall av stråleterapi. Ved å inkludere for mange variablar, innfører vi støy som gjer det vanskelegare å predikere utfall.

LDA med variablane som forklarer 95% av variansen gjev i kalibrering nøyaktigheit 75%, sensitivitet 78% og spesifisitet 72%. Etter full kryssvaldering er nøyaktigheita 58%, sensitiviteten 63% og spesifisiteten 50%. P-verdien er 0,1076 for gruppa av pasientar som vert friske, medan den er 0,1721 for dei som får tilbakefall. Dette er ikkje signifikant ved 5% signifikansnivå. Med andre ord er det berre modellen basert på 90% av variablane som skil signifikant mellom gruppene.

Tabell 22: LDA (kalibrering) der variablane som trengst for å forklare 95% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 75% Sensitivitet: 78% Spesifisitet: 72%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	38	9
	1	11	23

Tabell 23: LDA (validering) der variablane som trengst for å forklare 95% av variansen i data er nytta som forklaringsvariablar for å predikere behandlingsutfallet pfs. Pfs = 0 betyr at pasienten vart frisk, medan pasientar med pfs = 1 fekk tilbakefall.

Nøyaktigheit: 58% Sensitivitet: 63% Spesifisitet: 50%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	31	16
	1	18	16

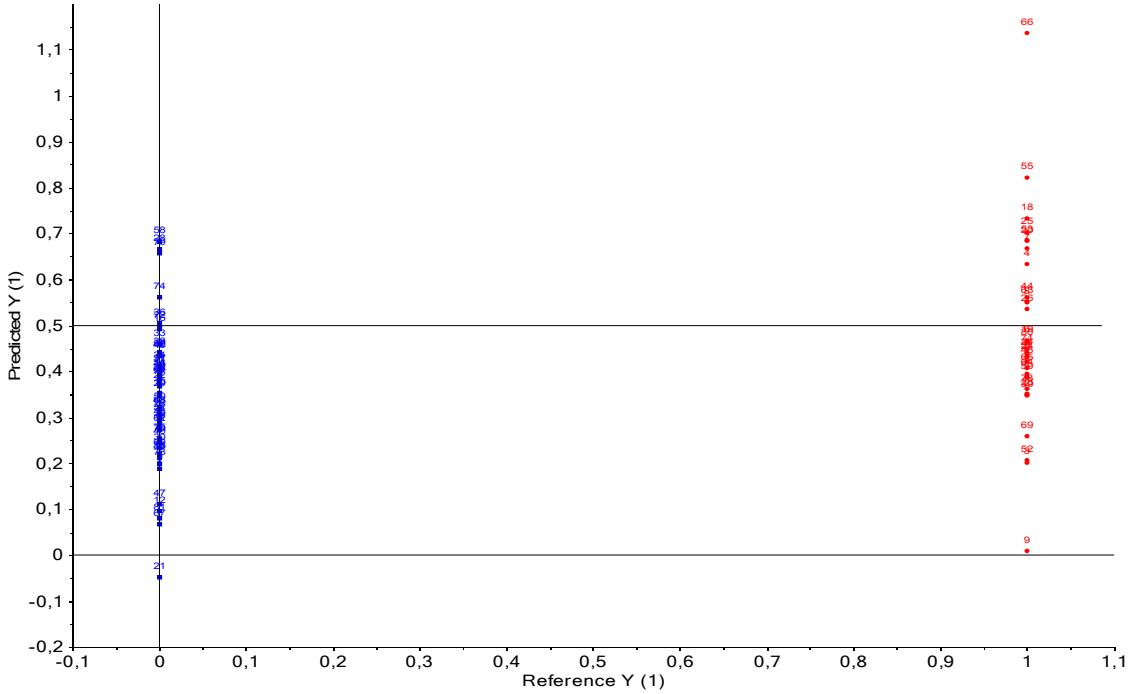
Variabelseleksjonen syner at få av variablane kan forklare store delar av variasjonen i data. Andel forklart varians for dei viktigaste variablane er synt i *tabell 24*. Dei to variablane som forklarer mest, er 90%-persentilen til A og 30%-persentilen til k_{ep}. Desse forklarar til saman 55% av variansen i forklaringsvariablane. Desse har også store ladningar i første komponent av PCA-modellen, synt i ladningsplottet i *figur 29*.

Tabell 24: Andel forklart varians for dei viktigaste variablane.

Variabel	Andel forklart varians	Samla forklart varians
A_90	37,51%	37,51%
kep_30	17,67%	55,18%
kel_mean	13,48%	68,66%
A_30	6,92%	75,58%
kep_7525	3,12%	78,70%
kep_skew	2,85%	81,55%
A_min	2,01%	83,56%
A_skew	1,82%	85,38%
li(lymfeknutar)	1,62%	87,00%
kel_25	1,55%	88,55%
Figo (stadie)	1,46%	90,01%
kel_max	1,35%	91,36%
Alder	1,27%	92,63%
Volum	1,16%	93,79%
kep_min	1,00%	94,79%
A_mode	0,95%	95,74%
kep_mode	0,74%	96,48%
kel_10	0,72%	97,20%
kel_60	0,51%	97,71%
kep_max	0,45%	98,16%
A_max	0,41%	98,57%
kel_kurt	0,39%	98,96%
A_7525	0,26%	99,22%

4.4 Regresjon

Dei mest lovande prinsipalkomponentane, PC-5, PC-8 og PC-10, nyttast som forklaringsvariablar i regresjonsanalyse med progresjonsfri overleving som respons. Til validering nyttast *leverage correction*, ein metode som for regresjon tilsvarar full kryssvalidering, *[42]*. Ei innføring i denne valideringsmetoden er gitt i Esbensen, *[33]*.



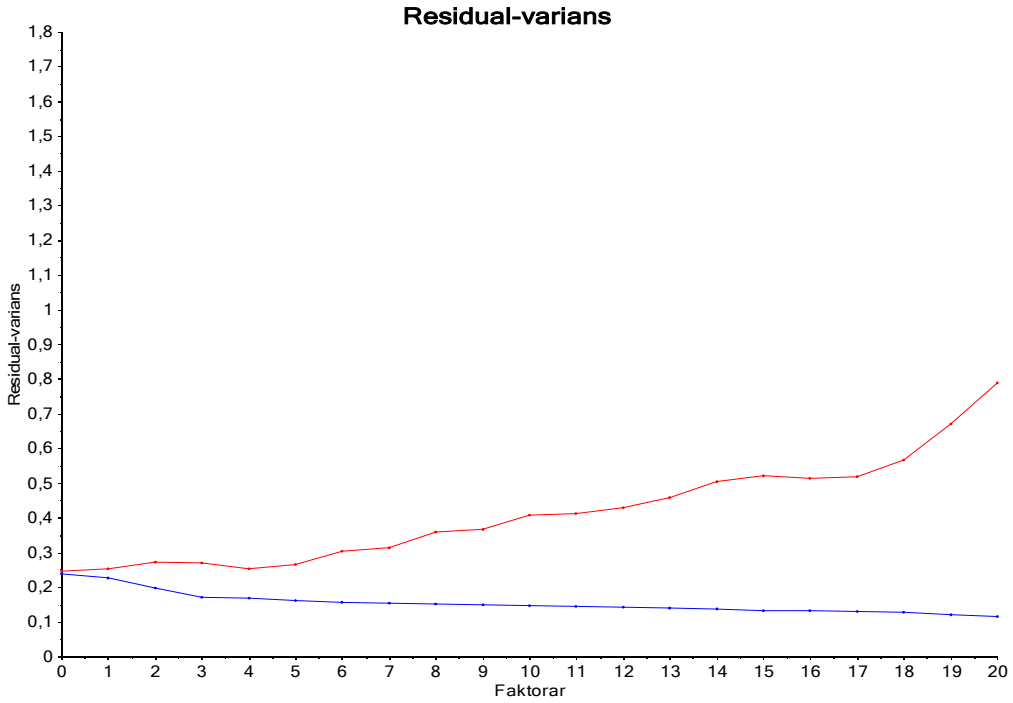
Figur 41: Predikert verdi (y-akse) mot faktisk verdi (x-akse) for regresjonsmodellen med PC-5, PC-8 og PC-10 som forklaringsvariablar og progresjonsfri overleving (pfs) som respons. Den svarte linja syner predikert y = 0,5. Raude punkt er pasientar med pfs = 1, medan blå punkt er pasientar med pfs = 0. Laga med Unscrambler.

Figur 41 syner predikert verdi mot faktisk verdi for modellen. Dersom modellen plasserer pasientane i riktig gruppe, skal den predikere ein verdi under 0,5 for pasientane med *pfs* = 0, og over 0,5 for pasientane med *pfs* = 1. Figuren syner at dei fleste pasientane, 44 av 49, eller 90%, med *pfs* = 0 er plassert riktig. For pasientane med *pfs* = 1, er derimot over halvparten plassert feil, 20 av 32, eller 63%. Modellen plasserer altså alt for mange pasientar i gruppa som bli friske att, og klarer ikkje å skilje ut dei som får tilbakefall. Ved å samanlikne *figur 41* med skårplottet i *figur 35*, kan ein sjå at dei pasientane som skil seg ut på skårplottet (det blå punkta som ligg blant dei raude, og dei raude punkta som ligg saman med dei blå), også vert feilklassifiserte i regresjonsmodellen.

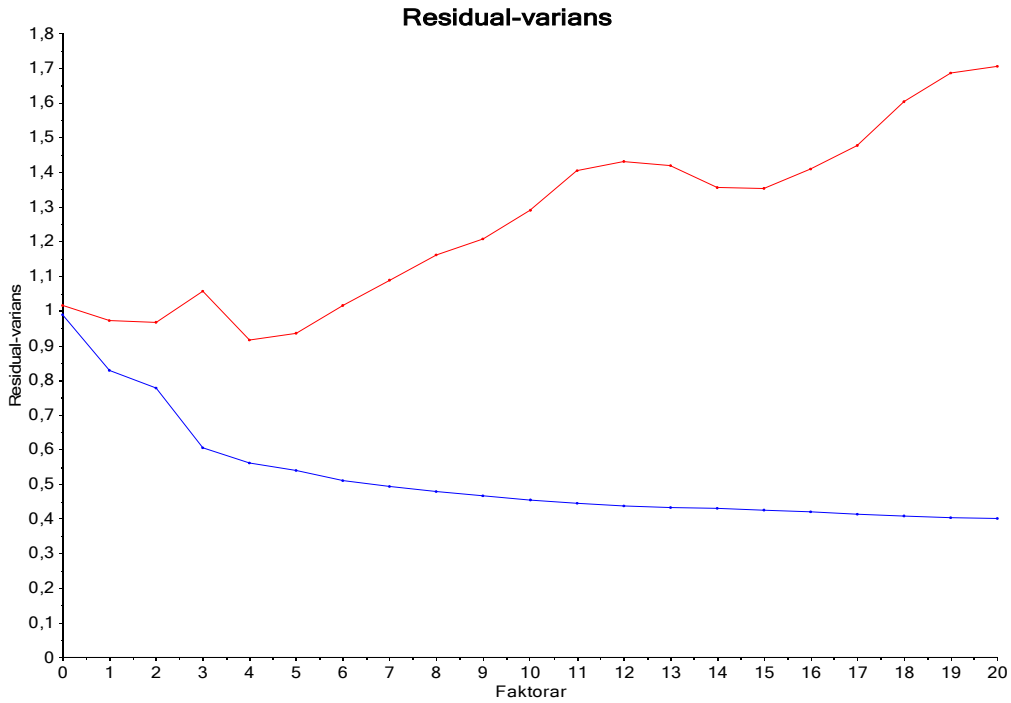
Vi har i tillegg gjort ein regresjon med dei ti første prinsipalkomponentane. Denne modellerer pasientane med tilbakefall noko betre, med 47% riktig i staden for 38% som i modellen med tre prinsipalkomponentar. Den er derimot noko svakare på pasientar som vert friske att, her plasserer den 80% riktig. Eit tilsvarande plott for denne modellen er synt i vedlegget, *avsnitt 7.3*.

4.5 PLS

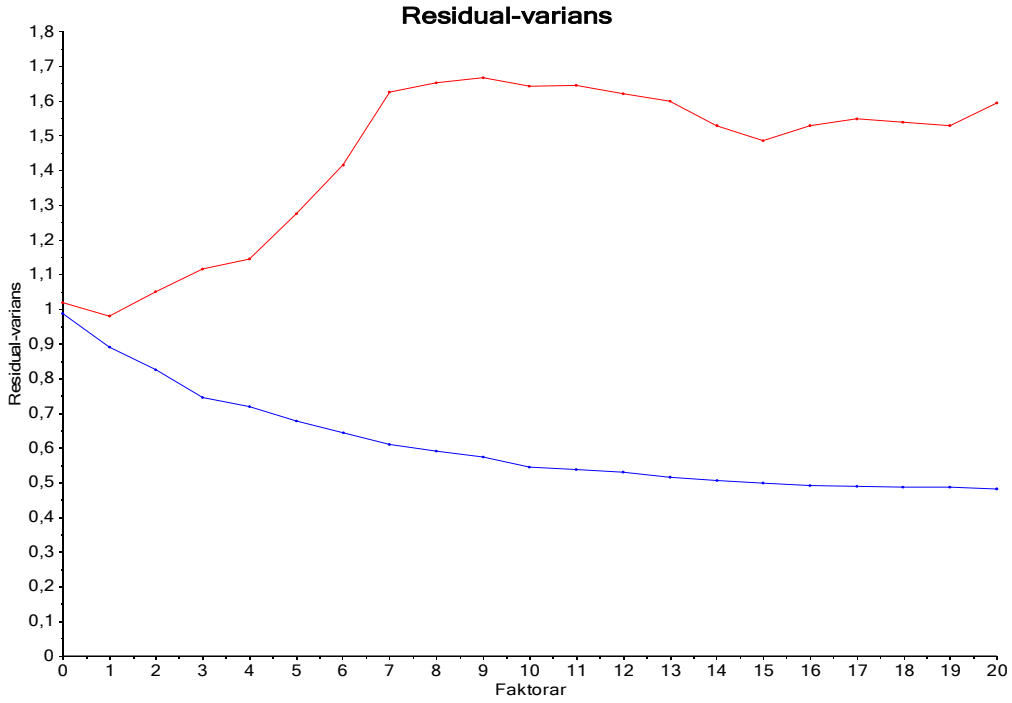
Det har blitt laga PLS-modellar med *pfs*, volum og stadie som responsvariablar og dei vanlege forklaringsvariablane. Alle modellane hadde låg forklaringsprosent, kring 50% til 60% av forklaringsvariabelen i kalibreringa, og særs svak validering, negativ forklaringsprosent for dei fleste komponentane, og kan difor ikkje nyttast. Residualvariansplotta for kalibrering og validering av dei tre modellane er synte i *figurane 42 - 44*.



Figur 42: Residualvariansplott for PLS-modell med alder, volum, stadie og statistiske parameterar for A, k_{ep} og k_{el} som forklaringsvariablar og progresjonsfri overleving som respons. Det er nytta full kryssvalidering. Den blå kurva syner residualvarians for kalibrering, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.



Figur 43: Residualvariansplott for PLS-modell med alder, stadie og statistiske parameterar for A, k_{ep} og k_{el} som forklaringsvariablar og volum som respons. Det er nytta full kryssvalidering. Den blå kurva syner residualvarians for kalibrering, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.



Figur 44: Residualvariansplott for PLS-modell med alder, volum og statistiske parameterar for A, k_{ep} og k_{el} som forklaringsvariablar og stadie som respons. Det er nytta full kryssvalidering. Den blå kurva syner residualvarians for sjølve modellen, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.

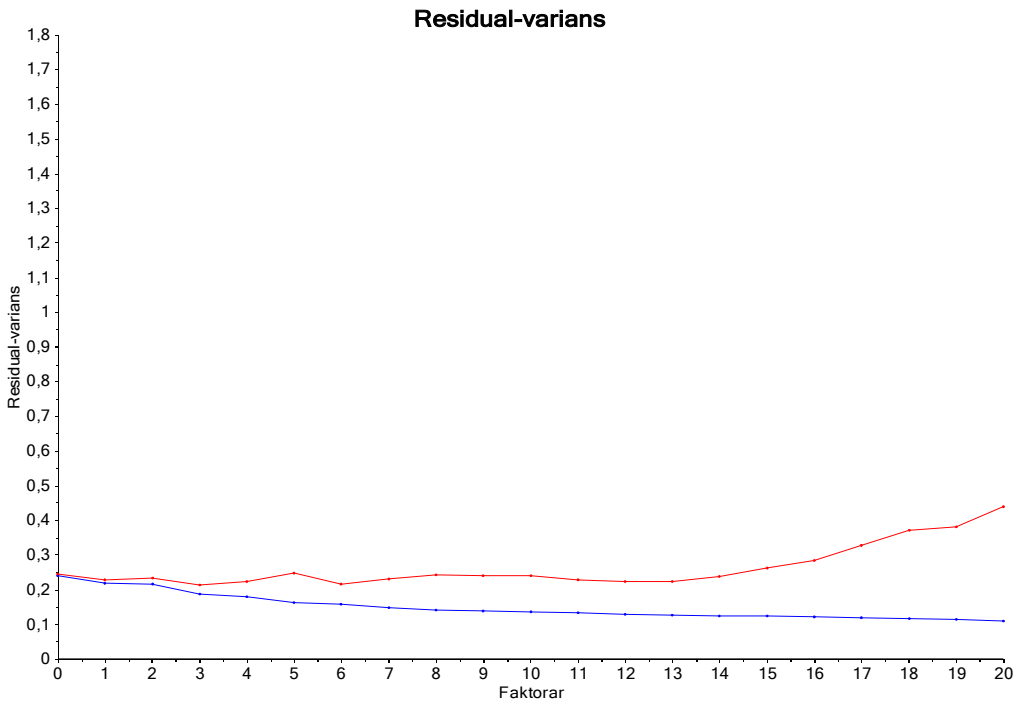
4.6 PLS med histogramverdiar

Brix-parameterane A, k_{ep} og k_{el} delast inn i intervall kvar for seg, for å lage histogram. Freedman og Diaconis sin regel, sjå *kapittel 3.4*, nyttast til å bestemme breidda av kvart intervall. Intervallbreidda er lik for alle pasientane, så om ein ser på det første intervallet til A for to pasientar, svarar dette til A-verdiar mellom 0 og 0,61 for begge svulstane.

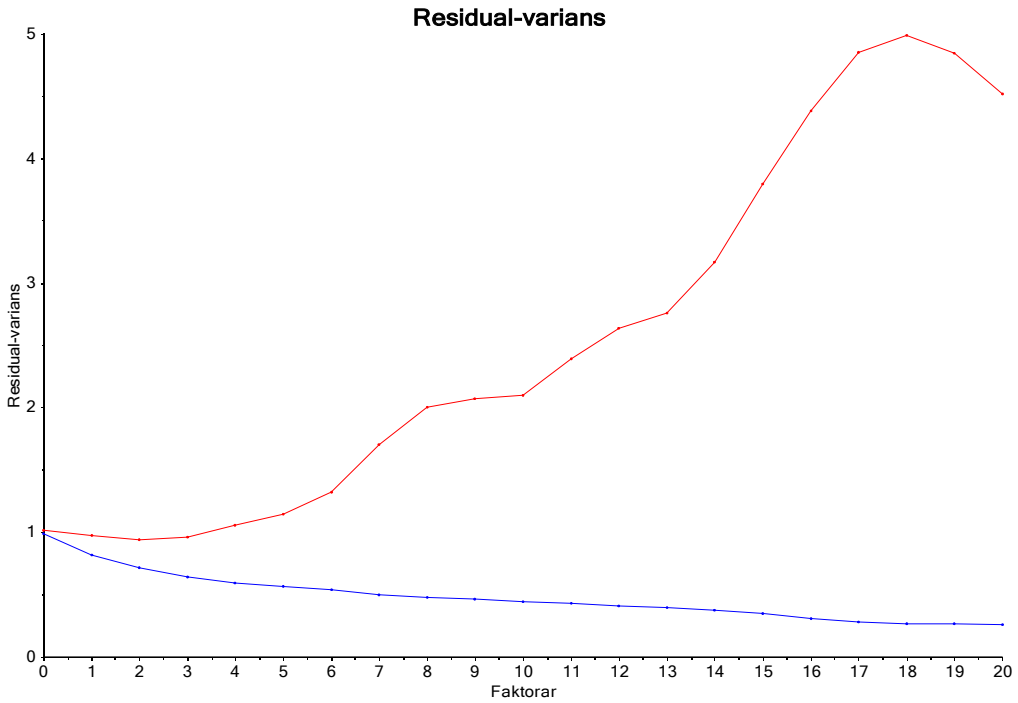
Tabell 25: Breidda av kvart intervall og talet på intervall i histogramma for kvar av parameterane A, k_{ep} og k_{el}. Berekna etter Freedman og Diaconis sin regel, [32].

	Minste verdi	**Største verdi**	**Intervallbreidd**	**Antal grupper**
A	0	20	0,61	33
k_{ep}	0	12	1,02	12
k_{el}	0	0,5	0,05	10

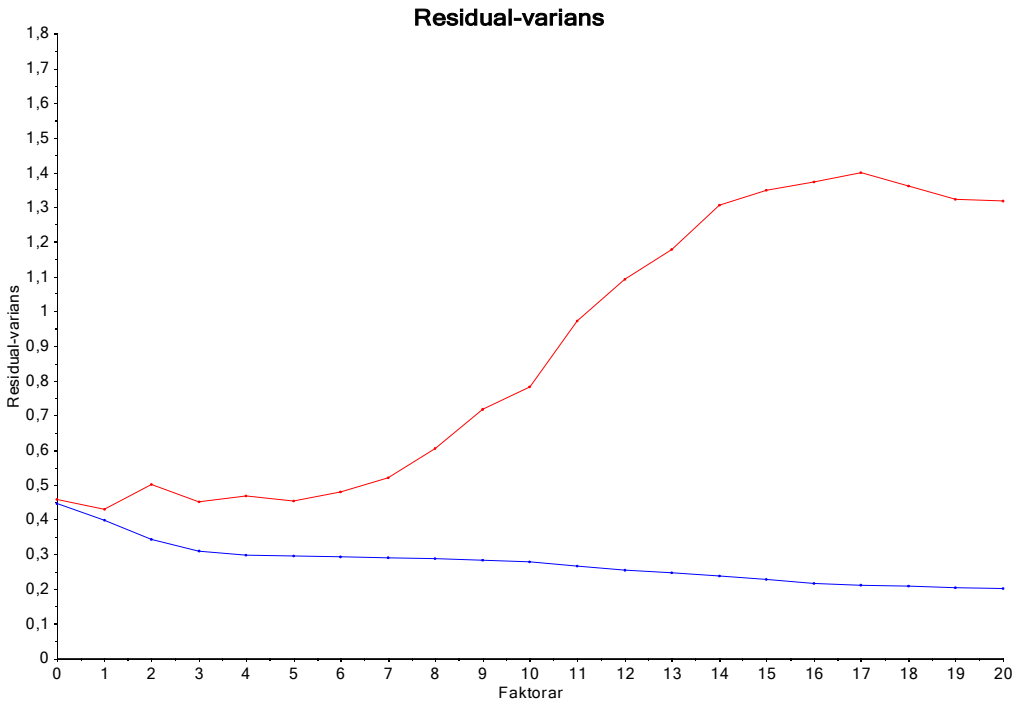
PLS med normaliserte histogramverdiar gav ikkje betre resultat enn dei andre PLS-analysane, noko ein kan sjå av residualvariansplotta i *figur 45 til 47*. Også her gjorde vi tre ulike analysar, der vi nytta progresjonsfri overleving, volum og stadie som respons.



Figur 45: PLS-analyse med alder, volum, stadie og histogramverdiar for A, k_{ep} og k_{el} som forklaringsvariablar og progresjonsfri overleving som respons. Det vart nytta full kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.



Figur 46: PLS-analyse med alder, stadie og histogramverdiar for A, k_{ep} og k_{el} som forklaringsvariablar og volum som respons. Det var nytta full kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.



Figur 47: PLS-analyse med alder, volum og histogramverdiar for A, k_{ep} og k_{el} som forklaringsvariablar og stadie som respons. Det vart nytta full kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude syner residualvarians for valideringa. Laga med Unscrambler.

4.7 SIMCA

SIMCA utførast i PLS_Toolbox med signifikansnivå 5%. Alder, stadie, volum, lymfeknuteinfiltrasjon og statistiske parameterar for A, k_{ep} og k_{el} nyttast som forklaringsvariablar, medan progresjonsfri overleving er respons. Variablane autoskalerast, det vil seie sentrerast og standardiserast som synt i likning (12), før analysen. Ved å undersøkje plott av residualvarians og PRESS (*predicted residual sum of squares*) for kvar av dei to PCA-modellane som vert laga, èin for *pfs* = 0 synt i *figur 48*, og èin for *pfs* = 1 synt i *figur 49*, finn vi at det bør nyttast 18 prinsipalkomponentar i modellen for pasientar med *pfs* = 0 og 15 prinsipalkomponentar i modellen for pasientar med *pfs* = 1.

Plott av Q- og T^2-verdiar er synte i *figur 50* (*pfs* = 0) og i *figur 51* (*pfs* = 1). Desse syner at 15 av dei 32 pasientane med *pfs* = 1, ligg innanfor signifikansgrensene til modellen for *pfs* = 0. Berre to av pasientane med *pfs* = 0 ligg innanfor signifikansgrensene til modellen med *pfs* = 1.

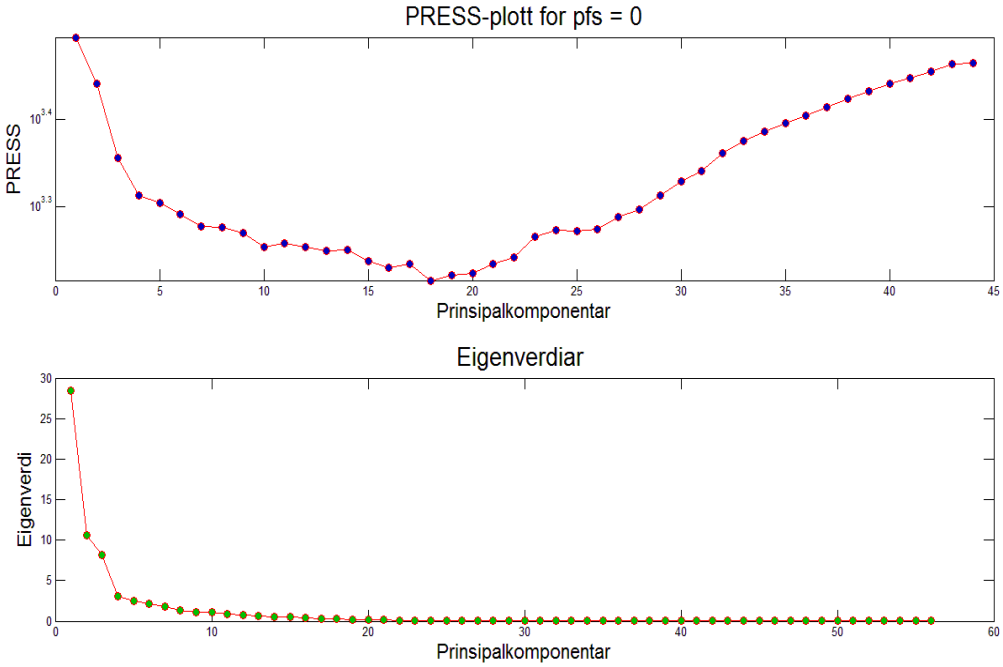
Fleire pasientar vert klassifiserte i begge klassar samtidig, eller i ingen av klassane. Dersom ein krev at alle pasientar skal plasserast i èin klasse, og såleis uansett plasserer kvar pasient i den gruppa den ligg nærmast, vert alle pasientar som vert friske att plasserte i riktig gruppe, medan sju av dei som fekk tilbakefall vert feilklassifiserte. Dette gjev ei nøyaktigheit på 91%, som synt i *tabell 26*. SIMCA klassifiserer med andre ord betre enn LDA og QDA.

Dei feilklassifierte objekta undersøkjast nærare for å sjå kva som skil desse frå dei andre objekta. Det ser ut til at desse sju er heilt gjennomsnittlege både når det gjeld alder, stadie, volum og Brix-parameterar. Ein mogleg årsak til at desse vert feilklassifiserte kan rett og slett vere at dei skil seg for lite ut, slik at dei ikkje vert plukka opp som tilbakefallspasientar av modellen.

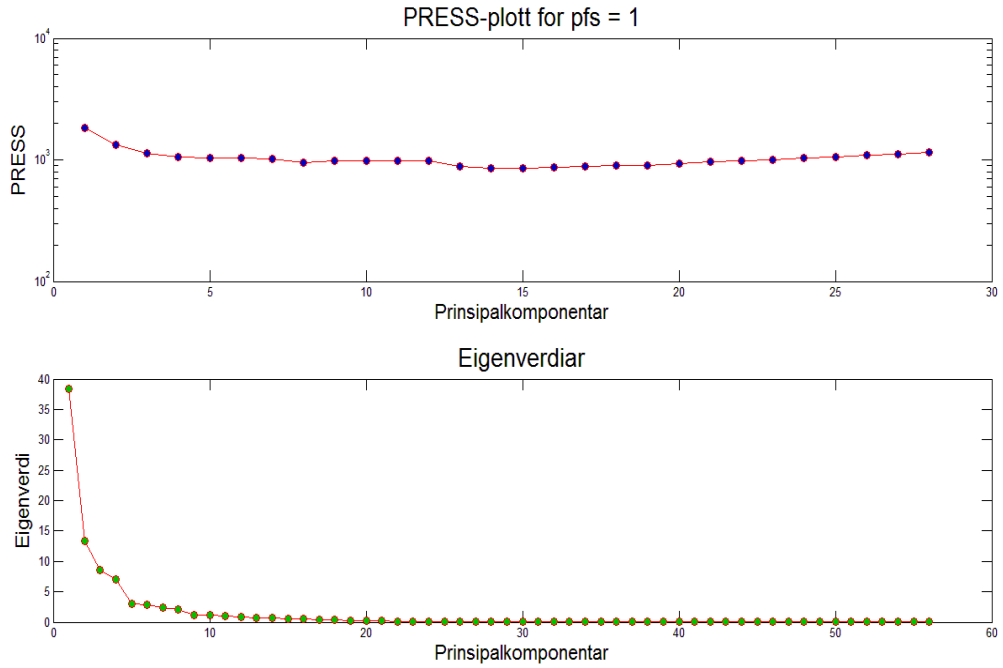
Tabell 26: Resultat av SIMCA (kalibrering). Gruppe 0 er pasientane som vart friske att, medan gruppe 1 er dei som fekk tilbakefall.

Nøyaktigheit: 91% Sensitivitet: 100% Spesifisitet: 78%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	49	7
	1	0	25

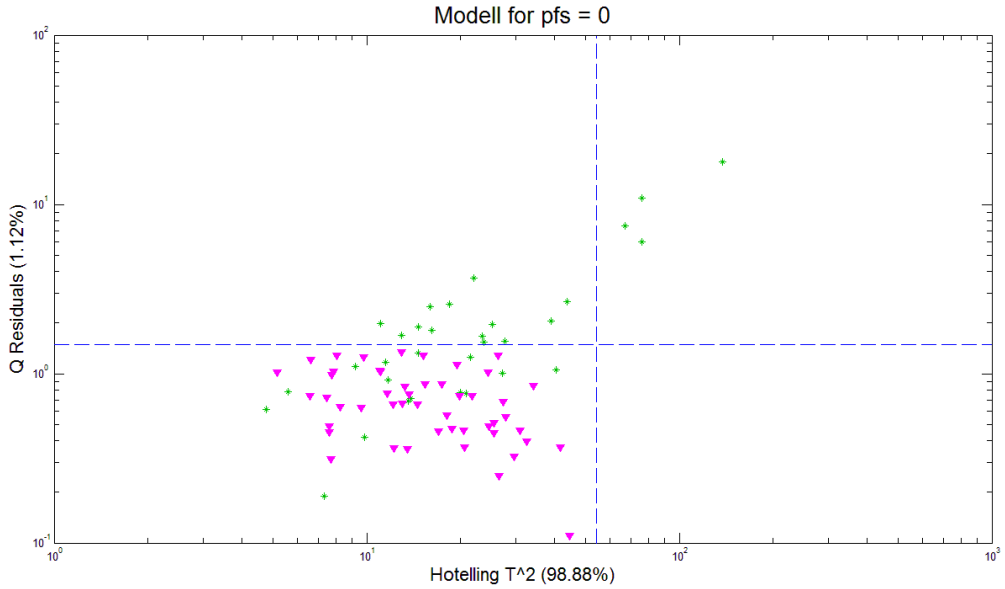
Matlab-skriptet nytta til SIMCA er gitt i vedlegget, *avsnitt 7.1*.



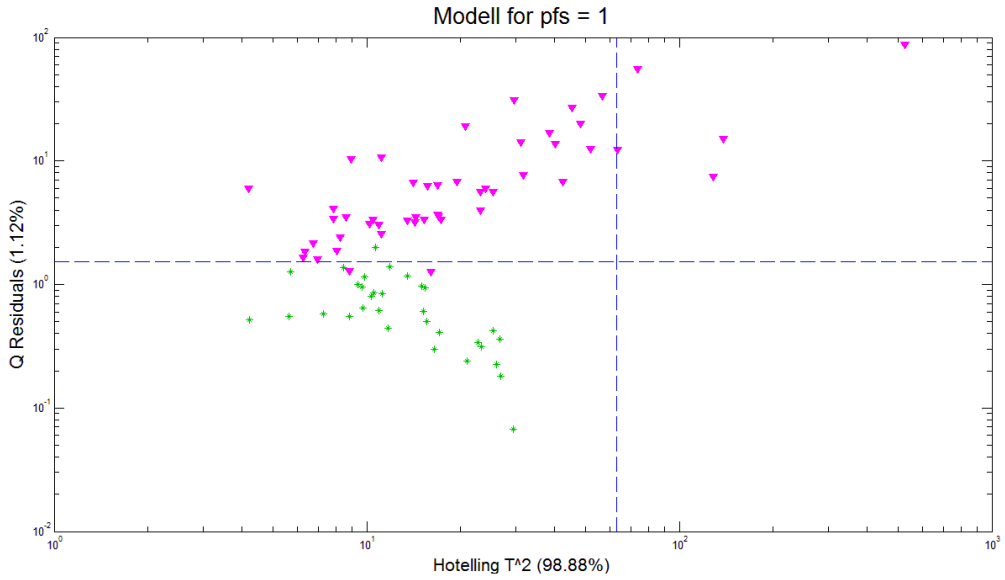
Figur 48: PRESS og eigenverdiar for kvar prinsipalkomponent i PCA-modellen av svulstane med pfs = 0, det vil seie dei som vart kurerte. Eigenverdien til kvar prinsipalkomponent syner kor stor del av variansen i datasettet som vert forklart av denne komponenten. Laga med PLS_Toolbox i Matlab.



Figur 49: PRESS og eigenverdiar for kvar prinsipalkomponent i PCA-modellen av svulstane med pfs = 1, det vil seie dei som fekk tilbakefall. Eigenverdien til kvar prinsipalkomponent syner kor stor del av variansen i datasettet som vert forklart av denne komponenten. Laga med PLS_Toolbox i Matlab.



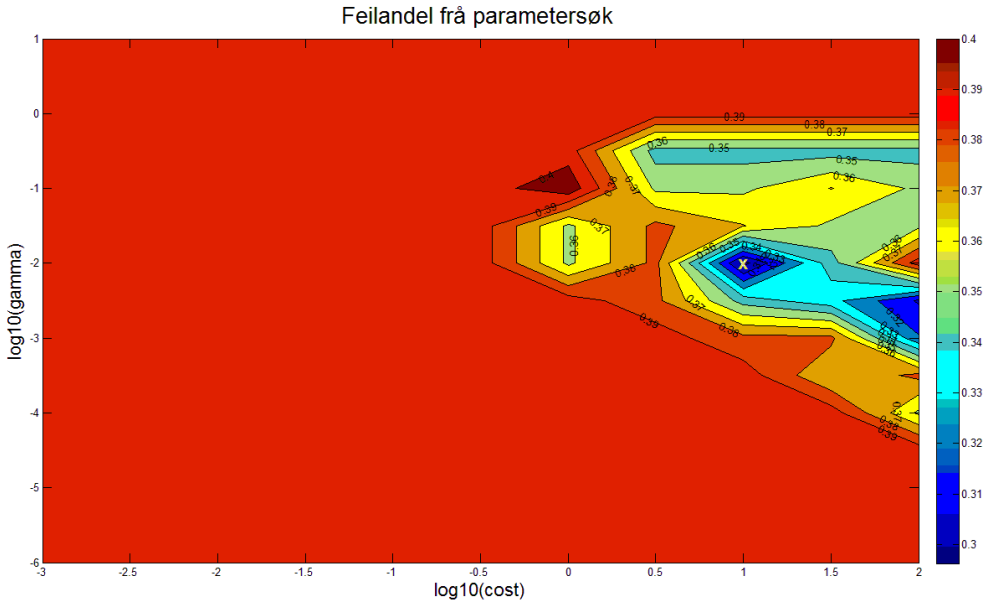
Figur 50: Q mot T^2 for PCA-modellen for pfs = 0, det vil seie for pasientane som vert friske att. Dei rosa punkta representerer pasientane med pfs = 0, det vil seie pasientane som denne modellen skal representere. Dei grøne punkta syner dei andre pasientane, som har pfs = 1. Dei stipla linjene syner grenseverdiane for Q og T^2 ved signifikansnivå 5%. Laga i Matlab med PLS_Toolbox.



Figur 51: Q mot T^2 for PCA-modellen for pfs = 1, det vil seie for pasientane med tilbakefall. Dei grøne punkta representerer pasientane med pfs = 1, det vil seie pasientane som denne modellen skal representere. Dei rosa punkta syner dei andre pasientane, som har pfs = 0. Dei stipla linjene syner grenseverdiane for Q og T^2 ved signifikansnivå 5%. Laga i Matlab med PLS_Toolbox.

4.8 SVM

I parametersøket for SVM nyttar ein 15 verdiar frå 10^{-6} til 10 (logaritmisk skala) for γ, og 11 verdiar frå 10^{-3} til 100 (også logaritmisk) for kostparameteren C. I søket etter det beste (C,γ)-paret nyttast full kryssvalidering.



Figur 52: Resultat av parameteroptimalisering for SVM. log(C) langs x-aksen, log(γ) langs y-aksen. Den beste parameterkombinasjonen, C = 10 og γ = 0,01, er markert med ein X. Raud farge betyr stor feilandel, medan blått indikerer liten feilandel. Laga med PLS_Toolbox i Matlab.

Resultatet frå parametersøket er synt i *figur 52*. Fargane indikerer feilklassifiseringsandel, der mørk raud syner høgast verdi og mørk blå syner lågast verdi. Punktet med lågast feilandel er markert med X, og gjev oss parameterane $C = 10$ og $\gamma = 0{,}01$.

Forvirringsmatrisa for SVM-modellen er gitt i *tabell 27*. Denne metoden har ei nøyaktigheit på 92,5%, og er såleis den modellen som separerer pasientane best. I motsetnad til fleire av dei andre klassifiseringane, finn denne modellen dei fleste pasientane som får tilbakefall, og plasserer desse i riktig gruppe.

Tabell 27: Kalibreringsresultat av SVM. Gruppe 0 er dei pasientane som vert friske, medan gruppe 1 er dei som får tilbakefall.

Nøyaktigheit: 93% Sensitivitet: 96% Spesifisitet: 88%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	47	4
	1	2	28

Modellen har 64 støttevektorar, det vil seie at 79% av objekta ligg nærme, eller på feil side av, grensa mellom klassane. Dette kan vere eit varsel om at modellen er veldig avhengig av desse 81 pasientane, og kanskje vil sjå heilt annleis ut dersom vi nytta målingar frå andre pasientar.

For å undersøkje kor avhengig modellen er av akkurat dette datasettet, gjennomførast ei full kryssvalidering. Resultatet er synt i *tabell 28*. Nøyaktigheita er 88%, og det er berre fire fleire feilklassifiserte pasientar enn i kalibreringa. Dette tyder på at modellen er robust, trass i det store talet på støttevektorar.

Tabell 28: Resultat av full kryssvalidering av SVM. Gruppe 0 er pasientane som vert friske, medan gruppe 1 er pasientane som får tilbakefall.

Nøyaktigheit: 88% Sensitivitet: 96% Spesifisitet: 75%		**Faktisk verdi**	
		0	**1**
Predikert verdi	**0**	47	8
	1	2	24

Matlab-skripta nytta til SVM, både kalibrering og validering, ligg som vedlegg til oppgåva.

5 Diskusjon

5.1 Formål

Målet med analysane i denne oppgåva har vore å undersøkje om parameterar frå ei tilpassing av DCE-MRI-målingar av livmorhalskreftsvulstar til Brix-modellen kan knyttast til utfall av stråleterapi. Dette har tidlegare blitt undersøkt ved hjelp av univariate metodar som logistisk regresjon, *[10]*. Univariate metodar undersøkjer kvar variabel for seg, og tek såleis ikkje omsyn til samspel mellom variablar. Multivariate metodar analyserer med fleire variablar på ein gong, og kan difor finne samanhengar som ikkje kjem fram ved univariat analyse. I denne oppgåva er målet difor å undersøkje om multivariate metodar kan gi ny innsikt i samanhengen mellom Brix-parametererar frå DCE-MRI-målingar og utfall av stråleterapi.

5.2 Vurdering av metodane

Prinsipalkomponentanalyse nyttast for å utforske variablane alder, stadie, svulstvolum og deskriptiv statistikk av Brix-parameterane A, k_{ep} og k_{el}. Analysen syner at ein stor del av variansen i desse variablane kan forklarast ved hjelp av få prinsipalkomponentar. Åtte prinsipalkomponentar kan forklare 90% av variansen i dei 64 variablane, jamfør *tabell 10*. Skårplotta syner at prinsipalkomponentane i svært liten grad deler pasientane inn i grupper etter utfall. Den beste separasjonen fann vi i det tredimensjonale skårplottet for komponentane PC-5, PC-8 og PC-10, synt i *figur 35*, og desse skårane nyttast som forklaringsvariablar i nokre av dei andre analysane.

PLS med progresjonsfri overleving, svulstvolum og stadie som responsvariablar, gav låg forklaringsprosent, kring 50% - 60% forklart varians i kalibreringa, og residualvarians på over 100% etter full kryssvalidering. Analysane vart utførte både med dei statististiske parameterane og med histogramverdiane, med om lag same resultat. Vanleg lineær regresjon nyttast på utvalde prinsipalkomponentar (PC-5, PC-8 og PC-10). Her får dei fleste svulstane predikert utfall på rundt 0,5, det vil seie at dei fleste hamnar om lag midt mellom dei to mogelege utfalla, som er *pfs* = 0 og *pfs* = 1. Desse regresjonsmetodane ser med andre ord ikkje ut til å vere ein god måte å klassifisere svulstane på.

PLS vart førsøkt både med deskriptiv statistikk som gjennomsnitt, standardavvik og persentilverdiar, og med histogramverdiar. Begge variantane gav låg forklaringsprosent, kring 50% - 60%, og presterte dårleg under full kryssvalidering. Det ser difor ikkje ut som histogramframstilling forklarer verken meir eller mindre enn parameterane frå deskriptiv statistikk.

K-means- og K-medians-klyngeanalyse nyttast for å dele svulstane inn i to grupper. Verken analysen av prinsipalkomponentar (PC-5, PC-8 og PC-10) eller analysen basert på alder, stadie, svulstvolum og deskriptiv statistikk gav grupper som samsvarte med behandlingsutfall. Det vil seie at gruppene som svulstane naturleg deler seg inn i i desse ikkje-overvaka analysane, ikkje syner utfall.

Diskriminant analyse i form av LDA og QDA søkjer å dele pasientane inn i grupper som tilsvarar behandlingsutfall. QDA med alder, stadie, volum og dei tre første prinsipalkomponentane frå PCA-modellen som forklaringsvariablar gav i kalibrering ei nøyaktigheit på 84%, sensitivitet 80% og spesifisitet 91%, det vil seie at den klassifiserer pasientane med tilbakefall betre enn dei som vart friske. LDA- og QDA-analysane med tre utvalde prinsipalkomponentar, PC-5, PC-8 og PC-10, som forklaringsvariablar gav i nokre tilfelle (QDA der sannsynet for kvar klasse er berekna ut i frå data) betre sensitivitet, men langt dårlegare spesifisitet. Det vil seie at informasjonen som ligg i alder, stadie, volum og dei andre prinsipalkomponentane gjer det enklare å klassifisere tilbakefallspasientane, men innfører meir forvirring kring dei som vert friske.

LDA er også utført etter variabelseleksjon. Vi har då valt ut dei variablane som skal til for å forklare ein gitt andel (99%, 95% og 90%) av variansen i datasettet bestående av alder, stadie, volum, lymfeknuteinfiltrasjon og deskriptiv statistikk for Brix-parameterane. Ei oversikt over andel forklart varians for dei viktigaste variablane er gitt i *tabell 24*. I desse analysane har vi også nytta full kryssvalidering.Ved å nytte variablane som forklarer 99% av variansen, fekk vi nøyaktigheit 77%, sensitivitet 76% og spesifisitet 78% i kalibreringa, og nøyaktigheit 62%, sensitivitet 63% og spesifisitet 59% etter validering. Ved å redusere talet på variablar til dei som forklarer 90% av variansen, gav analysen 77% nøyaktigheit, 78% sensitivitet og 75% spesifisitet i kalibrering og 67% nøyaktigheit, 67% sensitivitet og 65% spesifisitet i validering. P-verdiar berekna ved å samanlikne desse resultata med resultat frå tilfeldig gruppeinndeling, gav p-verdi på 0,316 for 99%-analysen og p-verdi på 0,011 for 90%-analysen. Det vil seie at LDA med variablane som trengst for å forklare 90% av variansen i datasettet skil signifikant mellom dei to behandlingsutfalla om ein vel eit signifikansnivå på 5%. Ved å analysere med fleire variablar, innfører ein støy som gjer klassifiseringa vanskelegare.

Kalibrering med SIMCA gav 91% nøyaktigheit, 100% sensitivitet og 78% spesifisietet. Det vil seie at alle pasientane som vert friske plasserast i riktig gruppe, men at sju av dei 32 som får tilbakefall vert klassifiserte som friske. Q mot T^2-plotta, gitt i *figur 50* og *51*, av dei to PCA-modellane som vert laga, èin for kvar klasse, syner at så mange som 15 av pasientane med *pfs* = 1 (tilbakefall) ligg innanfor signifikansgrensene til modellen for *pfs* = 0 (friske) ved 5% signifikansnivå. I modellen for *pfs* = 1 er det derimot berre to pasientar med *pfs* = 0 som ligg innanfor grensene. Det vil seie at ein del av av tilbakefallssvulstane liknar på svulstane som vert kurerte, medan få av svulstane som vert kurerte liknar på svulstane som gjev tilbakefall. Det vil seie at om ein pasient vert klassifisert som tilbakefallspasient, er det stort sannsyn for av vedkommane faktisk får tilbakefall. Dersom ein derimot vert klassifisert blant dei som blir friske, er det ikkje sikkert at ein unngår tilbakefall. Ut i frå dette, kan ein argumentere for at pasientar som vert klassifiserte til tilbakefall, bør få meir intensiv behandling, for å betre prognosane deira.

Støttevektormaskin (SVM) med RBF-kernel der vi nytta alder, stadie, volum, lymfeknuteinfiltrasjon og deskriptiv statistikk for Brix-parameterane som forklaringsvariablar, gav 93% nøyaktigheit, 96% sensitivitet og 88% spesifisitet. Etter full kryssvalidering var nøyaktigheita 88%, sensitiviteten 96% og spesifisiteten 75%. Dette er den beste klassifiseringa vi har fått i validering. Trass i at modellen har ein stor andel støttevektorar, 79% i kalibrering, klassifiserer den nesten like godt i validering som i kalibrering.

Det kjem tydeleg fram av analysane i denne oppgåva at overvaka metodar, det vil seie metodar der ein nyttar behandlingsutfallet som rettleiing for å dele svulstane inn i grupper, gjev riktigare klassifisering enn ikkje-overvaka metodar. Dette kan komme av at Brix-modellen, som DCE-MRI-målingane vert tilpassa til, ikkje er laga for å predikere utfall, men for å skildre blodgjennomstrøyminga i karnettverket i vevet, *[8]*.

SIMCA og SVM skil i stor grad pasientane med ulikt behandlingsutfall. Dette indikerer at det er samanheng mellom progresjonsfri overleving og variablane alder, stadie, svulstvolum og deskriptiv statistikk, og at denne er ikkje-lineær. LDA klassifiserer derimot signifikant dersom vi kraftig reduserer talet på variablar, noko som syner at datasettet inneheld mykje støy som ikkje gjev informasjon om utfall. Ved å redusere talet på variablar, kan ein såleis avdekkje ein langt enklare samanheng mellom Brix-parameterane og behandlingsutfallet.

5.3 Identifikasjon av viktige variablar

Variabelseleksjon syner at eit fåtal av variablane forklarer ein stor del av den samla variansen. Dei viktigast variablane er 90%-persentilen til A (37,5%), 30%-persentilen til k_{ep} (17,7%), gjennomsnittet av k_{el} over svulsten (13,5%) og 30%-persentilen til A (6,9%). Erlend Andersen fann i si masteroppgåve, *[10]*, at k_{el} var den av dei tre parameterane som best skil mellom dei to utfalla, og at det var persentilane frå 34% til 67% som gav signifikant skilje. Dette samsvarar med våre funn, der gjennomsnittet av k_{el}, altså 50%-persentilen, identifiserast som ein viktig parameter. Han fann også at A-persentilane frå 7% til 56% og frå 83% og 92% kan gi signifikant skilje mellom utfall, men at denne signifikansen fell vekk når ein korrigerer for svulstvolum og stadie. Amplitudane med desse persentilverdiane, 90% og 30%, peikast også ut i våre analysar. Når det gjeld utvaskingsraten k_{ep}, fann han at persentilane 10% til 36% gjev eit svakt signifikant skilje mellom utfalla, medan vi fann at 30%-persentilen til k_{ep} forklarer 17,7% av variansen. Dette betyr at persentila som han fann at var signifikante, peikast ut som dei variablane som forklarar størst del av variansen i våre analysar. Eit anna studium, gjennomført av Loncaster et al., *[12]*, syner at amplituden A i Brix-modellen korrelerer med utfall i form av progresjonsfri overleving. Også våre analysar peikar på A som den parameteren som forklarar mest varians.

I desse analysane har vi nytta statistiske parameterar for A, k_{ep} og k_{el} for kvar svulst. Dette fordi dei statistiske parameterane er enkle å tolke, samt at ein då har like mange tal per svulst, det vil seie at talet på variablar ikkje er avhengig av kor stor svulsten er. Ein fare ved dette, er at ein kan miste informasjon på vegen. Det er mogleg at sjølve parameterane, eventuelt forholdet mellom nabovokslar, inneheld informasjon ikkje kjem fram i gjennomsnitt, standardavvik og liknande.

5.4 Pasientklassifisering

Ved å undersøkje resultata frå klassifiseringane, gitt i *avsnitt 7.2* i vedlegget, ser ein at nokre svulstar går igjen som feilklassifiserte. Ei oversikt over kva svulstar som plasserast i feil klasse er synt i *tabell 29*. Pasientane 2 og 23 feilklassifiserast som friske i fire av dei fem analysane. Det ser ikkje ut til å vere noko som skil seg ut ved desse to. Dei har noko små svulstar (34 277 mm^3 og 24 674 mm^3), men dette er ikkje dei minste svulstane. Begge har stadie II på undersøkingstidspunktet, og begge får tilbakefall i form av metastasar. Alle pasientar som fekk tilbakefall, men vart feilklassifiserte i SVM og SIMCA, har tilbakefall i form av metastasar, så nær som pasient 24. Dei feilklassifiserte i LDA er meir jamt fordelt mellom lokalt tilbakefall og metastasar. Denne oppgåva har ikkje undersøkt dei to ulike formene for tilbakefall, lokalt og metastasar, men berre sett på skilnaden mellom pasientar som vert friske og tilbakefallspasientar. Eventuelle vidare analysar bør undersøke om det er skilnad på å predikere utfall for pasientar med lokalt tilbakefall og pasientar med metastasar.

Tabell 29: Liste over feilpredikerte pasientar i modellane SVM, SIMCA, LDA med variablar som forklarer 99% av variansen, 95% av variansen og 90% av variansen.

Modell	**Faktisk *pfs* = 1**	**Faktisk *pfs* = 0**
SVM	2, 6, 23, 59	28,42
SIMCA	2, 6, 24, 26, 27, 46, 69	
LDA (99%)	9, 23, 37, 39, 59, 64, 71	16, 17, 28, 32, 36, 42, 45, 54, 57, 58, 74, 79
LDA (95%)	2, 9, 19, 23, 27, 39, 59, 64, 71	16, 28, 32, 36, 42, 45, 57, 58, 74, 76, 79
LDA (90%)	2, 5, 8, 9, 23, 39, 64, 71	11, 12, 16, 28, 32, 38, 42, 45, 57, 58, 76

Pasientane 28 og 42 er pasientar som vert friske, men som i fire av dei fem modellane predikerast som tilbakefallspasientar. Desse pasientane har begge det høgaste stadiet, IV, ved DCE-MRI-undersøkinga. Dei kjem til undersøkinga med alvorleg kreftsjukdom, men vert likevel friske, noko som synast å vere vanskeleg for modellane å fange opp, sjølv om SIMCA lukkast i å identifisere også desse.

Det er mogleg at anna informasjon om pasientane kan forklare kvifor nokre pasientar er ekstra vanskelege å klassifisere. Oppfølgingsdata og sjukdomshistoria til desse pasientane kan undersøkjast for å sjå om ein finn likskapar mellom pasientane, som ikkje kjem fram av Brix-parameterane eller storleikane alder, stadie og svulstvolum.

5.5 Ulemper ved farmakokinetiske modellar

Farmakokinetiske modellar søkjer å redusere datamengda til nokre, typisk to eller tre, parameterar for kvar voksel. Dei tilpassar rådata, som er bilete som funksjon av tid, til ein matematisk modell. Når ein tilpassar DCE-MRI-målingane til farmakokinetiske modellar, er det ein risiko for å gjere føresetnader som ikkje stemmer. Andersen et al., *[14]*, påpeikar til dømes at DCE-MRI-bilete ikkje nødvedigvis har eit lineært forhold mellom signalauke og kontrastmiddelkonsentrasjon, slik som Brix-modellen føreset.

Eit mogleg problem med tilpassinga til Brix-modellen i denne oppgåva, er dei mange null-estimata av utvaskingsraten k_{el}. Alle svulstane har minst èin voksel der $k_{el} = 0$, noko som antydar at det ikkje vert vaska ut kontrastmiddel frå vokselen i det heile. Inga utvasking vil vere det same som å seie at det ikkje er blodgjennomstrøyming i vevet. Dette skuldast at tida undersøkinga tek er for kort til å estimere utvaskingsraten, sidan vokselen ikkje når utvaskingsfasen, eller har for få målingar i løpet av denne, *[10]*, jamfør *figur 11*. Fleire måletidspunkt i DCE-MRI-serien, eventuelt lenger tid mellom kvar måling, kunne ha gitt riktigare estimat av k_{el}.

Bruwer, MacGregor og Noseworthy, *[24]*, antydar at PCA-analyse av rådata, det vil seie den relative signalauken RSI, er betre eigna til å predikere behandlingsutfall enn analysar av parameterar frå farmakokinetiske modellar. Dei åtvarar om at dei farmakokinetiske modellane kan gjere for mange forenklingar, slik at ein mistar viktig informasjon. Sidan ein i prinsipalkomponentanalyse leitar etter latente samanhengar i datasettet i staden for å tilpasse data til parameterar ein har valt på førehand, kan ein her avdekke skilnader som ikkje inngår i modellen. Denne oppgåva nyttar parameterar frå Brix-modellen. Det er mogleg at rådata seier meir om behandlingsutfall enn det Brix-parameterane gjer. For å undersøkje dette, bør det gjennomførast analysar med rådata i tillegg til det som er gjort her med Brix-parameterane. Det er også mogleg at andre farmakokinetiske modellar, til dømes RR-modellen, *[9]*, kan vere betre eigna til å predikere utfall enn det Brix-modellen er.

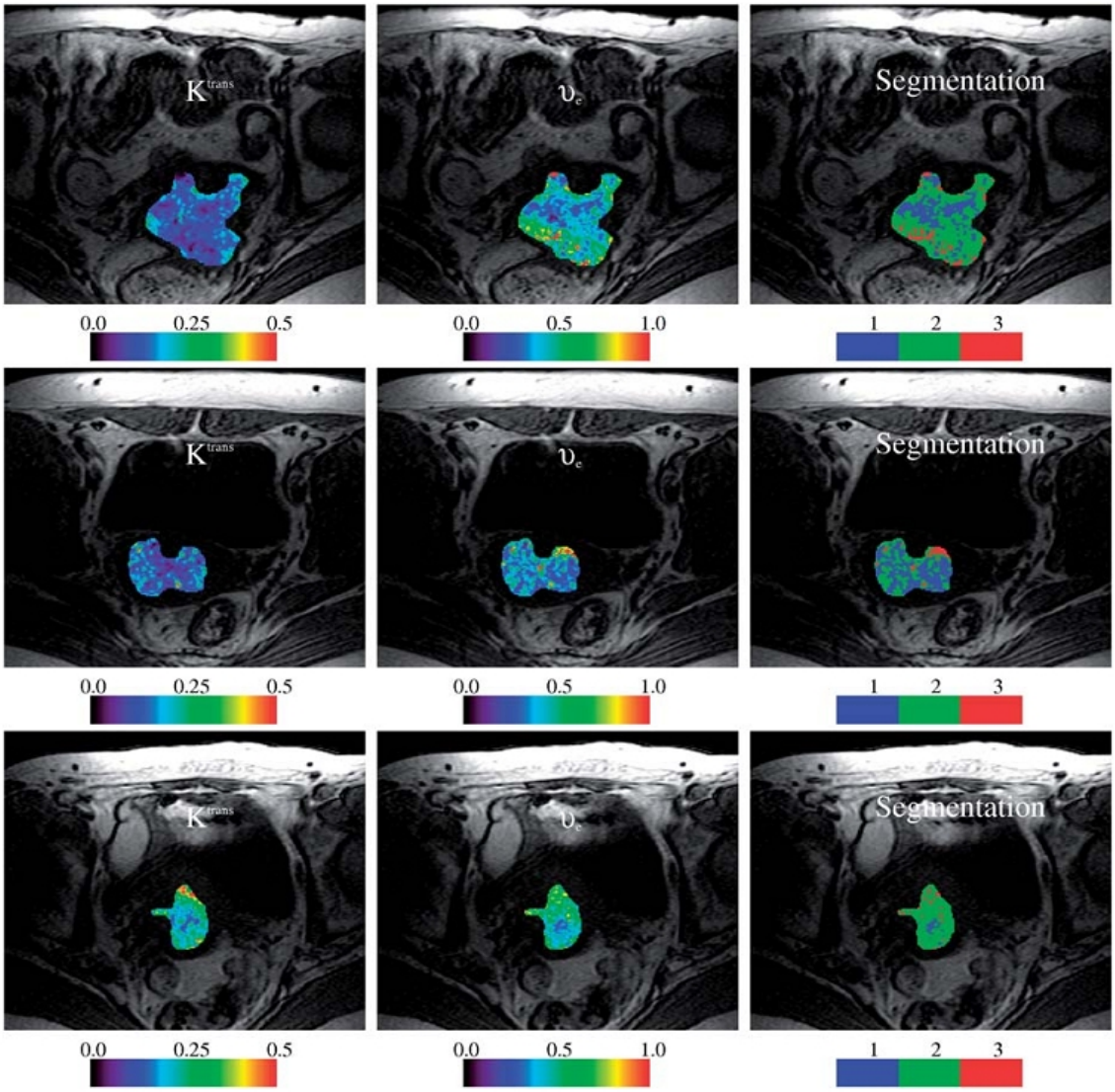
5.6 Rommlege analysar av svulstar

I analysane i denne oppgåva har vi ikkje undersøkt eventuelle rommlege samanhengar mellom vokslane, men berre sett på heile svulsten under eitt. Teksturanalyse, der ein ser på korleis vokslane varierer med nabovokslar, vil gi nye variablar som skildrar korleis vokslane forheld seg til kvarandre, *[43]*. Desse variablane kan nyttast som forklaringsvariablar i til dømes diskriminant analyse.

Multivariat biletanalyse, MIA, kan potensielt nyttast til å identifisere område i svulsten som responderer dårlegare på stråleterapi enn andre. Her nyttast PCA på eit sett treningsbilete til å rekne ut normale T^2-verdiar, *[43]*. Nye bilete projiserast inn i denne PCA-modellen, og T^2-verdiane vert berekna. Ut i frå dette kan ein identifisere vokslar som har unormalt høge T^2-verdiar, det vil seie vokslar som ikkje passar inn i den opprinnelege modellen. Dersom treningsbiletet som nyttast er eit bilete av ein svulst som vert kurerte, kan område med unormalt høge T^2-verdiar vere kritiske område av svulsten som ikkje responderer like godt på stråleterapi. Zahra et al., *[4]*, foreslår at DCE-MRI-målingar kan nyttast til å identifisere kristiske område av svulstar og tilpasse behandliga etter dette, til dømes ved å gi høgare stråledose til desse områda.

Fleire studiar, til dømes Cooper et al., *[11]*, og Loncaster et al., *[12]*, syner korrelasjon mellom DCE-MRI-målingar og oksygenmengde i vevet. Hypoksi, det vil seie lite oksygen i vevet, minkar effekten av stråleterapi, *[13]*. Multivariat biletanalyse kan nyttast til å identifisere dei hypoksiske områda i svulsten, slik at ein eventuelt kan gi desse meir intensiv behandling.

Andersen et al., *[14]*, nyttar K-means-klynger til å dele svulsten inn i område, og undersøkjer om desse indikerer utfall i form av lokalt tilbakefall. Dei nytta parameterane K^{trans} og v_e, sjå *kapittel 2.4*, som grunnlag for analysen, og delte vokslane i kvar svulst inn i tre klynger. Fordelinga av klyngene over tre ulike svulstar er gitt i *figur 53*, saman med kart over parameterane K^{trans} og v_e i dei same svulstane. Ved å dele svusltane inn i to grupper for kvar klynge, basert på om vokslane i klynga utgjorde ein stor eller liten del av det totale svulstvolumet, fann dei at for ei av klyngene, klynga med sentrum $K^{trans} = 0{,}20$ og $v_e = 0{,}45$, var det eit sigifikant skilje mellom behandlingsutfall for dei to gruppene. Pasientane med ein låg andel av denne klynga fekk hyppigare tilbakefall enn dei andre pasientane. Meir multivariat biletanalyse kan kanskje nyttast til å betre identifisere slike avgjerande område i svulstane.



Figur 53: Figuren syner fordelinga av K^{trans} (venstre), v_e (midt) og tre klynger frå K-means-klyngeanalyse (høgre). Dei to øvste svulstane fekk lokalt tilbakefall, medan den nederste ikkje fekk det. Klynge 2, markert med grønt på bileta til høgre, korrelerer negativt med lokalt tilbakefall. Henta frå Andersen et al., [14].

5.7 Vidare analysar

Dette studiet syner at multivariate teknikkar gjev moglegheit for vidare analysar av datasettet.

SIMCA-modellen er ikkje validert, så det bør gjennomførast ei validering, helst i form av full kryssvalidering slik som for SVM.

LDA med variabelseleksjon gav sigifikante resultat, men noko låg nøyaktigheit. QDA gav betre nøyaktigheit enn LDA. Ved å nytte QDA saman med variabelseleksjon, kan ein undersøkje om dette gjev signifikant klassifisering med større nøyaktigheit enn LDA.

Som antyda av Bruwer, MacGregor og Noseworthy, *[24]*, kan rådata, det vil seie relativ signalauke RSI som funksjon av tid, analyserast i staden for parameterar frå farmakokinetiske modellar.

Romlege samanhengar kan undersøkjast i form av teksturanalyse, og nyttast i klassifisering. I tillegg kan multivariat biletanalyse nyttast til å identifisere problemområde, til dømes hypoksiske område, i svulsten.

6 Konklusjon

Analysane i denne oppgåva indikerer at det er ein samanheng mellom parameterar frå ei tilpassing av DCE-MRI-bilete til Brix-modellen, og utfall av stråleterapi. Analysane syner at denne samanhengen ikkje er lineær.

Prinsipalkomponentanalyse av alder, stadie, svulstvolum og deskriptive statistiske parameterar for A, k_{ep} og k_{el}, syner at desse kan reduserast til få prinsipalkomponentar. Det trengst berre åtte komponentar for å skildre 90% av variansen, men det er ikkje mogleg å skilje klart mellom pasientar som vert friske og pasientar som får tilbakefall.

Verken lineær regresjon med prinsipalkomponentar som forklaringsvariablar, PLS med histogramverdiar eller PLS med dei deskriptive parameterane, lukkast i å skilje dei to gruppene av pasientar, dei med tilbakefall og dei som vart friske, frå kvarandre. PLS-modellar med svulstvolum eller stadie som respons gav heller ikkje betre resultat. Forklart varians er 50%-60% i kalibreringane, medan det er over 100% residualvarians etter full kryssvalidering.

Ikkje-overvaka klassifisering i form av K-means og K-medians-klyngeanalyse deler pasientane inn i to grupper, men desse gruppene samsvarar ikkje med behandlingsutfall.

LDA og QDA med utvalde komponantar frå PCA-modellen som forklaringsvariablar og progresjonsfri overleving som respons, gjev i det beste tilfellet (QDA der sannsynet for kvar klasse er kalkulert ut i frå data) nøyaktigheit på 70%, sensitivitet på 92% og spesifisitet på 38% i kalibrering. Ved å nytte fleire prinsipalkomponentar aukar nøyaktigheita til 84% og spesifisiteten til 91%, medan sensitiviteten går noko nedover til 80%.

Ein variabelseleksjon som vel ut dei variablane som forklarer mest av totalvariansen i datasettet, syner at vi berre treng 11 av dei 67 variablane for å forklare over 90% av variansen. Det er fire variablar som kvar for seg forklarer over 5% av variansen: 90%-persentilen til A, 30%-persentilen til k_{ep}, gjennomsnittsverdien til k_{el} og 30%-persentilen til A. Til saman dekkjer desse fire variablane nesten 76% av totalvariansen.

LDA med full kryssvalidering utført etter variabelseleksjon, syner at begge klassane, både pasientane som vert friske og dei som får tilbakefall, vert klassifiserte signifikant, med ein p-verdien 0,011 for begge klassar. Denne analysen hadde nøyaktigheit 77%, sensitivitet 78% og spesifisitet 75% i kalibrering og nøyaktigheit 67%, sensitivitet 67% og spesifisitet 65% etter full kryssvalidering.

SIMCA i kalibrering gjev nøyaktigheit 91%, sensitivitet 100% og spesifisitet 78%. Denne analysen klassifiserer med andre ord alle pasientane som vert friske riktig, men er noko svakare på dei som får tilbakefall.

Støttevektormaskiner (SVM) i kalibrering gav nøyaktigheit 93%, sensitivitet 96% og spesifisitet 88%, medan full kryssvalidering synte nøyaktigheit på 88%, sensitivitet på 96% og spesifisitet på 75%. Modellen har ein stor andel støttevektorar, 79%, men valideringa indikerer at den likevel ikkje er for avhengig av dette spesifikke datasettet.

Resultata frå analysene syner at det er enklare å identifisere pasientane som vert friske enn dei som får tilbakefall. Det er også nokre pasientar som går igjen som feilklassifiserte i fleire av analysane.

Multivariate metodar nytta på dette datasettet gjev kvantitative mål, som nøyaktigheit, sensitivitet og spesifisitet, på kor gode klassifiseringane er. Vi er også i stand til å identifisere kva svulstar som er vanskelege å predikere utfall for. Ein annan fordel i høve til univariate metodar, er at ein her tek omsyn til samspel mellom variablar, og ikkje treng korrigere for dette i etterkant. Multivariate metodar er såleis eit viktig bidrag til analysen av DCE-MRI-bilete.

Som oppfølging til desse analysane, kan pasientane som feilklassifiserast samanliknast for å finne ut om dei har noko til felles som ikkje kjem fram av alder, stadie, svulstvolum og DCE-MRI-målingar. Analysar som ikkje nyttar deskriptive statistiske parameterar, som til dømes teskturanalyse eller multivariat biletanalyse, bør forsøkjast i tillegg til analysane som er utførte i denne oppgåva. Det bør også vurderast om ein skal analysere rådata frå DCE-MRI-undersøkinga i staden for parameterar frå tilpassingar til farmakokinetiske modellar.

Kjelder

[1] Oncolex. Oslo Universitetssykehus HF. http://oncolex.no/GYN/Diagnoser/Livmorhals/Bakgrunn/Stadier.aspx.

[2] Mark H. Shiffman, Heidi M. Bauer, Robert N. Hoover, Andrew G. Glass, Diane M. Cadell, Brenda B. Rush, David R. Scott, Mark E. Sherman, Robert J. Kurman, Sholom Wacholder, Cynthia K. Stanton og M. Michele Manos (1993), Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia.. *Journal of the National Cancer Institute*, 85:958-964.

[3] Stewart C. Bushong (1996), *Magnetic Resonance Imaging: Physical and Biological Principles*. 2. utg. St. Louis, Missouri, USA: Mosby-Year Book.

[4] Mark A. Zahra, Kieren G. Hollingsworth, Evis Sala, David J. Lomas og Li T. Tan (2007), Dynamic contrast-enhanced MRI as a predictor of tumour response to radiotherapy. *Lancet Oncol*, 8:63-74.

[5] Peter Carmeliet og Rakesh K. Jain (2000), Angiogenesis in cancer and other diseases. *Nature*, 407:249-257.

[6] Olav Spigset (2010), G2 Farmakokinetikk og doseringsprinsipper, *Norsk legemiddelhåndbok*. Foreningen for utgivelse av Norsk Legemiddelhåndbok.

[7] Anwar R. Padhani (2002), Dynamic contrast-enhanced MRI in clinical oncology: Current status and future directions. *Journal of Magnetic Resonance Imaging*, 16:407-422.

[8] Gunnar Brix, Wolfhard Semmler, Rüdiger Port, Lothar R. Schad, Günther Layer og Walter J. Lorenz (1991), Pharmacokinetic Parameters in CNS Gd-DTPA Enhanced MR Imaging. *Computer Assisted Tomography*, 15:621-628.

[9] Thomas E. Yankeelov, Jeffrey J. Luci, Martin Lepage, Rui Li, Laura Debusk, P. Charles Lin, Ronald R. Price og John C. Gore (2005), Quantitative pharmacokonetic analysis of DCE-MRI data without an arterial input function: a reference region model. *Magnetic Resonance Imaging*, 23:519-529.

[10] Erlend Kristoffer Frivold Andersen (2009), *Dynamisk kontrastforsterket MRI av pasienter med livmorhalskreft. Korrelasjonsanalyse av bildeparametre mot langtidsoverlevelse etter stråleterapi*. Masteroppgåve. Oslo: Universitetet i Oslo, Fysisk Institutt.

[11] Rachel A. Cooper, Bernadette M. Carrington, Juliette A. Loncaster, Susan M. Todd, Susan E. Davidson, John P. Logue, Asha D. Luthra, Andrew P. Jones, Ian Stratford, Robert D. Hunter og Catharine M. L. West (2000), Tumour oxygenation levels correlate with dynamic contrast-enhanced magnetic resonance imaging parameters in carcinoma of the cervix. *Radiotherapy and Oncology*, 57:53-59.

[12] Juliette A. Loncaster, Bernadette M. Carrington, Johnathan R. Sykes, Andrew P. Jones, Susan M. Todd, Rachel Cooper, David L. Buckley, Susan E. Davidson, John P. Logue, Robin D. Hunter og Catharine M. L. West (2002), Prediction of radiotherapy outcome using dynamic contrast enhanced mri of carcinoma of the cervix. *Int. J. Radiation Oncology Biol. Phys.*, 54:759-767.

[13] Eric J. Hall og Amato J. Giaccia (2006), *Radiobiology for the radiologist*. 6. utg. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins.

[14] Erlend K. F. Andersen, Gunnar B. Kristensen, Heidi Lyng og Eirik Malinen (2011), Pharmacokinetic analysis and k-means clustering of DCEMR images for radiotherapy outcome prediction of advanced cervical cancers. *Acta Oncologica*, 50:859-865.

[15] Erlend K.F. Andersen, Knut Håkon Hole, Kjersti V. Lund, Kolbein Sundfør, Gunnar B. Kristensen, Heidi Lyng og Eirik Malinen. (2011), Dynamic Contrast-Enhanced MRI of Cervical Cancers: Temporal Percentile Screening of Contrast Enhancement Identifies Parameters for Prediction of Chemoradioresistance. *International Journal of Radiation Oncology Biology Physics*.

[16] Richard A. Johnson og Dean W. Wichern (2002), *Applied Multivariate Statistical Analysis*. 5. utg. Upper Saddle River, New Jersey: Pearson Education.

[17] Joseph P. Hornak (1996-2010), *The Basics of MRI*. Interactive Learning Software. http://www.cis.rit.edu/htbooks/mri/index.html.

[18] John Lilley (2001), *Nuclear Physis: Principles and Applications*. Chichester, England: John Wiley & Sons.

[19] David J. Griffiths (2005), *Introduction to Quantum Mechanics*. 2. utg. Upper Saddle River, New Jersey, USA: Pearson Education.

[20] Mark A. Brown og Rickhard C. Semelka (2010), *MRI: Basic principles and applications*. 4. utg. Hoboken, New Jersey, USA: Wiley-Blackwell.

[21] Ray H. Hashemi og William G. Bradley jr. (1997), *MRI: The Basics*. Baltimore, Maryland, USA: Williams & Wilkins.

[22] Paul A. Tipler og Gene Mosca (2008), *Physics for Scientists and Engineers*. 6. utg. New York, USA: W.H. Freeman and Company.

[23] Atle Bjørnerud (2008), *FYS-KJM-4740: The Physics of Magnetic Resonance Imaging*. Kompendium. Oslo: Universitetet i Oslo, Fysisk Institutt.

[24] Mark-John Bruwer, John F. MacGregor og Michael D. Noseworthy (2008), Dynamic contrast-enhanced MRI diagnostics in oncology via principal component analysis. *Journal of Chemometrics*, 22:708-716.

[25] Cheng Yang, Walter M. Stadler, Gregory S. Karczmar, Michael Milosevic, Ivan Yeung og Masoon A. Haider (2010), Comparison of Quantitative Parameters in Cervix Cancer Measured by Dynamic Contrast– Enhanced MRI and CT. *Magnetic Resonance in Medicine*, 63:1601-1609.

[26] Anthea Maton (1997), *Human Biology and Health*. 3. utg. Englewood Cliffs, New Jersey: Pearson Prentice Hall.

[27] Paul S. Tofts, Gunnar Brix, David L. Buckley, Jeffrey L. Evelhoch, Elizabeth Henderson, Michael V. Knopp, Henrik B. W. Larsson, Ting-Yim Lee, Nina A. Mayr, Geoffrey J. M. Parker, Ruedinger E. Port, June Taylor og Robert M. Weisskoff (1999), Estimating Kinetic Parameters From Dynamic Contrast-Enhanced T1-Weighted MRI of a Diffusable Tracer: Standardized Quantities and Symbols. *Journal of Magnetic Resonance Imaging*, 10:223-232.

[28] Personleg kommunikasjon (2011). Professor II Eirik Malinen, Fysisk institutt, Universitetet i Oslo, Oslo, Noreg.

[29] Matlab. http://www.mathworks.se/.

[30] Gunnar G. Løvås (2004), *Statistikk for universitet og høgskoler.* 2. utg. Oslo, Noreg: Universitetsforlaget.

[31] David W. Scott (1979), On Optimal and Data-Based Histograms. *Biometrika*, 66:605-610.

[32] David Freedman og Persi Diaconis (1981), On the Histogram as a Density Estimator: L2 Theory . *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:453-476.

[33] Kim Esbensen (2009), *Multivariate Data Analysis in Practice*. 5. utg. Oslo, Noreg: Camo Software.

[34] David L. Olson og Dursun Delen (2008), *Advanced Data Mining Techniques*. Berlin, Heidelberg, Tyskland: Springer-Verlang.

[35] Barry M. Wise, Neal B. Gallagher, Rasmus Bro, Jeremy M. Shaver, Willem Windig og R. Scott Koch (2006), *Chemometrics Tutorial for PLS_Toolbox and Solo*. Wenatchee, Washington, USA: Eigenvector Research.

[36] Ovidiu Ivanciuc (2007), Applications of Support Vector Machines in Chemistry. *Reviews in Computational Chemistry*, 23:291-400.

[37] Chih-Wei Hsu, Chih-Chung Chang og Chih-Jen Lin (2010), *A Practical Guide to Support Vector Classification*. Taiwan: National Taiwan University, Department of Computer Science.

[38] Eigenvector Documentation Wiki. http://wiki.eigenvector.com/index.php?title=Main_Page.

[39] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (2006), *Introduction to Linear Regression Analysis*. 4. utg. Hoboken, New Jersey, USA: Wiley.

[40] Michael Haenlein og Andreas M. Kaplan (2004), A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics*, 3:283-297.

[41] Personleg kommunikasjon (2011). F.aman. Ulf Geir Indahl, Institutt for matematiske realfag og teknologi, Universitetet for Miljø- og Biovitskap, Ås, Noreg.

[42] The Unscrambler X 10.1. http://www.camo.com/index.html.

[43] Fernando López-García, Gabriela Andreu-García, José Blasco, Nuria Aleixos og José-Miguel Valiente (2010), Automatic detection of skin defects in citrus fruits using a multivariate image analysis approach. *Computers and Electronics in Agriculture*, 71:189-197.

7 Vedlegg

7.1 Matlab-skript

Lese data frå tekstfil og lagre som stuktur

```
% Read information from txt-file and store it in a struct.
% Turid Torheim
% 16.03.11

% txt-file contains patient information:
% #name MM001
% #FIGO 2b
% #birthyear[DDMMYYYY] 27021956
% #progression_free_survival        0
% #locoregional_control        0
% #Lymphnode_infiltration[1=True]        1
% #study_date[DDMYYYY] 27022001
% #Degrees_of_freedom           12
% # X Y Z A kep kel A_stddev kep_stddev kel_stddev chisqr fit_status
% Data for each patient end with #EOP.
% File end with #EOF.

% Read every line from the file into the cell array A.
A = importdata('data_to_txtMM.txt','\t');

% Find all EOP (end-of-patient).
% eop is a vector. The numbers in the vector shows wich lines in the file
% that has EOP.
eop = strmatch('#EOP',A);

% Total number of patients.
num_patients = length(eop);

% Separate the patients.
for i = 1:num_patients
    if i == 1
        patient{i} = {A{1:eop(i)}};
    else
        patient{i} = {A{eop(i-1)+1:eop(i)}};
    end
end

info(num_patients) = struct('mri',[],'summary',[]);

for I = 1:num_patients  % For every patient.

    % Find patient name.
    index = strmatch('#name',patient{I});
    name = sscanf(patient{I}{index},'#name %s');

    % Find figo status.
    index = strmatch('#FIGO',patient{I});
```

```
    figo = sscanf(patient{I}{index},'#FIGO %s');

    % Find birth year.
    index = strmatch('#birthyear[DDMMYYYY]',patient{I});
    birth = sscanf(patient{I}{index},'#birthyear[DDMMYYYY] %d');

    % Find progression_free_survival.
    index = strmatch('#progression_free_survival',patient{I});
    pfs = sscanf(patient{I}{index},'progression_free_survival %d');

    % Find locoregional_control.
    index = strmatch('#locoregional_control',patient{I});
    lc = sscanf(patient{I}{index},'#locoregional_control %d');

    % Find lymphnode_infiltration.
    index = strmatch('#Lymphnode_infiltration[1=True]',patient{I});
    li = sscanf(patient{I}{index},'#Lymphnode_infiltration[1=True] %d');

    % Find study date.
    index = strmatch('#study_date[DDMYYYY]',patient{I});
    study = sscanf(patient{I}{index},'#study_date[DDMYYYY] %d');

    % Find degrees of freedom for chi-square.
    index = strmatch('#Degrees_of_freedom',patient{I});
    df = sscanf(patient{I}{index},'#Degrees_of_freedom %d');

    % Read mri-data (that is, info about each voxel).
    [m,n] = size(patient{I});
    ind = 1;
    for i = 1:n
        C = textscan(patient{I}{i}, '%f64', 11, 'CommentStyle', '#'); %
Search for 11 number in a row.
        if isempty(C{1})==0 % That is, if we have found numbers,
            for j = 1:11;
                data(ind,j) =  C{1}(j); %Put mri-data into the matrix data.
            end
            ind = ind + 1;
        end
    end

    % Calculate volume of tumor.
    [m n] = size(data);
    volume = m*0.78*0.78*5; % Gives tumor volume in (mm)^3.

    % Calculate age of patient.
    S = num2str(study);
    B = num2str(birth);
    Syear = S(end-3:end); % The year is given by the four last digits.
    Byear = B(end-3:end); % The year is given by the four last digits.
    study_year = str2num(Syear);
    birth_year = str2num(Byear);
    age = study_year - birth_year;

    % Find the center of the tumor.
    center(1) = mean(data(:,1)); % x
    center(2) = mean(data(:,2)); % y
    center(3) = mean(data(:,3)); % z

    % Information about the patient is stored in the struct summary.
    summary =
```

```
struct('name',name,'figo',figo,'birthdate',birth,'studydate',study,'age',ag
e,'volume',volume,'lymphnode_infiltration',li,'locoregional_control',lc,'pr
ogression_free_survival',pfs,'df',df,'center',center);

    % Struct with mri-data and patient information.
    info(I).mri = data;
    info(I).summary = summary;

end

% info(I) contains everything to do with patient I.
% info(I).mri gives the mri-data.
% info(I).summary gives the patient information.

save('info.mat','info');
```

Lese data frå tekst-fil til Unscrambler

```
% Reads information from txt-file, calculates statistical parameters,
% and stores it in a way suitable for exporting to Unscrambler.
% Turid Torheim
% 16.06.11

% The scripts reads patient information and Brix-paramters from the file
'data_to_txtMM.txt'.
% The information is stored in 'mridata.mat'.
% This file contains 7 matrixes:
% 'info','A','kep','kel','A_persentil','kep_persentil' and 'kel_persentil'

% txt-file contains patient information:
% #name MM001
% #FIGO 2b
% #birthyear[DDMMYYYY] 27021956
% #progression_free_survival        0
% #locoregional_control        0
% #Lymphnode_infiltration[1=True]        1
% #study_date[DDMYYYY] 27022001
% #Degrees_of_freedom           12
% # X Y Z A kep kel A_stddev kep_stddev kel_stddev chisqr fit_status
% Data for each patient end with #EOP.
% File end with #EOF.

% Read every line from the file into the cell array A.
A = importdata('data_to_txtMM.txt','\t');

% Find all EOP (end-of-patient).
% eop is a vector. The numbers in the vector shows which lines in the file
% that has EOP.
eop = strmatch('#EOP',A);

% Total number of patients.
num_patients = length(eop);

% Separate the patients.
for i = 1:num_patients
    if i == 1
        patient{i} = {A{1:eop(i)}};
```

```
    else
        patient{i} = {A{eop(i-1)+1:eop(i)}};
    end
end

% Creates matrixes to store the information.
info = nan(num_patients,6);
A = nan(num_patients,8);
kep = nan(num_patients,8);
kel = nan(num_patients,8);
A_persentil = nan(num_patients, 13);
kep_persentil = nan(num_patients, 13);
kel_persentil = nan(num_patients, 13);

for I = 1:num_patients  % For every patient.

    % Find patient name.
    index = strmatch('#name',patient{I});
    name = sscanf(patient{I}{index},'#name %s');

    % Find figo status.
    index = strmatch('#FIGO',patient{I});
    figo = sscanf(patient{I}{index},'#FIGO %d');

    % Find birth year.
    index = strmatch('#birthyear[DDMMYYYY]',patient{I});
    birth = sscanf(patient{I}{index},'#birthyear[DDMMYYYY] %d');

    % Find progression_free_survival.
    index = strmatch('#progression_free_survival',patient{I});
    pfs = sscanf(patient{I}{index},'#progression_free_survival %d');

    % Find locoregional_control.
    index = strmatch('#locoregional_control',patient{I});
    lc = sscanf(patient{I}{index},'#locoregional_control %d');

    % Find lymphnode_infiltration.
    index = strmatch('#Lymphnode_infiltration[1=True]',patient{I});
    li = sscanf(patient{I}{index},'#Lymphnode_infiltration[1=True] %d');

    % Find study date.
    index = strmatch('#study_date[DDMYYYY]',patient{I});
    study = sscanf(patient{I}{index},'#study_date[DDMYYYY] %d');

    % Read mri-data (that is, info about each voxel).
    data = [];
    [m,n] = size(patient{I});
    ind = 1;
    for i = 1:n
        C = textscan(patient{I}{i}, '%f64', 11, 'CommentStyle', '#'); %
Search for 11 number in a row.
        if isempty(C{1})==0 % That is, if we have found numbers,
            for j = 1:11;
                data(ind,j) =  C{1}(j); % Put mri-data into the matrix
data.
            end
            ind = ind + 1;
        end
    end
```

```
    % Calculate volume of tumor.
    [m n] = size(data);
    volume = m*0.78*0.78*5; % Gives tumor volume in (mm)^3.

    % Calculate age of patient.
    S = num2str(study);
    B = num2str(birth);
    Syear = S(end-3:end); % The study year is given by the four last
digits.
    Byear = B(end-3:end); % The birth year is given by the four last
digits.
    study_year = str2num(Syear);
    birth_year = str2num(Byear);
    age = study_year - birth_year;

    % Remove voxels with large values of A, kep and kel, before calulating
    % statistical parameters.
    [m n] = size(data);
    for i = 1:m
        if data(i,4) > 10
            data(i,4) = nan; % Removes voxels with A>10.
        end
        if data(i,5) > 12
            data(i,5) = nan; % Removes voxels with kep>12.
        end
        if data(i,6) > 0.5
            data(i,6) = nan; %Removes voxels with kel>0.5.
        end
    end

    % Statistical parameters for A.
    A(I,1) = nanmean(data(:,4)); % Mean.
    A(I,2) = nanmedian(data(:,4)); %Median.
    A(I,3) = mode(data(:,4)); % Most frequent value.
    A(I,4) = min(data(:,4)); % Minimum.
    A(I,5) = max(data(:,4)); % Maximum.
    A(I,6) = nanstd(data(:,4)); % Standard deviation.
    A(I,7) = skewness(data(:,4)); % Skewness.
    A(I,8) = kurtosis(data(:,4)); % Kurtosis.

    % Statistical parameters for kep.
    kep(I,1) = nanmean(data(:,5)); % Mean.
    kep(I,2) = nanmedian(data(:,5)); %Median.
    kep(I,3) = mode(data(:,5)); % Most frequent value.
    kep(I,4) = min(data(:,5)); % Minimum.
    kep(I,5) = max(data(:,5)); % Maximum.
    kep(I,6) = nanstd(data(:,5)); % Standard deviation.
    kep(I,7) = skewness(data(:,5)); % Skewness.
    kep(I,8) = kurtosis(data(:,5)); % Kurtosis.

    % Statistical parameters for kel.
    kel(I,1) = nanmean(data(:,6)); % Mean.
    kel(I,2) = nanmedian(data(:,6)); %Median.
    kel(I,3) = mode(data(:,6)); % Most frequent value.
    kel(I,4) = min(data(:,6)); % Minimum.
    kel(I,5) = max(data(:,6)); % Maximum.
    kel(I,6) = nanstd(data(:,6)); % Standard deviation.
    kel(I,7) = skewness(data(:,6)); % Skewness.
    kel(I,8) = kurtosis(data(:,6)); % Kurtosis.
```

```
    % Patient information.
    info(I,1) = age;
    info(I,2) = volume;
    info(I,3) = li;
    info(I,4) = lc;
    info(I,5) = pfs;
    info(I,6) = figo;

    % A percentiles.
    A_persentil(I,1) = quantile(data(:,4),.10); % 10%.
    A_persentil(I,2) = quantile(data(:,4),.20); % 20%.
    A_persentil(I,3) = quantile(data(:,4),.25); % 25%.
    A_persentil(I,4) = quantile(data(:,4),.30); % 30%.
    A_persentil(I,5) = quantile(data(:,4),.40); % 40%.
    A_persentil(I,6) = quantile(data(:,4),.50); % 50%.
    A_persentil(I,7) = quantile(data(:,4),.60); % 60%.
    A_persentil(I,8) = quantile(data(:,4),.70); % 70%.
    A_persentil(I,9) = quantile(data(:,4),.75); % 75%.
    A_persentil(I,10) = quantile(data(:,4),.80); % 80%.
    A_persentil(I,11) = quantile(data(:,4),.90); % 90%.
    A_persentil(I,12)= iqr(data(:,4)); % 75% - 25%.
    A_persentil(I,13)= quantile(data(:,4),.90) - quantile(data(:,4),.10); %
90% - 10%.

    % kep percentiles.
    kep_persentil(I,1) = quantile(data(:,5),.10); % 10%.
    kep_persentil(I,2) = quantile(data(:,5),.20); % 20%.
    kep_persentil(I,3) = quantile(data(:,5),.25); % 25%.
    kep_persentil(I,4) = quantile(data(:,5),.30); % 30%.
    kep_persentil(I,5) = quantile(data(:,5),.40); % 40%.
    kep_persentil(I,6) = quantile(data(:,5),.50); % 50%.
    kep_persentil(I,7) = quantile(data(:,5),.60); % 60%.
    kep_persentil(I,8) = quantile(data(:,5),.70); % 70%.
    kep_persentil(I,9) = quantile(data(:,5),.75); % 75%.
    kep_persentil(I,10) = quantile(data(:,5),.80); % 80%.
    kep_persentil(I,11) = quantile(data(:,5),.90); % 90%.
    kep_persentil(I,12)= iqr(data(:,5)); % 75% - 25%.
    kep_persentil(I,13)= quantile(data(:,5),.90) - quantile(data(:,5),.10);
% 90% - 10%.

    % kel percentiles.
    kel_persentil(I,1) = quantile(data(:,6),.10); % 10%.
    kel_persentil(I,2) = quantile(data(:,6),.20); % 20%.
    kel_persentil(I,3) = quantile(data(:,6),.25); % 25%.
    kel_persentil(I,4) = quantile(data(:,6),.30); % 30%.
    kel_persentil(I,5) = quantile(data(:,6),.40); % 40%.
    kel_persentil(I,6) = quantile(data(:,6),.50); % 50%.
    kel_persentil(I,7) = quantile(data(:,6),.60); % 60%.
    kel_persentil(I,8) = quantile(data(:,6),.70); % 70%.
    kel_persentil(I,9) = quantile(data(:,6),.75); % 75%.
    kel_persentil(I,10) = quantile(data(:,6),.80); % 80%.
    kel_persentil(I,11) = quantile(data(:,6),.90); % 90%.
    kel_persentil(I,12)= iqr(data(:,6)); % 75% - 25%.
    kel_persentil(I,13)= quantile(data(:,6),.90) - quantile(data(:,6),.10);
% 90% - 10%.

end

% The file mridata.mat contains the matrixes info, A, kep, kel,
A_persentil, kep_persentil and kel_persentil.
```

```
%
% info gives patient information. Each row represents 1 patient, and has 6
% colums:
% age volume lymphnode_infiltration locoregional_control
% progression_free_survival figo
%
% A, kep and kel contains statistical parameters.
% 8 columns:
% mean median mode min max std skewness kurtosis
%
% A_persentil, kep_persentil and kel_persentil contains the percentile
values for A, kep and kel.
% 13 columns:
% 10% 20% 25% 30% 40% 50% 60% 70% 75% 80% 90% 75%-25% 90%-10%

save('mridata.mat','info','A','kep','kel','A_persentil','kep_persentil','ke
l_persentil');
```

SIMCA (kalibrering)

```
% Simca med PLS-toolbox.
% Turid Torheim
% 21.10.2011

% Skilje mellom pfs=0 og pfs=1 med alder, stadie, volum og statistiske
% parameterar for A, kep og kel som forklaringsvariablar.

% Tal på pasientar.
P = 81;

% % Les inn data frå fil.
% load('MRI.mat');

% Lagrar forklaringsvariablane i matrisa X.
X = [info(:,1:3) info(:,8:end)];

% Lagrar utfall i vektoren pfs (binær, 0/1).
pfs = info(:,7);

% Endrar pfs frå 0/1 til 1/2 for å passe betre til simca.
pfs_endra = pfs;
for i = 1:P
    if pfs_endra(i) == 0
        pfs_endra(i) = 1;
    else
        pfs_endra(i) = 2;
    end
end

% Lagar modellen.
SIMCAmodell = simca(X,pfs_endra);

% Predikerer på grunnlag av modellen.
SIMCApred  = simca(X,SIMCAmodell);

% Predikert klasse for kvar pasient.
resultat = SIMCApred.nclass;
```

```
% Lagar forvirringstabellen.
confusiontable(pfs_endra,resultat)
% Eventuelt: [tabell klassar] = confusiontable(pfs_endra,resultat);
```

SVM (kalibrering)

```
% C-SVM med PLS-toolbox.
% Turid Torheim
% 09.11.2011

% Skilje mellom pfs=0 og pfs=1 med alder, stadie, volum,
lymfeknuteinfiltrasjon og statistiske
% parameterar for A, kep og kel som forklaringsvariablar.
% C-SVM med autoskalering av X-data, RBF-kernel og full kryssvalidering i
paramtersøket.
% Parameteren gamma styrer forma på grensa mellom klassane,
% medan C styrer kor straffbart det skal vere å feilklassifisere.

% Tal på pasientar.
P = 81;

% % Les inn data frå fil.
% load('MRI.mat');

% Lagrar forklaringsvariablane i matrisa X.
X = [info(:,1:3) info(:,8:end)];

% Lagrar utfall i vektoren pfs (binær, 0/1).
pfs = info(:,7);

% Endrar pfs frå 0/1 til 1/2.
pfs_endra = pfs;
for i = 1:P
    if pfs_endra(i) == 0
        pfs_endra(i) = 1;
    else
        pfs_endra(i) = 2;
    end
end

% Val for svm, vel her ingen plotting, autoskalering av X-data og P
oppdelingar i kryssvalideringa til parametersøket (dvs full
kryssvalidering).
% C-SVM er default, seier difor ikkje noko om dette.
val = struct('plots','none','preprocessing','autoscale','splits',P);

% Lagar modellen.
SVMmodell = svmda(X,pfs_endra,val);

% Lagar forvirringsmatrisa.
confusiontable(SVMmodell)
% Eventuelt: [tabell klassar] = confusiontable(SVMmodell);
```

Full kryssvalidering av SVM.

```
% SVM med PLS_Toolbox.
% Turid Torheim
% 06.12.2011

% Full kryssvalidering for SVM.
% Vil skilje mellom klassane pfs=0 og pfs=1 med alder, stadie, volum,
lymfeknuteinfiltrasjon og statistiske
% parameterar for A, kep og kel som forklaringsvariablar.
% Nyttar C-SVM med autoskalering av X-data og RBF-kernel.
% Parameteren gamma styrer forma på grensa mellom klassane, medan C styrer
kor straffbart det skal vere å feilklassifisere.

% Spesifiserer talet på pasientar, P.
P = 81;

% Les inn data frå fil.
% MRI.mat inneheld objektnamn (dvs label for kvar pasient), variabelnamn
% samt matrisa info. info har 81 rader (ein for kvar pasient) og 71 søyler.
% info(:,1:3) er alder, volum og lymfeknuteinfiltrasjon, info(:,7) inneheld
% utfall (0 eller 1) og info(:,8:P) inneheld dei statistiske parameterane
% for A, kep og kel.
load('MRI.mat');

% Lagrar forklaringsvariablane i matrisa X.
X = [info(:,1:3) info(:,8:end)];

% Lagrar utfall i vektoren pfs (binær, 0/1).
pfs = info(:,7);

% Endrar pfs frå 0/1 til 1/2.
pfs_endra = pfs;
for i = 1:P
    if pfs_endra(i) == 0
        pfs_endra(i) = 1;
    else
        pfs_endra(i) = 2;
    end
end

SP = 0;
SN = 0;
FP = 0;
FN = 0;

% Full kryssvalidering.
for i = 1:P

    Xt = [X(1:i,:); X((i+1):P,:)]; % Kalibrering.
    Xv = X(i,:); % Validering.
    pfs_endrat = [pfs_endra(1:i); pfs_endra((i+1):P)]; % Kalibrering.
    pfs_endrav = pfs_endra(i); % Validering.


    % Valg for svm, vel her ingen plotting, autoskalering av X-data og N
oppdelingar i kryssvalideringa til parametersøket (dvs full
kryssvalidering).
    % C-SVM er default.
```

```
    val = struct('plots','none','preprocessing','autoscale','splits',P-1);

    % Lagar modellen.
    SVMmodell = svmda(Xt,pfs_endrat,val);

    % Predikerer på grunnlag av modellen.
    SVMpred  = svmda(Xv,SVMmodell,val);

    % Vektoren med resulat av prediksjonen.
    pfs_pred = SVMpred.pred{2};

    if pfs_endrav == pfs_pred
        if pfs_endrav == 1
            SP = SP+1;
        else SN = SN+1;
        end
    else
        if pfs_endrav == 1
            FP = FP+1;
        else FN = FN+1;
        end
    end

end

ForvirringsTabell = [SP FP; FN SN];
```

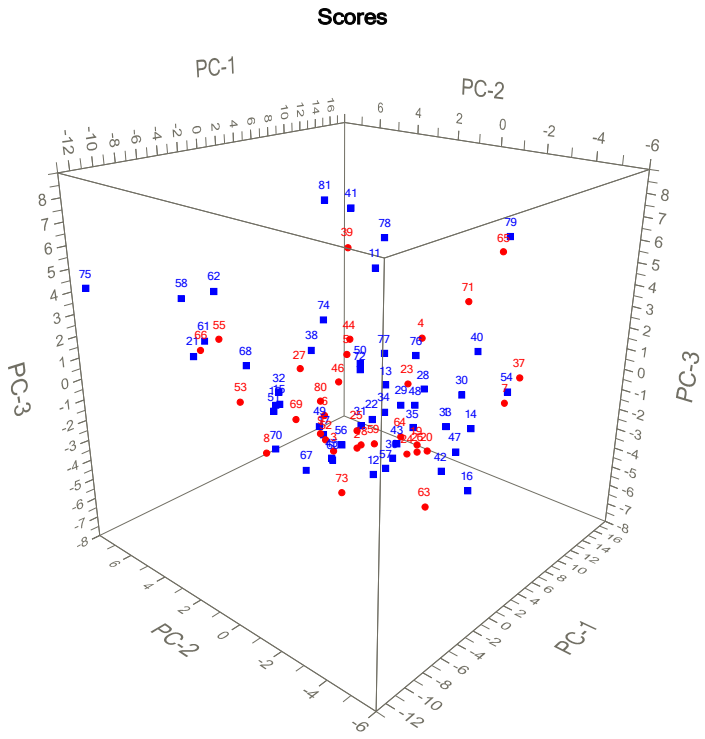
7.2 Resultat frå prediksjonar

Pasient	Pfs	Kmeans PC5, 8 og 10	Kmeans Alle	Kmedian PC5, 8 og 10	Kmedian Alle	LDA 99%	LDA 95%	LDA 90%	SVM	SIMCA
1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	1	1	0	0	0	0
3	1	1	0	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	0	1	1	1	1	0	1	1
6	1	1	0	1	0	1	1	1	0	0
7	1	1	1	1	1	1	1	1	1	1
8	1	0	1	0	1	1	1	0	1	1
9	1	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	1	0	0
12	0	0	0	0	0	0	0	1	0	0
13	0	1	0	1	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	0	1	1	0	1	1	1	1	0	0
17	0	1	0	1	0	1	0	0	0	0
18	1	1	1	1	1	1	1	1	1	1
19	1	1	0	1	0	1	0	1	1	1
20	1	1	1	1	1	1	1	1	1	1
21	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	1	1	0	1	0	0	0	0	0	1
24	1	0	0	0	1	1	1	1	1	0
25	1	1	1	1	1	1	1	1	1	1
26	1	1	1	1	1	1	1	1	1	0
27	1	1	0	1	0	1	0	1	1	0
28	0	1	0	1	1	1	1	1	1	0
29	0	0	0	0	1	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	1	1	1	1	0	0
33	0	1	0	1	1	0	0	0	0	0
34	0	1	0	1	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
36	0	0	1	0	1	1	1	0	0	0
37	1	0	1	0	1	0	1	1	1	1
38	0	1	0	1	0	0	0	1	0	0
39	1	0	1	0	1	0	0	0	1	1
40	0	0	0	0	1	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0
42	0	1	0	1	0	1	1	1	1	0
43	0	1	1	1	1	0	0	0	0	0
44	1	1	1	1	1	1	1	1	1	1
45	0	1	0	1	0	1	1	1	0	0
46	1	1	0	1	0	1	1	1	1	0
47	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0
49	0	1	0	1	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0	0	0	0
52	1	1	0	0	1	1	1	1	1	1
53	1	1	0	1	0	1	1	1	1	1
54	0	0	1	0	1	1	0	0	0	0
55	1	1	0	1	0	1	1	1	1	1
56	0	0	0	0	0	0	0	0	0	0
57	0	1	1	1	1	1	1	1	0	0
58	0	1	0	1	0	1	1	1	0	0
59	1	1	0	1	1	0	0	1	0	1
60	0	0	0	0	0	0	0	0	0	0
61	0	1	0	1	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0	0
63	1	1	1	1	1	1	1	1	1	1
64	1	0	0	0	0	0	0	0	1	1
65	1	0	0	0	0	1	1	1	1	1
66	1	1	0	1	1	1	1	1	1	1
67	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0
69	1	0	0	0	0	1	1	1	1	0
70	0	0	0	0	0	0	0	0	0	0
71	1	1	1	1	1	0	0	0	1	1
72	0	1	0	1	1	0	0	0	0	0
73	1	1	0	1	1	1	1	1	1	1
74	0	1	0	1	0	1	1	0	0	0
75	0	1	0	1	0	0	0	0	0	0
76	0	0	0	0	0	0	1	1	0	0
77	0	0	0	0	0	0	0	0	0	0
78	0	0	0	0	0	0	0	0	0	0
79	0	1	1	1	1	1	1	0	0	0
80	1	0	0	0	1	1	1	1	1	1
81	0	0	0	0	0	0	0	0	0	0

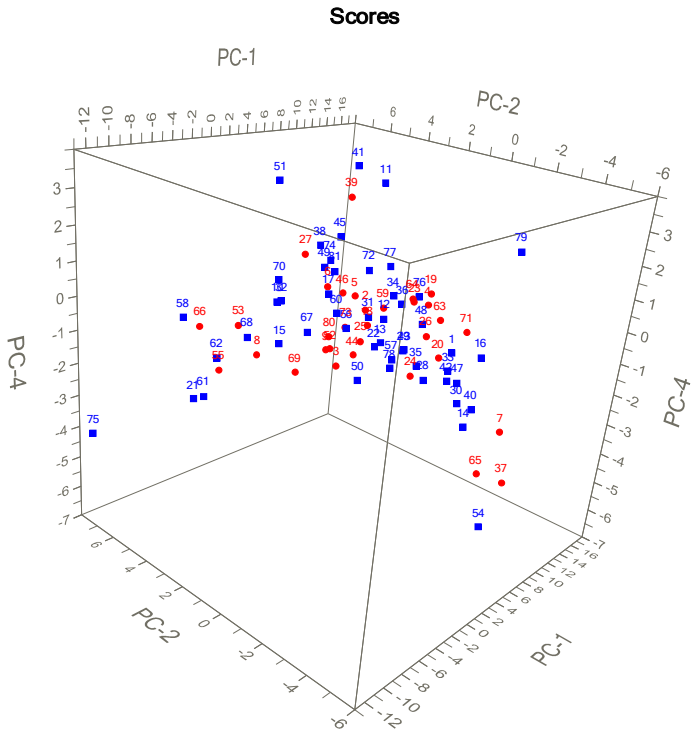
7.3 Plott

Skårplott frå PCA-modellen

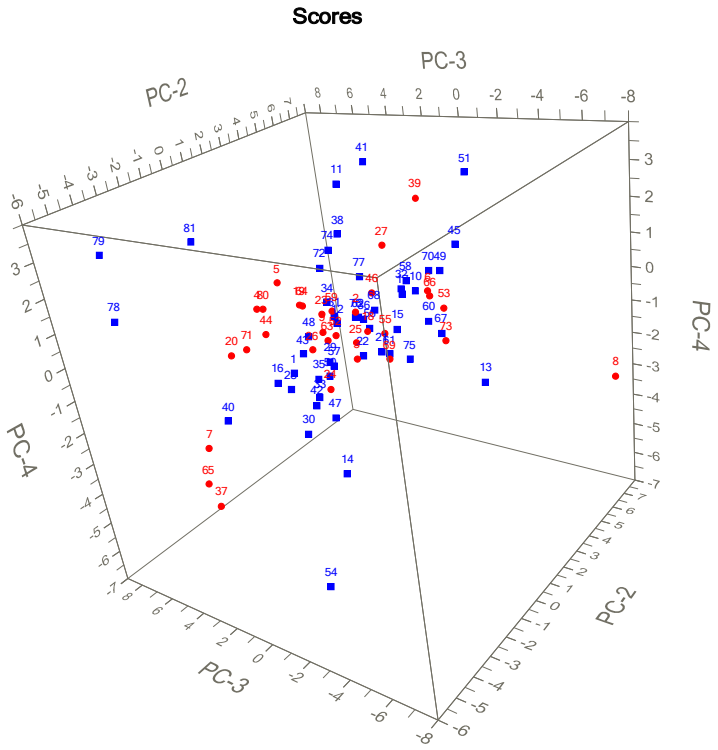
Fleire skårplott for PCA-modellen av alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}.



Figur 54: Skårplott frå PCA-modell av alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Plottet syner skårane for dei tre første prinispalkomponentane. Dei raude punkta syner pasientar som får tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske att (pfs = 0). Laga med Unscrambler.

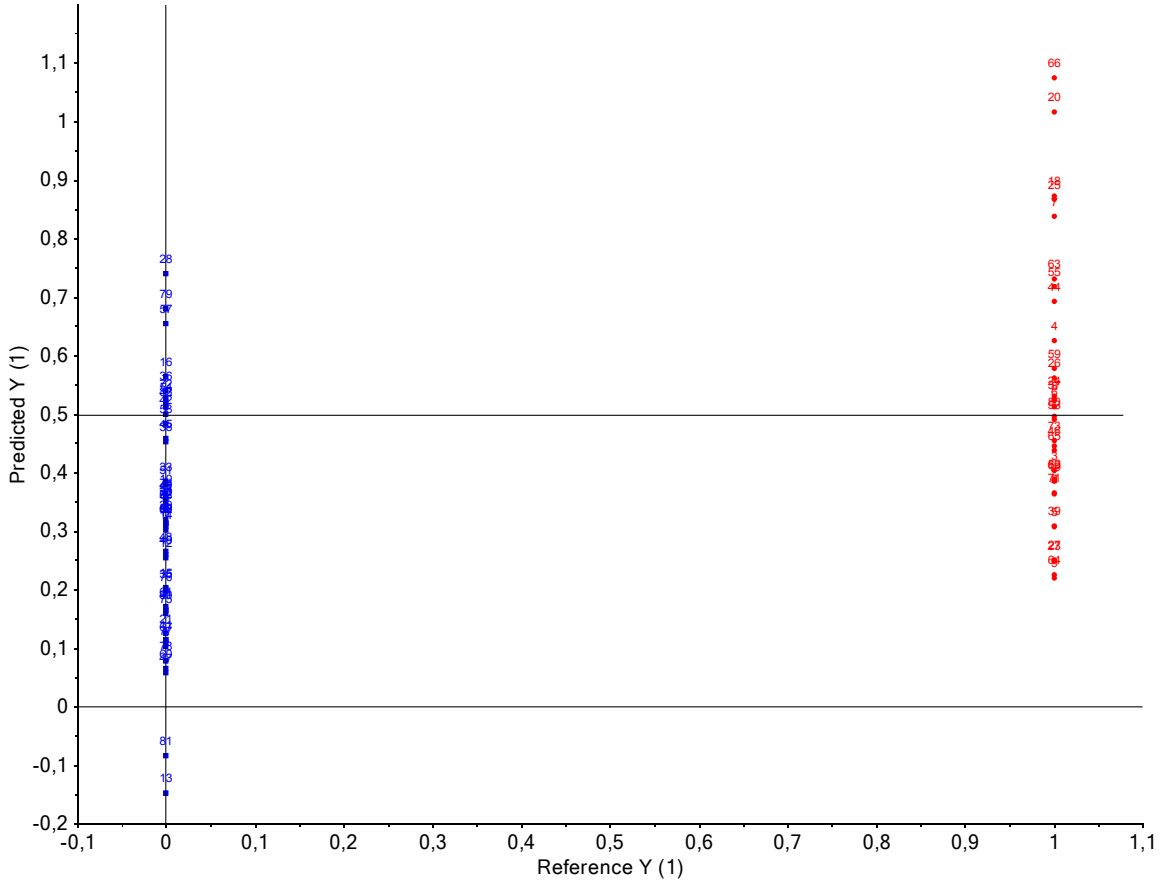


Figur 55: Skårplott frå PCA-modell av alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Plottet syner skårane for PC-1, PC-2 og PC-4. Dei raude punkta syner pasientar som får tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske att (pfs = 0). Laga med Unscrambler.



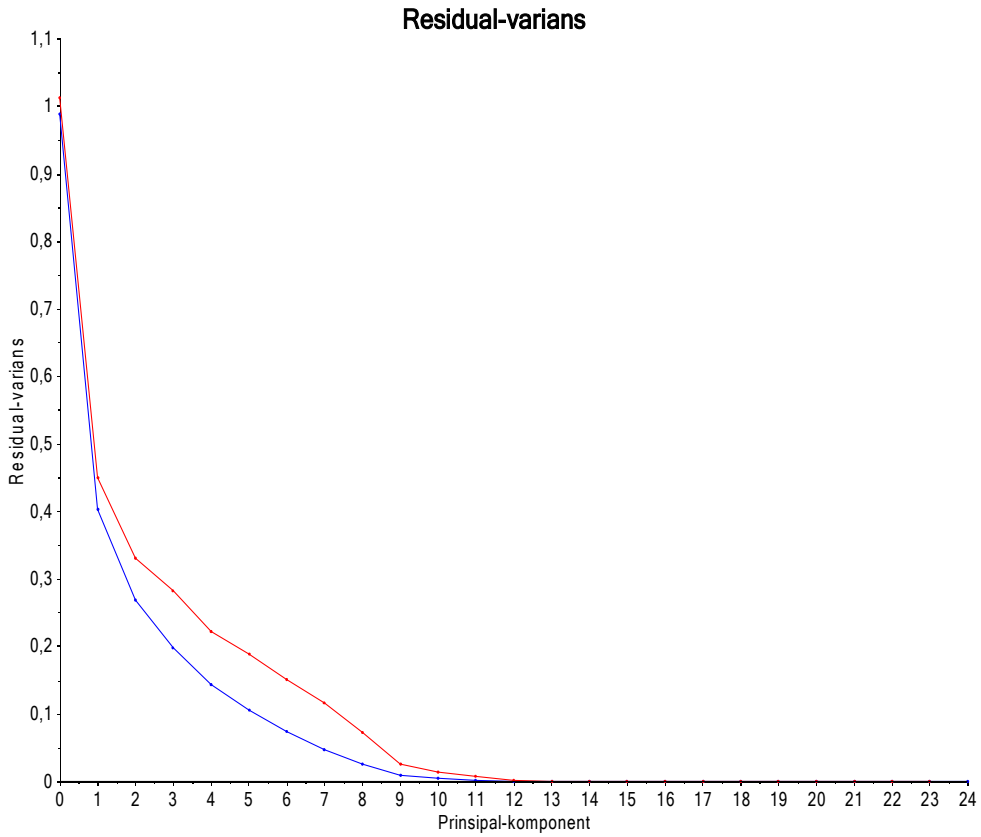
Figur 56: Skårplott frå PCA-modell av alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Plottet syner skårane for PC-2, PC-3 og PC-4. Dei raude punkta syner pasientar som får tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske att (pfs = 0). Laga med Unscrambler.

Regresjon med dei 10 første komponentane frå PCA

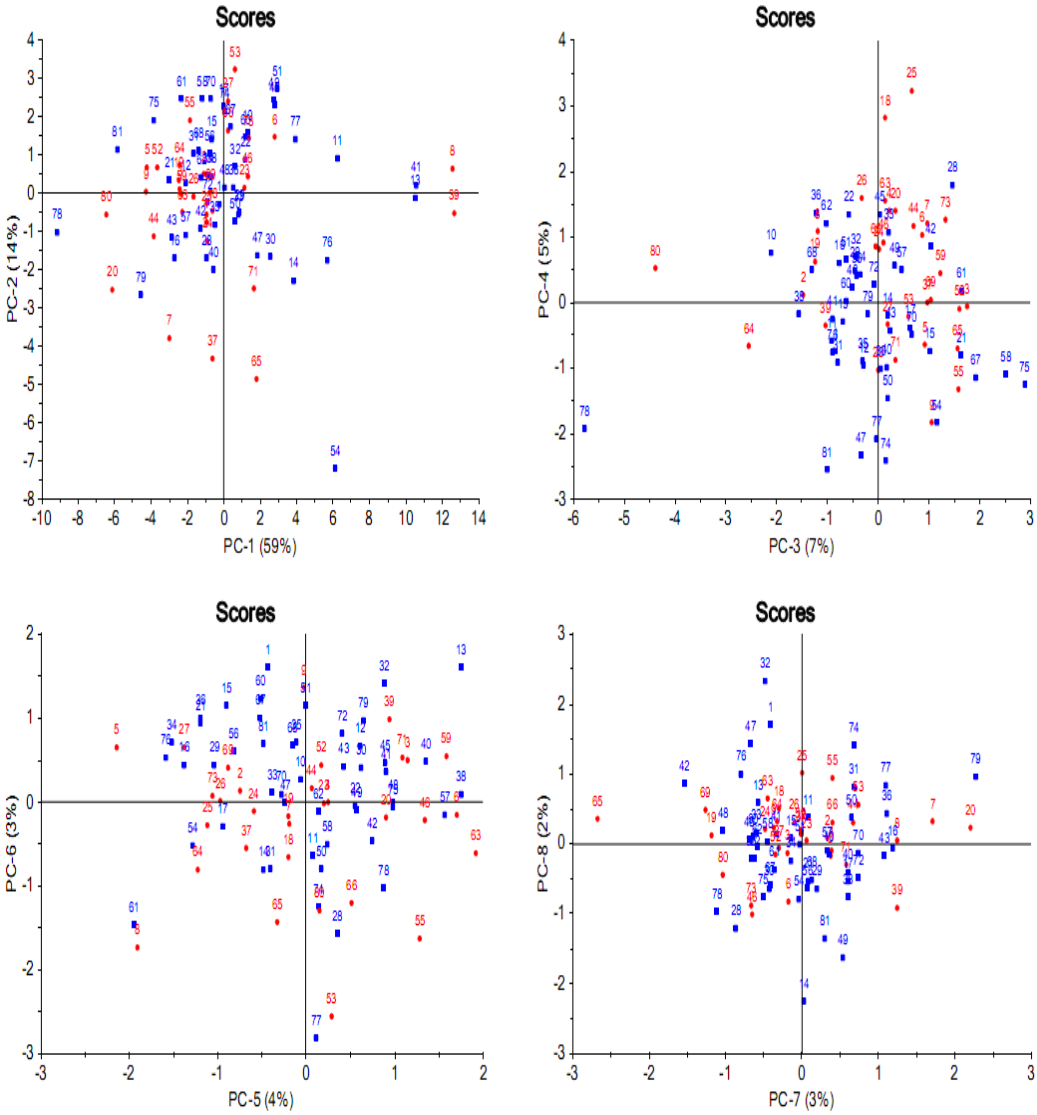


Figur 57: Predikert verdi (y-akse) mot faktisk verdi (x-akse) for regresjonsmodellen med dei 10 første prinispalkomponentane frå PCA-analysen som forklaringvariablar og progresjonsfri overleving (pfs) som respons. Den svarte linja syner y = 0,5. Raude punt er pasientar med pfs = 1, det vil seie pasientar med tilbakefall, medan blå punkt er pasientar med pfs = 0, det vil seie dei som vert friske. Laga med Unscrambler.

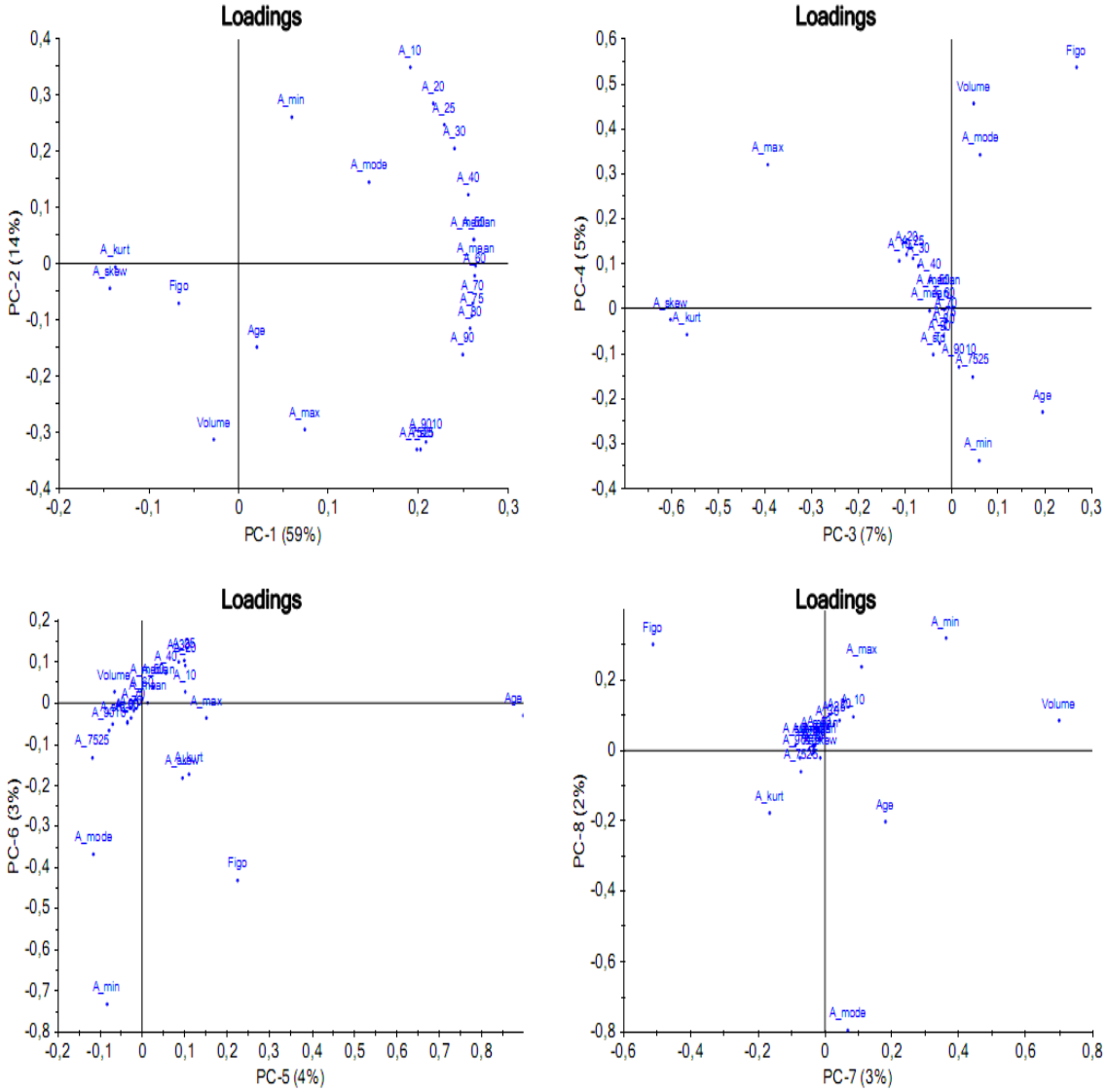
PCA med alder, volum, stadie, og statistiske parameterar for *A*



Figur 58: Residualvariansplott for PCA-modellen med alder, stadie, volum og statistiske parameterar for A som forklaringvariablar, og progresjonsfri overleving (pfs) som respons. Det er nytta fyll kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude gjeld validering. Laga med Unscrambler.

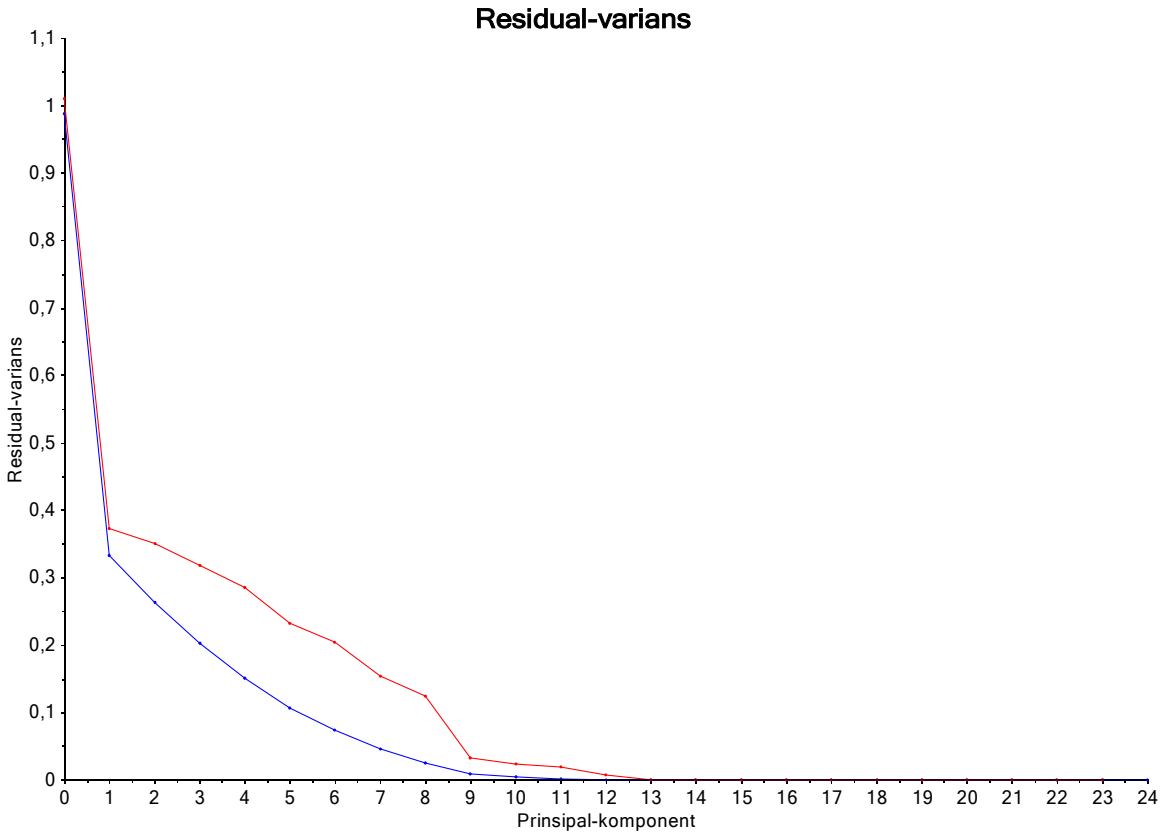


Figur 59: Skårplott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for A som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons. Dei raude punkta representerer pasientar med tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske (pfs = 0).
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

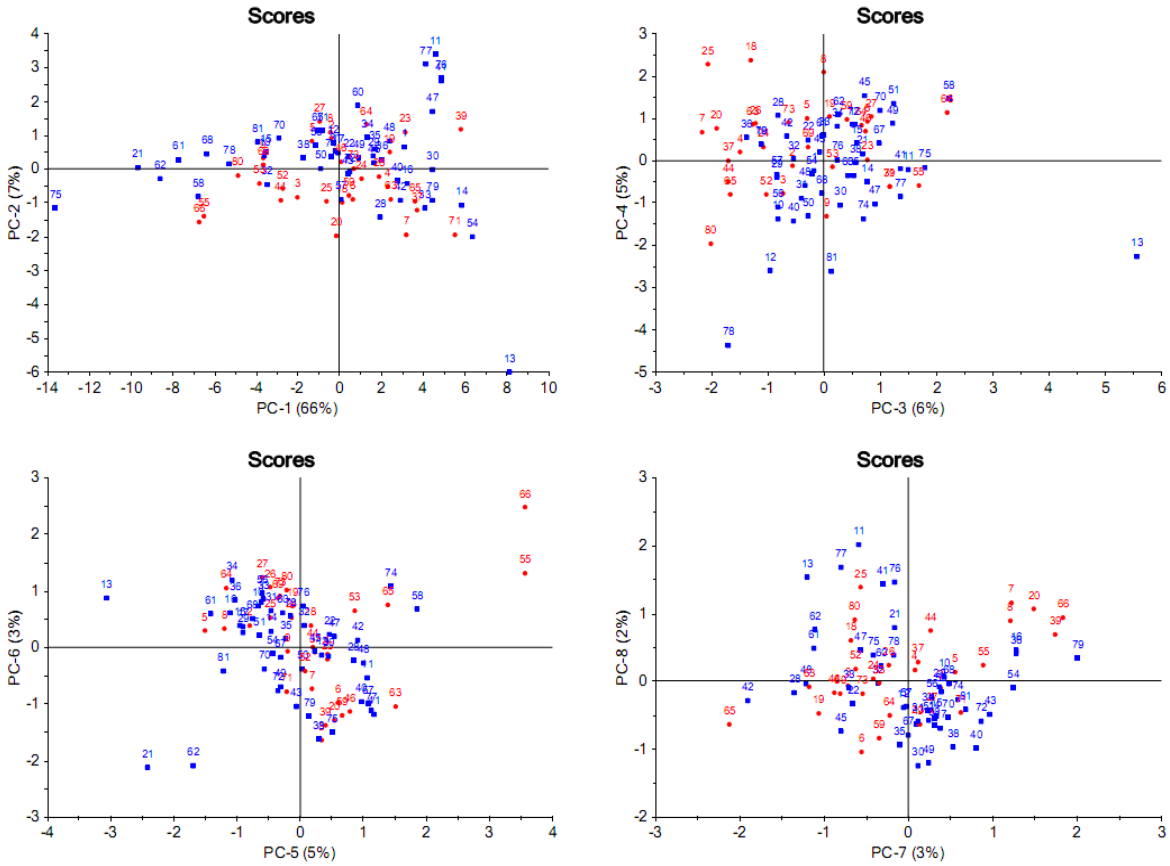


Figur 60: Ladningsplott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for A som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons.
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

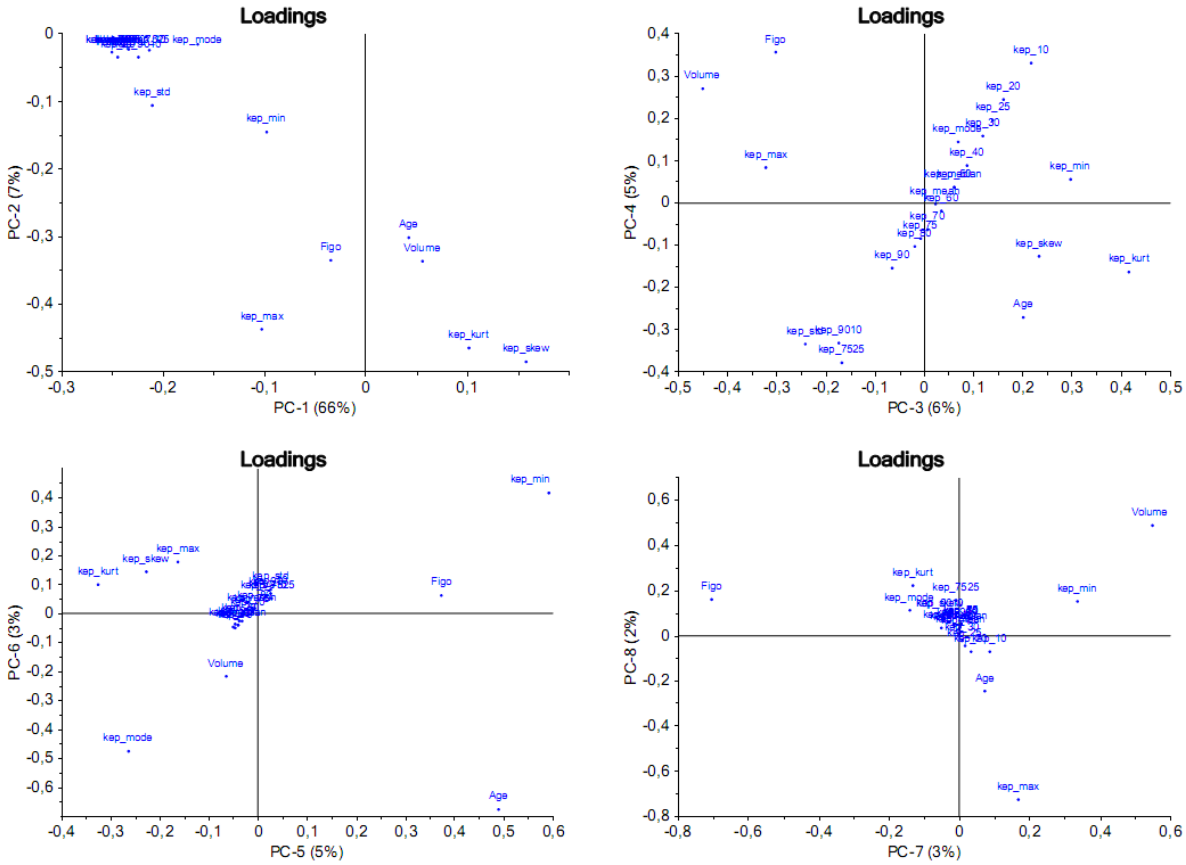
PCA med alder, volum, stadie og statistiske parameterar for k_{ep}



Figur 61: Residualvariansplott for PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{ep} som forklaringvariablar, og progresjonsfri overleving (pfs) som respons. Det er nytta fyll kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude gjeld validering. Laga med Unscrambler.

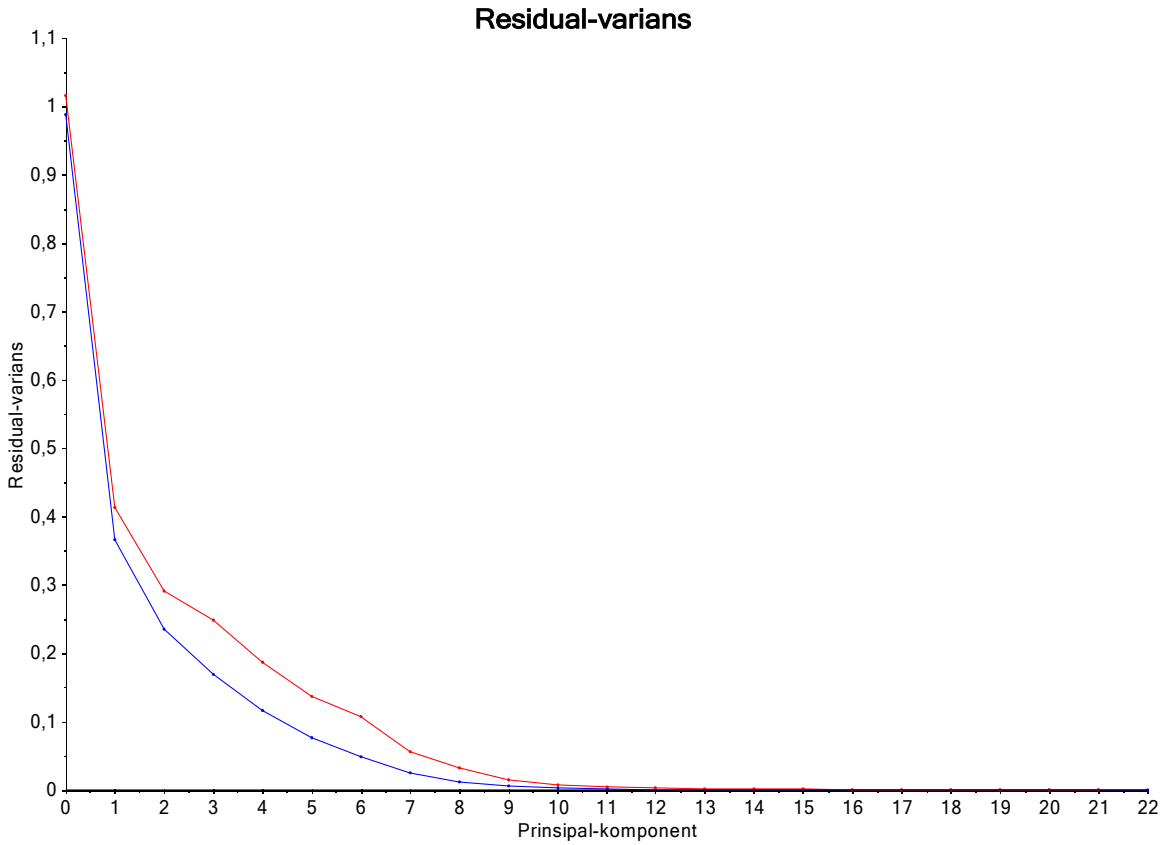


Figur 62: Skår-plott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{ep} som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons. Dei raude punkta representerer pasientar med tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske (pfs = 0).
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

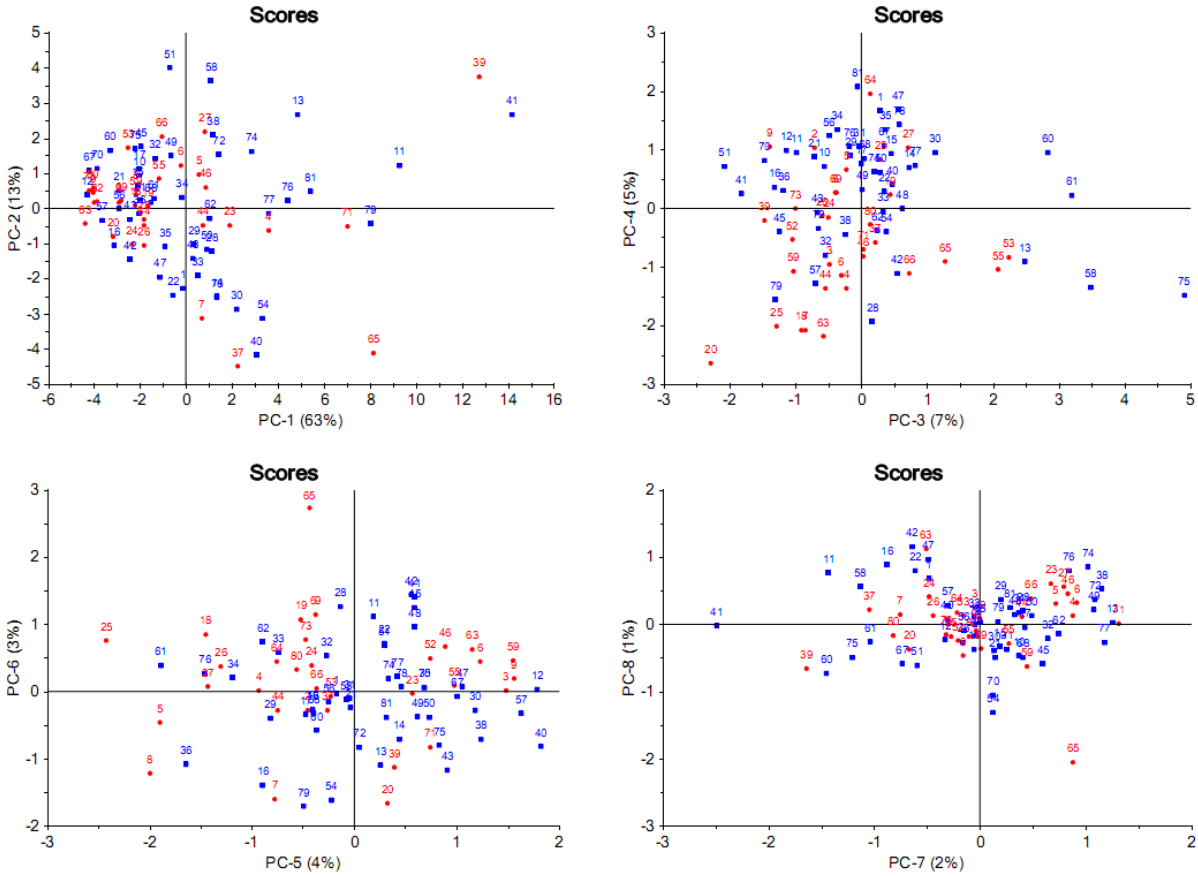


Figur 63: Ladningsplott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{ep} som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons.
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

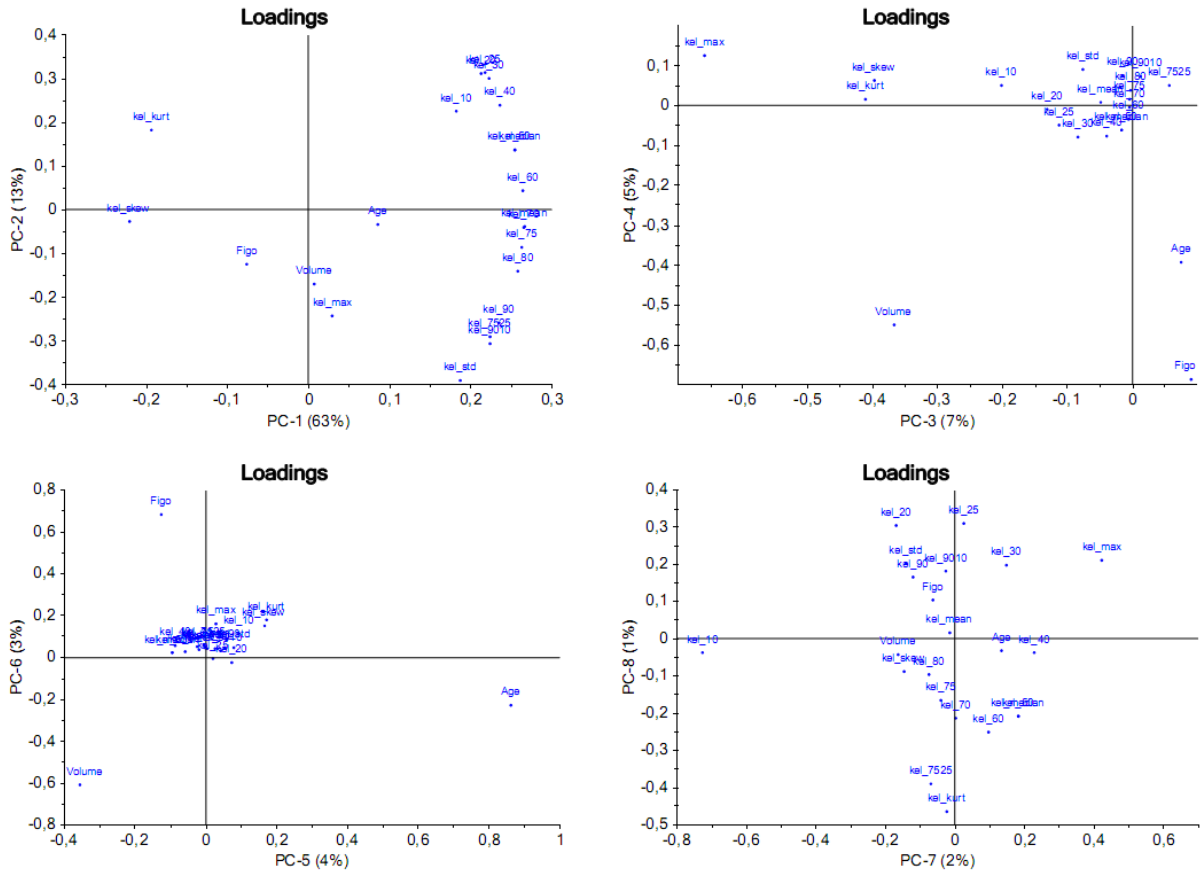
PCA med alder, stadie, volum og statistiske parameterar for k_{el}



Figur 64: Residualvariansplott for PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{el} som forklaringvariablar, og progresjonsfri overleving (pfs) som respons. Det er nytta fyll kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude gjeld validering. Laga med Unscrambler.

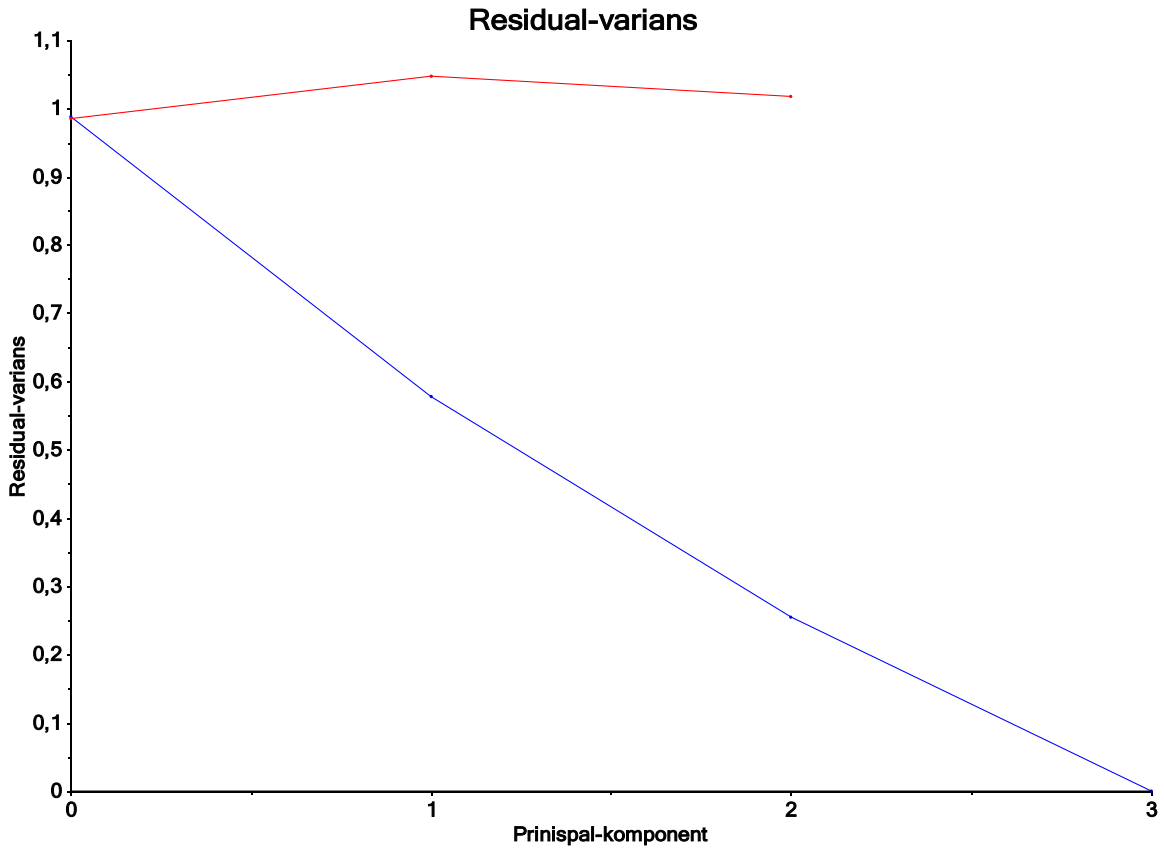


Figur 65: Skårplott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{el} som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons. Dei raude punkta representerer pasientar med tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske (pfs = 0).
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

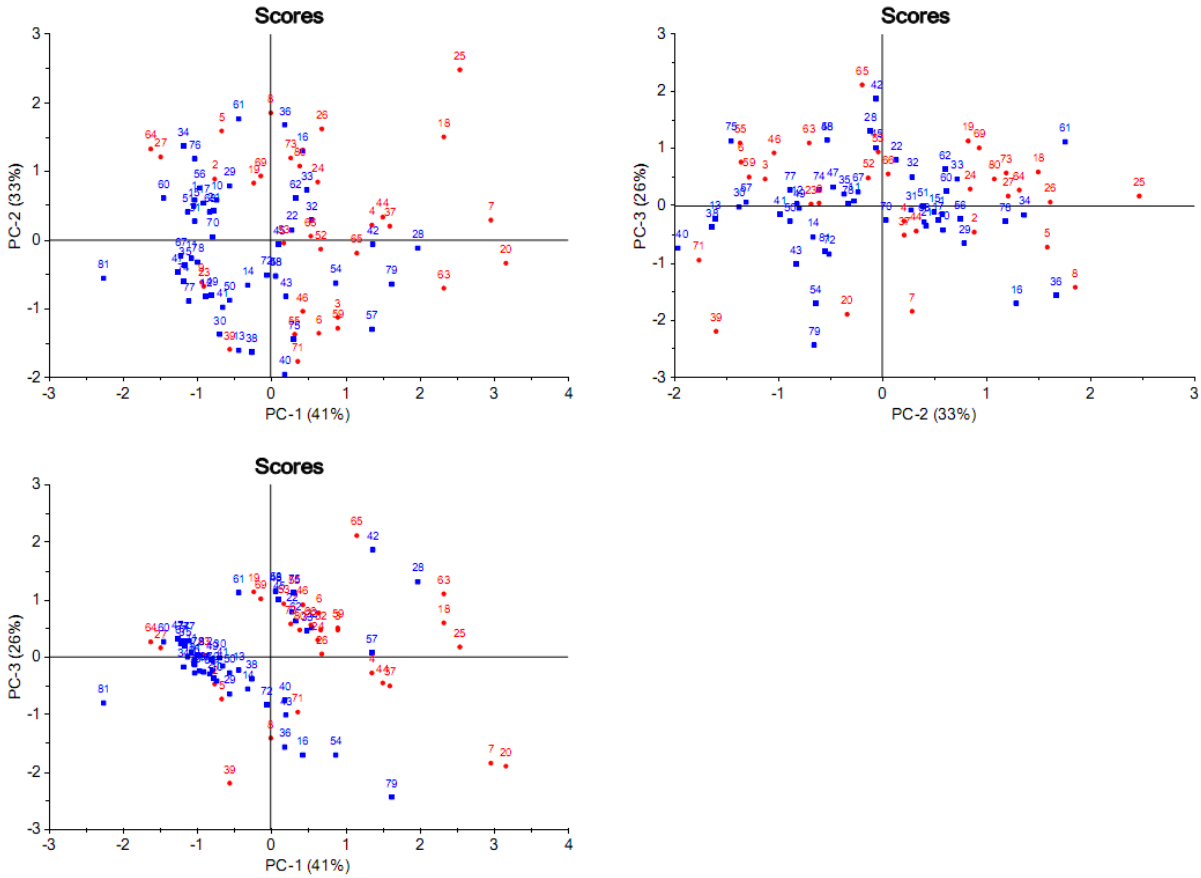


Figur 66: Ladningsplott for dei åtte første prinsipalkomponentane i PCA-modellen med alder, stadie, volum og statistiske parameterar for k_{el} som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons.
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-4 mot PC-3.
Nederst til venstre: PC-6 mot PC-5.
Nederst til høgre: PC-8 mot PC-7.
Laga med Unscrambler.

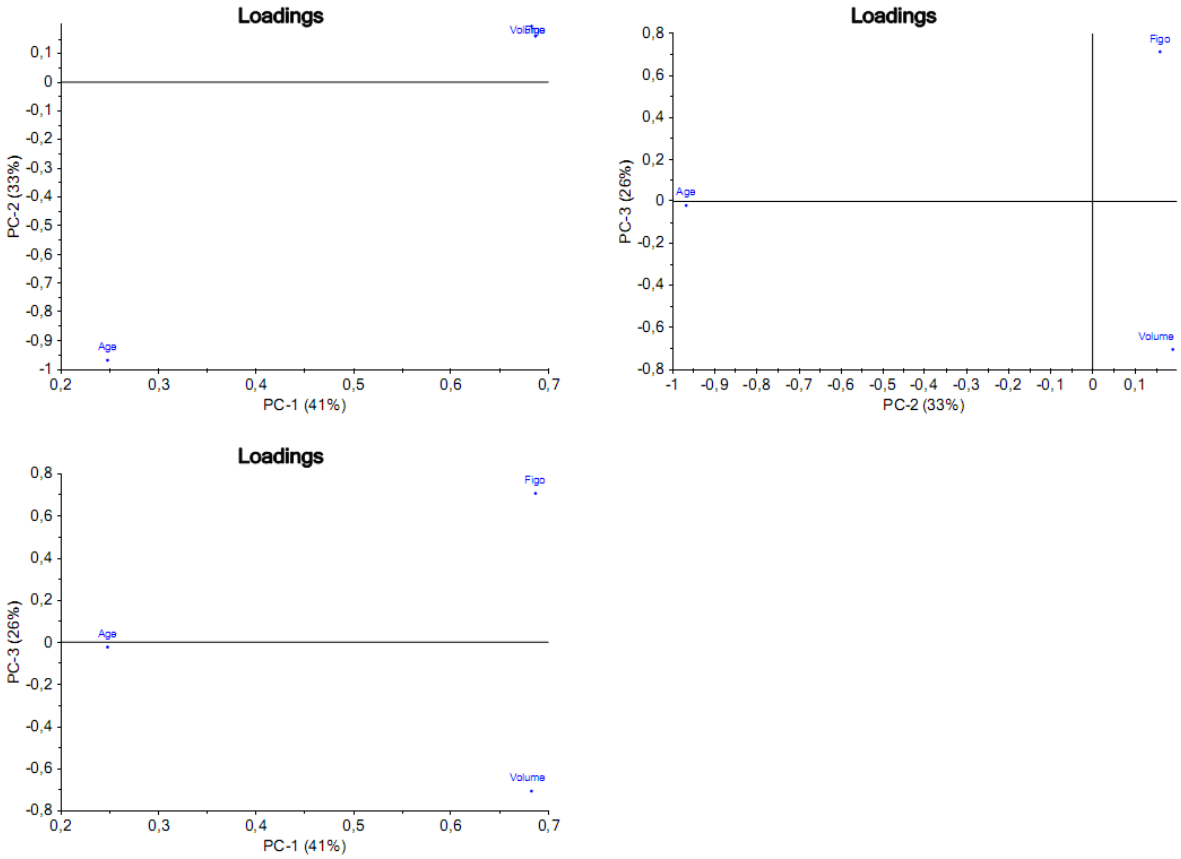
PCA med alder, stadie og volum



Figur 67: Residualvariansplott for PCA-modellen med alder, stadie og volum som forklaringvariablar, og progresjonsfri overleving (pfs) som respons. Det er nytta fyll kryssvalidering. Den blå kurva syner residualvarians for kalibreringa, medan den raude gjeld validering. Laga med Unscrambler.

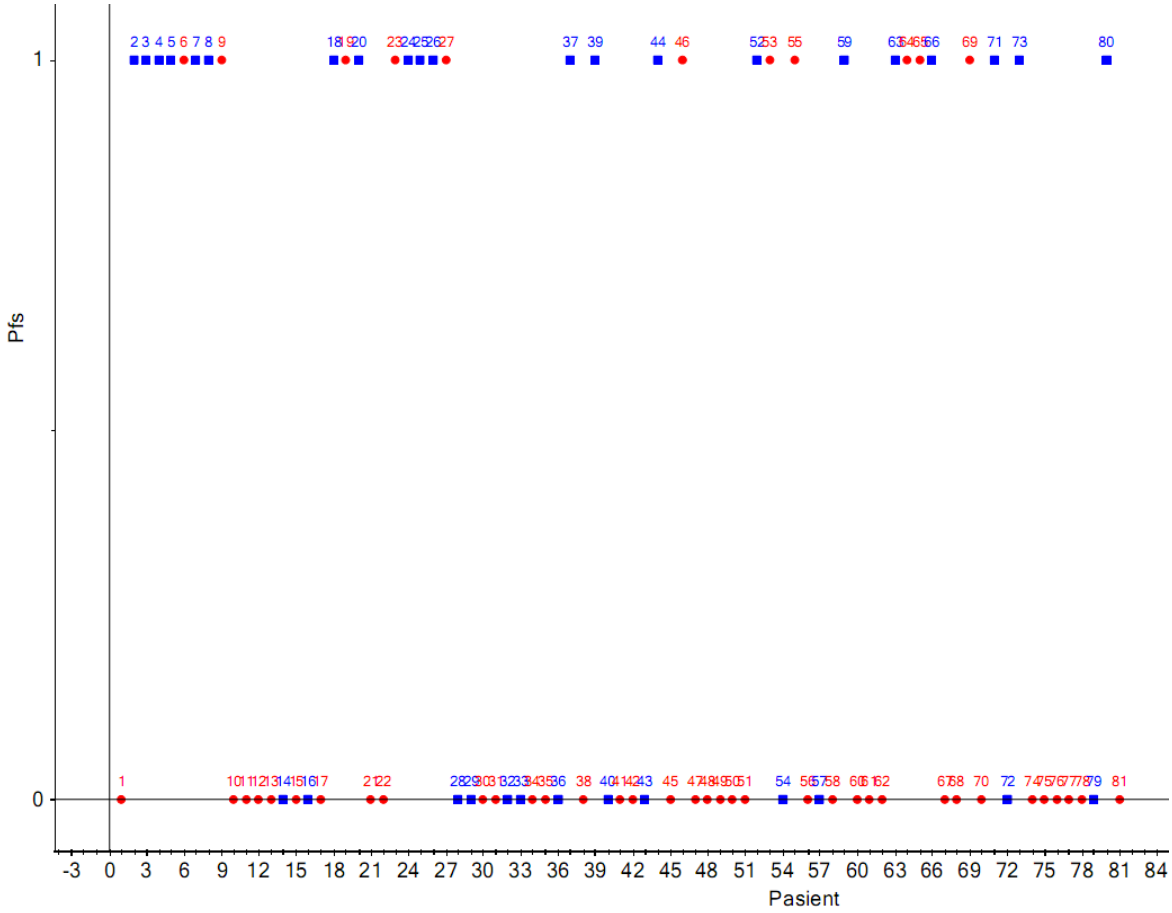


Figur 68: Skår-plott for dei tre prinsipalkomponentane i PCA-modellen med alder, stadie og volum som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons. Dei raude punkta representerer pasientar med tilbakefall (pfs = 1), medan dei blå syner pasientar som vert friske (pfs = 0).
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-3 mot PC-1.
Nederst til venstre: PC-3 mot PC-1.
Laga med Unscrambler.

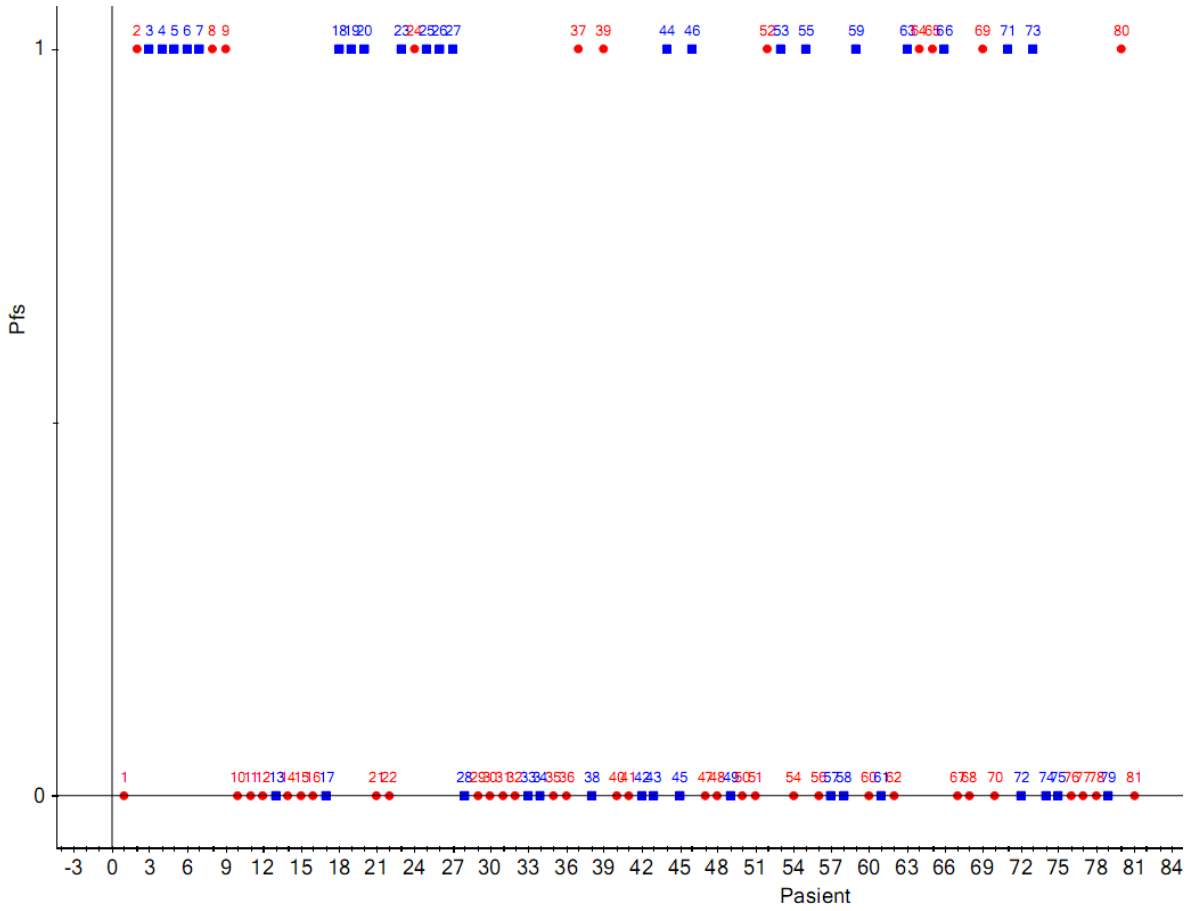


Figur 69: Ladningsplott for dei tre prinsipalkomponentane i PCA-modellen med alder, stadie og volum som forklaringsvariablar, og progresjonsfri overleving (pfs) som respons.
Øvst til venstre: PC-2 mot PC-1.
Øvst til høgre: PC-3mot PC-2.
Nederst til venstre: PC-3 mot PC-1.
Laga med Unscrambler.

K-medians-klyngeanalyse



Figur 70: K-medians-klynger basert på alder, stadie, volum og statistiske parameterar for A, k_{ep} og k_{el}. Pasientnummer langs x-aksen og progresjonsfri overleving langs y-aksen. Dei ulike fargane (raud og blå) syner dei to klyngene. Laga med Unscrambler.



Figur 71: K-medians-klynger basert på tre prinispalkomponentar (5, 8 og 10) frå PCA-modellen. Pasientnummer langs x-aksen og progresjonsfri overleving langs y-aksen. Dei ulike fargane (raud og blå) syner dei to klyngene. Laga med Unscrambler.