CHALMERS

GÖTEBORGS UNIVERSITET

Klassificering av neuropati baserat på svettmönster

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Johan Broberg
Hussein Hamoodi
Henrik Håkansson
Jonathan Kerr

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2018

Klassificering av neuropati baserat på svettmönster

Examensarbete för kandidatexamen i matematisk statistik vid Göteborgs universitet
Hussein Hammodi

Kandidatarbete i matematik inom civilingenjörsprogrammet Automation och mekatronik vid Chalmers
Johan Broberg

Kandidatarbete i matematik inom civilingenjörsprogrammet Informationsteknik vid Chalmers
Henrik Håkansson

Kandidatarbete i matematik inom civilingenjörsprogrammet Kemiteknik vid Chalmers
Jonathan Kerr

Handledare: Aila Särkkä och Anders Hildeman
Examinator: Maria Roginskaya och Ulla Dinger

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2018

Förord

Vi vill tacka våra handledare Anders Hildeman och Aila Särkkä för betydande stöttning och ett stort engagemang.

För projektet har tidslogg kontinuerligt först individuellt per gruppmedlem. Även en projektdagbok har förts på veckobasis där en gruppmedlem per vecka har summerat det utförda arbetet. Här följer en bidragsrapport som förtydligar gruppmedlemmarnas individuella prestationer.

Ansvarsfördelning

Gruppmedlemmarna har sällan haft återkommande ansvarspunkter - däremot har arbetet delats upp så att vissa uppgifter har genomförts gemensamt i grupp och andra har delats ut till par eller enskilda gruppmedlemmar. Ett ansvarsområde sträckte sig oftast över en till två veckor. Exempel på ett ansvarsområde kunde vara att skriva en kodsnutt i R som skulle utföra en viss sak eller skriva på ett stycke i rapporten. Vilka som huvudsakligen jobbade med vad syns i styckena Metod och implementation samt rapportskrivande.

Skrivande av dagbok gjordes veckovis i av alla gruppmedlemmar i roterande ordning.

Planering

Gruppen lade vecka till vecka upp en plan för vad som behövde göras och fördelade sedan gemensamt arbetet sinsemellan. Här bestämdes också arbetstider och att göra listor i Trello, ett webbaserat planeringsverktyg, vilket mestadels sköttes av Henrik. För rapportskrivningen skedde mindre gemensam planering av vad som skulle göras och gruppmedlemmarna skrev nya stycken eller fyllde befintlig text där det ansågs behövas.

Metod och implementation

Vilka metoder som skulle användas diskuterades och beslutades gemensamt i gruppen. Själva implementeringarna, i vårt fall kodsnuttar i R, gjordes antingen i par eller enskilt. Tabellen nedan visar de huvudsakliga implementeringar som varje person gjorde.

Metod	**Jonathan**	**Johan**	**Henrik**	**Hussein**
Sätta samma yttre och inre korsvalidering	X	X	X	
Testa olika modellfelsuppskattningar	X		X	X
Dimensionsreducering (PCA och Backward stepwise)	X	X	X	X
Splittra data till träning och validering		X		X
Grundläggande undersökning av data	X	X	X	X
Möjliggöra modeller baserat på olika delmängder från data	X		X	
Generering av resultat och plottar till rapport			X	
Logistisk regression	X	X		X
Visualisera korrelationen				X
Skapa och testa linjärkombinationer av kovariater			X	X

Rapportskrivande

Tabellen nedan visar vilka personer som huvudsakligen har skrivit på de olika kapitlen eller delkapitlen. Dock har samtliga gruppmedlemmar kontinuerligt läst igenom alla delar i rapporten och kommit med förbättringsförslag.

Kapitel	Delkapitel	Jonathan	Henrik	Hussein	Johan
Abstract		X		X	X
Populärvetenskaplig			X		
Inledning	Bakgrund		X		
	Syfte				X
	Avgränsningar				X
Data		X	X	X	X
Teori och Metod	Klassificering			X	X
	Logistisk Regression			X	X
	Modellfel	X	X	X	X
	Standardisering	X			X
	Dimensionsreducering		X	X	X
	Implementation av modellval	X	X		
	Undersökning av variationer...	X	X		
Resultat			X		
Diskussion	Dimensionsreducering och...		X	X	X
	Val av kroppsdel	X	X		
	Hantering av tröskelvärde		X	X	
	Jämförelse med tidigare forskning				X
	Framtida utveckling	X			X
Slutsats		X			X
Appendix	Jämförelse mellan alla...		X		
	ROC-kurvor		X		
	Matematisk förklaring...	X			X
	Spridning av kovariaterna			X	

Populärvetenskaplig presentation

Att upptäcka sjukdomen perifer neuropati hos personer som är tidigt i sjukdomsförloppet har länge varit svårt eftersom det har saknats enkla och snabba metoder. Vi konstruerade en matematisk modell som bygger på mätningar av patienters svettningar på vad eller fot för att avgöra om patienten bär på sjukdomen eller inte. Modellen presterade mycket bra, vilket innebär att tidigare testmetoder kan komma att ersättas i framtiden.

Perifer neuropati innebär att en eller flera perifera nerver, vilket är nervtrådar som inte är en del av hjärnan eller ryggmärgen, inte fungerar som de ska. Symptom som uppkommer tidigt är stickningar och domningskänslor i armar och ben. Senare uppkommer ofta allvarligare besvär som avsaknad av känsel eller försvagade muskler. Orsaken till att sjukdomen uppkommer är olika från fall till fall, vanliga exempel är diabetes, cellgiftbehandling eller alkoholism. Det är själva orsaken som avgör vilken behandling en patient ska få - men innan det kan ske måste läkaren få reda på om patienten ens har perifer neuropati eller inte. De mest exakta metoder som används idag bygger på omfattande läkarundersökningar, vilka ofta först påbörjas när patienten har haft besvär ett tag. Förhoppningen med den nya testmetoden är att den ska vara lika precis men enklare och mindre tidskrävande, så att patienter redan i första läkarkontakten testas.

Det finns olika typer av neuropati beroende på hur många och vilka typer av nerver som påverkas. I flera typer av perifer neuropati blir små nervtrådar som inte isoleras med ett hölje av substansen myelin tidigt påverkade av sjukdomen. Eftersom dessa typer av nerver påverkar svettningsfunktionen kan onormal svettning tyda på perifier neuropati.

Man kan aktivera svettning hos en patient med hjälp av en ström av ämnen som stimulerar receptorer i svettkörtlarna. I redan etablerade testmetoder har man bland annat försökt uppskatta hur mycket svett som produceras och sedan använda det för att avgöra huruvida patienten har neuropati eller inte. Den nya metoden är att istället använda en specialtillverkad kamera som filmar hela svettförloppet i hög upplösning. Det möjliggör mer information om svettningen än vad de tidigare metoderna kunde ge, exempelvis hur mycket svett varje separat svettkörtel producerar under testet.

I tidigare undersökningar har man sett en stor skillnad i svettförlopp från patienter som lider av neuropati mot kontrollpersoner. Det som vi istället fokuserade på var att ta fram en matematisk modell som använde en mätning från den specialtillverkade kameran för att avgöra om mätningen gjordes på en frisk eller sjuk individ. Inspelningar med kameran utförda på testpersonernas vader och fötter användes, och resultaten tyder på att mätningar från vaderna fungerar bäst för att upptäcka personer som har sjukdomen.

Modellen använde stillbilder från en inspelning med kameran vid tre olika tidpunkter under testet. Från varje bild beräknades ett antal mått som relaterade till hur mycket patienten svettades vid den tidpunkten. Sedan användes dessa uträknade mått som indata till modellen. I det dataunderlag som vi använde fanns det även information om personernas sjukdomstillstånd, det vill säga om de hade neuropati eller inte. Sjukdomstillståndet hade undersökts med andra testmetoder än mätning av svett. När modellen beräknades utnyttjades både information man kände till om svettningarna från de tre stillbilderna och om personernas sjukdomstillstånd.

För att avgöra hur bra modellen fungerade i praktiken skickades bara de mått som beräknats från stillbilderna in till modellen, och informationen om sjukdomstillståndet var dolt. Modellen gav i sin tur tillbaka ett svar hur sannolikt det var att personen som mätningen gjordes på hade perifer neuropati. Modellsvaren jämfördes då med den kända informationen om personens verkliga sjukdomstillstånd. Trots att den matematiska modellen som användes byggde på en relativt enkel metod kunde den i 96 % av fallen korrekt avgöra om en en inspelning gjorts på en sjuk eller frisk person - utan att alltså känna till den informationen på förhand.

Dock finns det en del arbete kvar att göra innan den nya metoden kan tillämpas i vården. Exempelvis behöver man ta ställning till exakt vad man förväntar sig att den matematiska modellen ska åstadkomma. Det går att justera metoden så att den blir bättre på att korrekt ge svaret att de är sjuka för personer som lider av perifer neuropati - men då på bekostnad av att fler som inte bär på sjukdomen felaktigt får svaret är att de är sjuka från modellen. Det går också att justera

så att mätningar från friska personer med stor sannolikhet bedöms vara friska - nackdelen blir då att fler sjuka felaktigt få svaret att de är friska. Om syftet med testet främst är att fånga upp många sjuka är den förstnämnda bäst, men om man istället vill filtrera ut de som säkert är sjuka från resten skulle den senare vara att föredra.

Sammanfattning

Syftet med undersökningen var att avgöra klassificerbarheten av patienter med perfier neuropati baserat på svettmönster med hjälp av logistisk regression. Vår data innehöll tre grupper: kontroller, neuropatiska och obekräftat neuropatiska, individer som misstänks lida av neuropati men ännu inte fått det bekräftat. De obekräftat neuropatiska användes bara i träningmängden och inte i valideringsmängden. Data som användes har mätts på fot eller vad.

Klassificerbarheten undersöktes för data uppmätt på patienters fot, vad samt för båda kroppsdelarna tillsammans. Undersökningen gjordes med två korsvalideringar, en inre för att bestämma ett lämpligt kovariatrum och en yttre för att avgöra den faktiska klassificerbarheten.

Det bästa sättet att klassificera enligt undersökningen var att använda data från enbart vader och att använda dimensionsreducering med principialkomponentanalys för 15 kovariater. Med hundra simuleringar av vår modell blev medelvärdet av arean under grafen från receiver operating characteristic-kurvan 0.96 med en standardavvikelse på 0.01. Om de två olika klassificeringsfelen värderades lika högt och modellen designades så att båda feltyperna hade lika stor sannolikhet kunde den använda metoden klassificera med ca 10 % fel. Under undersökningen fanns problem med att datamängden innehöll få neuropatiska patienter. För framtida forskning hade det varit intressant att utöka mängden sjuka.

Abstract

The purpose of this investigation was to determine the ability to classify peripheral neuropathy patients based on data from sweat patterns using logistic regression. Our data contained three groups: controls, neuropathics and individuals believed to be neuropathic but were not yet confirmed. Subjects from the last group was only used as part of a training set and not as validation set. The data was measured from calves and feet.

The ability to classify patients was examined by using data from feet, calves, or both. Our investigation was conducted using two nested crossvalidations, one inner to determine the appropriate dimensional space and one outer to evaluate the performance of the classification.

The best way to classify was determined to be on data from only calves with dimensionality reduction using principal component analysis from 15 covariates. With one hundred simulations of this method the area under the curve for the receiver operating characteristic-curve was on average 0.96 with a standard deviation of 0.01. If the two possible types of classification errors were considered equal and the model was designed to have the same proportion of errors the method was able to classify with only about 10% error. A problem during the investigation was that the data being used had very few neuropathic patients. For future research it would be interesting to expand the data to contain more neuropathics.

Innehåll

1 Inledning

Perifer neuropati[1] är ett generellt begrepp för dysfunktionalitet av en eller flera perifera nerver, nervtrådar som inte är en del av hjärnan eller ryggmärgen. Tidiga symptom är exempelvis stickningar och domningskänslor i armar och ben. Senare uppkommer allvarligare komplikationer så som avsaknad av känsel eller försvagade muskler. Den underliggande orsaken till besvären är oftast helt individuell för varje fall, men vanliga exempel är diabetes, cellgiftbehandling eller alkoholism. Vilken typ av behandling som är tillämplig beror helt på orsaken till besvären. Eftersom tillståndet förvärras med tiden är det viktigt att behandling inleds så tidigt som möjligt.

Det finns olika typer av neuropati beroende på hur många och vilka typer av nerver som påverkas. I flera typer av perifer neuropati blir små nervtrådar som inte isoleras med ett hölje av substansen myelin, *omyeliniserade* nervtrådar[2], tidigt påverkade av sjukdomen. Därför kan abnormitet i funktionen hos de omyeliniserade nervtrådarna indikera ett tidigt stadium av perifer neuropati. Då svettkörtlar stimuleras av omyeliniserade nervtrådar kan mätning av svettutsöndring[3] användas för att detektera abnormitet i funktionen, vilket då kan innebära både under- och överproduktion av svett.

Idéen att mäta svettutsöndring har tillämpats i flera olika varianter av tester. En av de mest använda metoderna är *Quantitative sudomotor axon reflex test* (QSART). Metoden går ut på att först stimulera svettkörtlarna på en liten yta med hjälp av en ström av ämnen som binder till receptorer i svettkörtlarna. Därpå mäts luftfuktigheten över ytan som stimulerats för att uppskatta volymen svett som produceras över tid från ytan. I en studie[4] visades det att 74% av 125 personer som led av neuropati uppvisade anormala resultat på QSART jämfört med en kontrollgrupp.

Provitera et al. [5] presenterade en modifierad variant av svettestet där en specialtillverkad kamera filmade svettningen i hög upplösning under 5 minuter. Denna metod ger även information om svettningen hos varje svettkörtel individuellt. Loavenbruck et al. [6] undersökte resultat från kameratestet mer ingående med fokus på måtten total svettning, svetthastighet per svettkörtel samt densitet av svettkörtlar på olika kroppsdelar. Det konstaterades att den största skillnaden mellan kontrollgruppen och de neuropatiska personerna var som störst för stimuleringar på vad och fot. Dessutom visades det att jämfört med kontrollgruppen hade de neuropatiska lägre svetthastighet per svettkörtel i 90% av fallen på vaden och 80 % av fallen på foten.

Resultaten i Loavenbrucks undersökning tyder på att personer med perifer neuropati uppvisar resultat från svettester som i hög grad är skilda från friska. Med den kunskap om maskininlärning som numera finns lättillgänglig kan det tänkas att filminspelningar från svettestet lämpar sig för att automatiskt bestämma patientens hälsotillstånd. Vi undersöker därför om det, med data från svettester, går att ta fram en klassificeringsmodell som kan avgöra om en ny inspelning med kameran från ett svettest kommer från en frisk eller neuropatisk person.

1.1 Syfte

Syftet är att undersöka hur väl klassificeringsmetoder kan tillämpas för att identifiera patienter som lider av perifer neuropati, baserat på observationer med de 15 olika måtten. Vi vill även ta reda på vilka mätningar, utifrån om de är uppmätta på vad, fot eller båda kroppsdelarna sammanslaget, som lämpar sig bäst för klassificering.

1.2 Avgränsningar

Klassificering och maskininlärning är breda områden som omfattar fler relevanta metoder än vad som kunnat behandlas i detta projekt. Vi har valt att utgå från klassificeringsmetoden logistisk regression som är en relativt enkel metod, för att snabbt kunna utföra och få förståelse för klassificering. Vi har fokuserat på att utforska olika varianter av logistisk regression för att hitta en så bra slutgiltig modell som möjligt för denna klassificeringsmetod.

Under arbetets gång fick vi tillgång till nytt filmmaterial av patienters svettproduktion som inte har används. Detta material kom så pass sent och för att kunna använda det hade de spatiella måtten behövts räknas ut även för dessa filmer så att informationen skulle kunna användas tillsammans

med den andra datan. Hade vi haft möjlighet att generera ytterligare data från dessa filmer så kunde det möjligtvis ha bidragit till en bättre klassificeringsmodell.

2 Data från svettmönster

Arbetet har gjorts på en datamängd som består av fem olika spatiala mått uträknade från de bilder som tagits med Loavenbrucks kamera, på någon av försökspersonens fötter eller vader. Data kommer från 401 filminspelningar av svettmönster från 185 olika försökspersoner, där en inspelning motsvarar en observation. Försökspersonerna består av både personer som lider av neuropati och personer som inte gör det. För varje observation har de spatiala måtten beräknats vid tre olika tidpunkter: 1, 10 samt 30 sekunder. Ett spatialt mått för en viss tidpunkt benämns här som kovariat och eftersom det finns fem mått vid tre olika tidpunkter finns det alltså sammanlagt 15 kovariater.

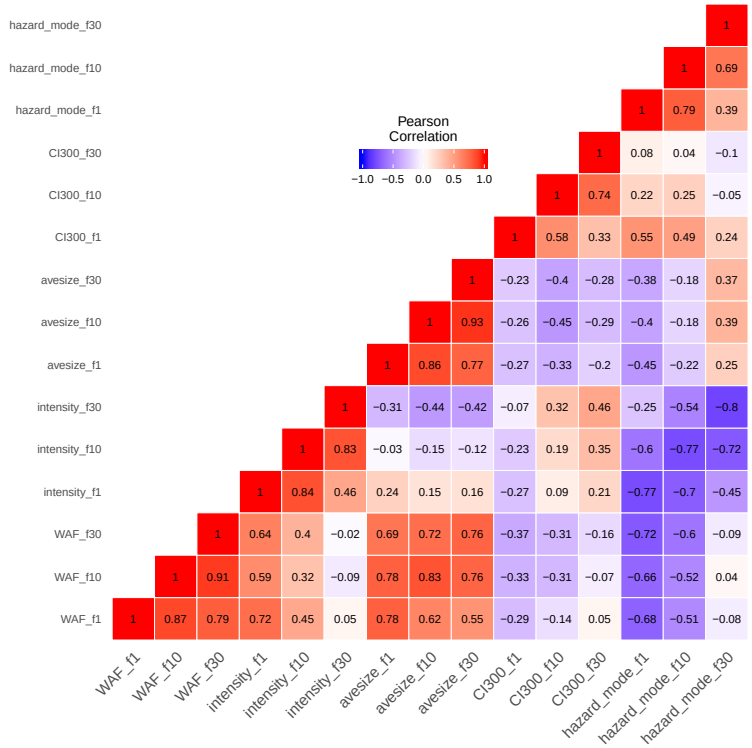
De fem spatiala måtten är:

- **WAF (Wetness Area Fraction):** Andel area av bild som täcks med svett
- **Intensity:** Antal separata svettfläckar som andel av den totala bildarean. (Då två svettfläckar växer ihop räknas de som en enskild fläck)
- **Avesize (Average Size):** Den genomsnittliga arean av svettfläckar på en bild mätt i antal pixlar
- **CI300:** Ett klusterindex där högt värde indikerar att svettfläckar tenderar att existera nära varandra och lågt värde indikerar att svettfläckar är utspridda. (Se definition i C.0.1)
- **Hazard Mode:** Ett mått på den genomsnittliga tomma ytan mellan fläckar på en bild (Se definition i C.0.2

För samtliga observationer noteras, förutom de 15 måtten, även om patienten har neuropati eller om denna tillhör en kontrollgrupp med friska. Kontrollgruppen består av 120 personer från vilka det finns 301 observationer. 153 av observationerna i kontrollgruppen är uppmätta på patieternas vad och 148 på deras fot. De personer som är diagnostiserade med neuropati uppgår till 18 personer från vilka det totalt finns 27 observationer, 18 på vad och 9 på fot. Gruppen neuropatiska är alltså klart underrepresenterad. Det finns även observationer från 47 personer som själva uppgett att de har symptom av neuropati, men som inte fått diagnosen perifer neuropati bekräftad. Antalet observationer från den gruppen uppgår till 73 observationer, 45 på vad och 28 på fot. Dessa observationer har klassificerats som neuropatiska, men då personernas tillstånd inte är helt säkra har de särbehandlats från de neuropatiska med bekräftad diagnos i samband med klassificeringen. Tabell 1 visar en översikt över dessa observationer.

		Sjukdomsstatus			
		Frisk	Bekräftat Neuropatisk	Obekräftat Neuropatisk	Totalt
Kroppsdel	Fot	148	9	28	185
	Vad	153	18	45	216
	Totalt	301	27	73	**401**

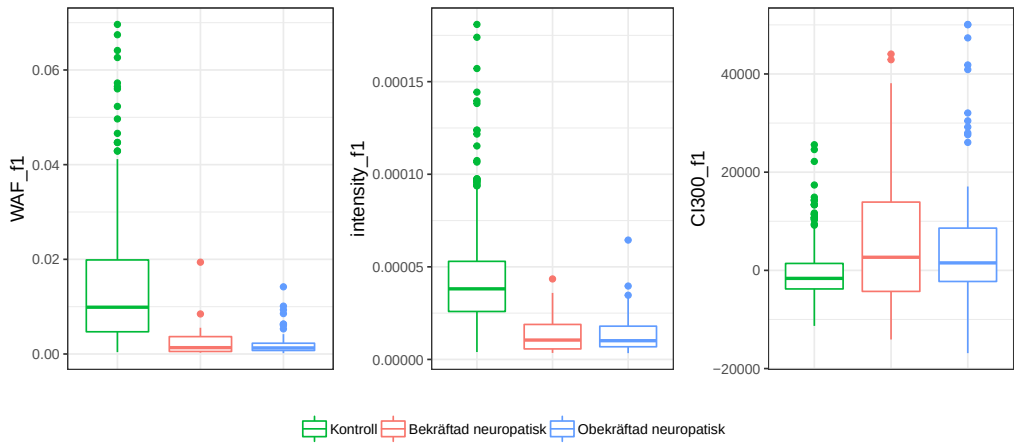
Tabell 1: Tabell som visar antalet observationer för de olika patienterna samt för de olika kroppsdelarna mätningar gjorts på



Figur 1: Visualisering av korrelationsmatrisen av kovariaterna. Blå indikerar negativ korrelation och röd indikerar positiv korrelation. Ljusare färg innebär korrelation närmare 0.

Många av kovariaterna har en stark korrelation till varandra, vilket kan ses i korrelationsmatrisen i figur 1. Exempel på korvariater med stark korrelation är WAF, Avesize och Hazard mode. CI300 uppvisar däremot en relativt svag korrelation till de övriga.

För att jämföra de tre grupperna och se hur de beter sig med olika mätningar, så har vi skapat låddigram som visar hur mätningen på kontroll, bekräftad neuropatiska och obekräftad neuropatiska sprider sig vilket vi kan se i figur 2 med tre olika typer av mätningar. Från figuren ser vi tydligt att mätningarna på de obekräftade patienter ligger mycket nära på de neuropatiska patienter.



Figur 2: Låddigram som visar hur olika kovariater sprider sig för de olika patienter beroende på deras hälsoläge . I den vänstra figuren har vi WAF_{f1}, mittersta har vi intensity$_{f1}$ och högra har vi $CI300_{f1}$.

3 Teori och metod

Problemet som undersöktes var kortfattat att utveckla en metod för att förutsäga om nya observationer, med okänt tillstånd, kommer från friska eller neuropatiska personer. Detta kan brytas ned till 3 huvudsakliga delproblem: vilken klassificeringsmetod som skall användas, hur valet av kovariater som används i klassificeringsmetoden genomförs samt hur modellens noggrannhet kan uppskattas.

I denna del beskrivs hur klassificering och logistisk regression tillämpades. Modellfel uppskattades med antingen korsvalidering tillsammans med AUC av ROC eller AIC_c. Dimensionsreducering gjordes med Stepwise Backward eller principalkomponentanalys.

3.1 Klassificering

I klassificeringsproblem vill man tilldela en observation av uppmätta värden $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ till en av D diskreta klasser $y_i = c_d$ där $d = 1, \cdots, D$ [7]. För N stycken observationer ges varje observation som en rad i en matris $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ och motsvarande klasser ges som en kolumnvektor $\mathbf{y} = (y_1, \ldots, y_N)^T$ enligt tabell 2. Varje vektorelement x_{ij} motsvarar här värdet för en av k kovariater.

respons $\mathbf{y}$	kovariater $\mathbf{X}$		
y_1	x_{11}	$\ldots$	x_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_N	x_{N1}	$\ldots$	x_{Nk}

Tabell 2: Matris för realiserad data där en rad motsvarar en observation.

y kallas här för responsvariabel och kan ses som en funktion av observationen $\mathbf{x}$. I praktiken är det ofta omöjligt att hitta denna funktion exakt så istället approximeras y med

$$\hat{y} = f(\mathbf{x}),$$

där f utgörs av en klassificeringsmodell och $\hat{y}$ är den estimerade klassen observationen $\mathbf{x}$ tillhör. I vårt fall har observationerna klassificerats beroende på om mätning utförs på en neuropatisk eller frisk person och därför har y två möjliga klasser:

$$y = \begin{cases} 0 & \text{om observationen uppmätts på frisk person} \\ 1 & \text{om observationen uppmätts på neuropatisk person} \end{cases}$$

För att bestämma $f(\mathbf{x})$ så att den beskriver sambandet mellan $\mathbf{x}$ och y väl används tidigare uppmätta observationer där den korrekta klassen är känd. Dessa observationer med tillhörande respons används för att träna klassificerningsmodellen, vilket innebär att $f(\mathbf{x})$ anpassas till observationernas riktiga klasser.

3.2 Logistisk regression

Logistisk regression är en klassisk och väl beprövad klassificeringsmetod [8]. Den utgår från att observationerna $\mathbf{x}$ och y kommer från en stokastisk vektor χ respektive en stokastisk variabel Y och returnerar ett estimat av den betingade sannolikheten $P(Y = 1 \mid \chi = \mathbf{x})$. Att metoden estimerar sannolikheten att observationen tillhör en viss klass istället för bara klassen gör att metoden blir både flexibel och tolkningsbar av användaren Logistisk regression kan även användas för ett godtyckligt antal klasser men i vårt fall räcker det med endast klasserna frisk, $Y = 0$ och neuropatisk, $Y = 1$.

Med logistisk regression vill man använda en linjär funktion av $\mathbf{x}$ för att beskriva $P(Y = 1 \mid \chi = \mathbf{x})$. Samtidigt behöver $P(Y = 1 \mid \chi = \mathbf{x}) + P(Y = 0 \mid \chi = \mathbf{x}) = 1$ gälla och det returnerade värdet från modellen ska vara i intervallet $[0, 1]$ för att resultatet skall motsvara en sannolikhet [9]. Så $P(Y = 1 \mid \chi = \mathbf{x}) = \beta_0 + \beta_1 x_1 \ldots \beta_k x_k$ är nödvändigtvis inte en giltig modell. Istället antas det

linjära sambandet gälla för logaritmen av oddsen av sannolikheten $P(Y = 1 \mid \chi = \mathbf{x})$, kallad *logit*, vilket innebär att

$$logit = log(odds) = \ log\left(\frac{P(Y = 1 \mid \chi = \mathbf{x})}{1 - P(Y = 1 \mid \chi = \mathbf{x})}\right) = \ \beta_0 + \beta_1 x_1 \dots \beta_k x_k =$$

$$log\left(\frac{P(Y = 1 \mid \chi = \mathbf{x})}{1 - P(Y = 1 \mid \chi = \mathbf{x})}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \implies$$

$$\frac{P(Y = 1 \mid \chi = \mathbf{x})}{1 - P(Y = k \mid \chi = \mathbf{x})} = \exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right) \implies$$

$$g(\boldsymbol{x}) := \quad P(Y = 1 \mid \chi = \mathbf{x}) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right)}. \tag{1}$$

Funktionen $g(\boldsymbol{x})$ är alltså resultatet från den logistiska regressionsmodellen. Själva modellantagandet i logistisk regression bygger på att responsvariabeln för varje observation Y_i, $i = 1, \cdots, N$ följer en Bernoulli-fördelning. Y_i antar alltså värdet 1 med sannolikhet π_i och värdet 0 med sannolikheten $(1 - \pi_i)$ där $P(Y_i = 1 | \chi = x) = \pi_i$. Vi har att:

$$Y_i \sim \text{Bernoulli}(\pi_i),$$

$$P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2}$$

För att uppskatta parametrna i $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ används *maximum likelihood*-metoden, vilket innebär att likelihoodfunktionen $\mathcal{L}(\boldsymbol{\beta})$ maximeras med avseende på parametrarna $\boldsymbol{\beta}$. Likelihoodfunktionen beskriver hur troligt det är att den observerade datan kommer från en given distribution med parametrarna $\boldsymbol{\beta}$ och genom att maximera funktion ges de mest troliga värdena $\hat{\boldsymbol{\beta}}$. Givet observerad data $\mathbf{x}_1, \dots, \mathbf{x}_N$ med tillhörande klasser $y_1, \dots, y_N$ är likelihoodfunktionen definierad som

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N} P(Y_i) = \prod_{i=1}^{N} \pi_i^{y_i} \left(1 - \pi_i\right)^{1-y_i}.$$

Ofta används logaritmen av likelihoodfunktionen

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\pi_i^{y_i} \left(1 - \pi_i\right)^{1-y_i}\right) \tag{3}$$

då denna ofta har maxima i samma punkt som likelihoodfunktionen men är enklare att optimera. För att skriva (3) som en funktion av parametrarna $\boldsymbol{\beta}$ används (1) och (2):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log\left(\pi_i^{y_i} \left(1 - \pi_i\right)^{1-y_i}\right) = \sum_{i=1}^{N} y_i \log \pi_i + (1 - y_i) \log\left(1 - \pi_i\right)$$

$$= \sum_{i=1}^{N} y_i \log \pi_i - y_i \log(1 - \pi_i) + \log(1 - \pi_i) = \sum_{i=1}^{N} y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log\left(1 - \pi_i\right)$$

$$= \sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij}\right) - \log\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij}\right)\right)$$

där vi i den sista likheten har använt att $(1 - \pi_i) = \frac{1}{1+\exp(\beta_0 + \sum \beta x)}$. För att maximera $\ell(\boldsymbol{\beta})$ tas

gradienten fram och sätts lika med noll,

$$\mathbf{0} = \nabla(\ell) = \begin{cases} \sum_{i=1}^{N} y_i - \frac{e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}} \\ \sum_{i=1}^{N} y_i x_{i1} - \frac{x_{i1}e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}} \\ \ddots \\ \sum_{i=1}^{N} y_i x_{ip-1} - \frac{x_{ip-1}e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}} \end{cases},$$

vilket ger k stycken ickelinjära ekvationer, lika många som antalet kovariater i modellen. Låt varje kvot $\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_i x_{ij}}} = p(y_i = 1|\chi = \boldsymbol{x_{i\cdot}}, \boldsymbol{\beta})$. I matrisform kan $\nabla(\ell)$ skrivas som:

$$\nabla(\ell) = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

För att bestämma $\hat{\boldsymbol{\beta}}$ används ofta Newton–Raphson algoritmen för vilken en iteration kan uttryckas

$$\boldsymbol{\beta}^{\text{ new}} = \boldsymbol{\beta}^{\text{ old}} - \left(\nabla^2(\ell)\right)^{-1}\nabla(\ell)$$
$$\nabla^2(\ell) = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

Där $\nabla^2(\ell)$ är hessianmatrisen och $\boldsymbol{W} = \boldsymbol{p(1-p)}$, och kallas för viktmatrisen [10]. Vi får:

$$\boldsymbol{\beta}^{\text{ new}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

Där $\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$. Som begynnelsevärde brukar $\boldsymbol{\beta} = \mathbf{0}$ användas. Algoritmen konvergerar eftersom log-likelihood funktionen är konkav.

Med de erhållna värdena på $\hat{\boldsymbol{\beta}}$ kan alltså modellen estimera sannolikheten $P(y = 1 \mid \chi = \mathbf{x})$. För att klassificera observationer till en distinkt grupp, 0 eller 1, kan ett tröskelvärde p^* användas. Klassificeringsmodellen kan då skrivas som

$$f(\mathbf{x}) = h(g(\mathbf{x}), p^*) = \begin{cases} 0 & \text{om} \quad g(\mathbf{x}) < p^* \\ 1 & \quad\quad \text{annars} \end{cases} \tag{4}$$

där $g(\mathbf{x})$ betecknar modellen för logistisk regression vilken ger den estimerade sannolikheten för att observationen kommer från en neuropatisk. $h(\mathbf{x}, p^*)$ betecknar klassificeringen som görs vilket är beroende av värdet på p^*. Tröskelvärdet p^* kan alltså ses som en ytterligare parameter till modellen.

3.3 Uppskattning av modellfel

För att undersöka hur väl en tränad modell med logistisk regression $g_{\mathcal{T}}(\mathbf{x})$ fungerar krävs mått som kan användas för att dels jämföra modeller emellan samt beskriva modellens prediktionsförmåga. Dessa mått kan beskrivas som estimat av en funktion

$$\text{Err} = L(Y, g_{\mathcal{T}}(\chi)) \tag{5}$$

där L kan väljas till olika funktioner och $g_{\mathcal{T}}(\chi)$ är en modell vars parametrar har skattats med $n(\mathcal{T})$ antal observationer som finns i mängden $\mathcal{T} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_{n(\mathcal{T})}, \mathbf{x}_{n(\mathcal{T})})\}$. Mängden $\mathcal{T}$ benämns här som träningsmängd. L brukar väljas så att ett litet värde betyder att $g_{\mathcal{T}}$ är en bra prediktionsmodell för distributionen.

3.3.1 Tränings- och valideringsmängd

När parametrarna $\boldsymbol{\beta}$ hos modellen $g_{\mathcal{T}}(\chi)$ skattas, när modellen tränas, maximeras likelihoodfunktionen i ekvation (3) med hjälp av träningsmängden. Parametervärdena $\hat{\boldsymbol{\beta}}$ är alltså beroende av observationerna i träningsmängden, men det är inte garanterat att dessa observationer representerar den sanna distributionen. Modellfelet i ekvation (5) syftar till att beskriva felet av en godtycklig dragning från den sanna distributionen, oberoende av vilka observationer som ingått i träningsmängden.

Det finns olika tillvägagångssätt att beräkna ett väntevärdesriktigt estimat av ekvation (5). Om observationer $(y_i, \mathbf{x}_i) \in \mathcal{T}$ används för att estimera Err kommer de i de flesta fall att ge ett bias och det verkliga felet underskattas[10]. Detta eftersom modellen är optimerad med avseende på träningsmängden. Om det finns mycket data är uppdelning av observationer till träningsmängd och mängden som används för felestimering, valideringsmängd, en vanlig metod. Det innebär att observationer antingen tillhör träningsmängden $\mathcal{T}$ och används för parameterskattning, eller att de tillhör valideringsmängden för att estimera felet. Modellfelet estimeras då genom

$$\widehat{\text{Err}} = \hat{L}(\mathbf{y}, g_{\mathcal{T}}(\mathbf{X})) \text{ med } \mathbf{X} = (\mathbf{x}_{n(\mathcal{T})+1}, ..., \mathbf{x}_N)^T \text{ och } \mathbf{y} = (y_{n(\mathcal{T})+1}, ..., y_N)^T$$

Med en slumpvis vald tränings- och valideringsmängd blir då observationer som används för träning och validering oberoende av varandra.

3.3.2 Receiver Operating Characteristic (ROC) och Area Under Curve (AUC)

En modell med logistisk regression som fungerar bra kommer att ge små sannolikheter för observationer från friska personer ($y_i = 0$) och stora sannolikheter för observationer från neuropatiska personer ($y_i = 1$). Genom att välja

$$L(Y, g_{\mathcal{T}}(\chi)) = 1 - P(g_{\mathcal{T}}(\chi_a) < g_{\mathcal{T}}(\chi_b) \mid Y_a = 0, Y_b = 1) = P(g_{\mathcal{T}}(\chi_a) \geq g_{\mathcal{T}}(\chi_b) \mid Y_a = 0, Y_b = 1) \tag{6}$$

där (χ_a, Y_a) och (χ_b, Y_b) är två oberoende dragningar från distributionen, uppskattas modellens förmåga att separera klasserna [11]. En fördel med detta mått på modellfel är att det beskriver hur väl modellen klassificerar utan att p^* behöver bestämmas på förhand. Därför användes det både för att jämföra modeller sinsemellan samt för att utvärdera klassificeringsförmågan hos enskilda modeller. Detta gjordes med hjälp av arean under kurvan, förkortat AUC, för en så kallad *Receiver Operating Characteristic* (ROC).

För att förenkla beskrivningen av klassificeringsresultat kan en förvirringsmatris användas, vars utseende visas i tabell 3. Där kan man se modellens klassificeringar i relation till dess korrekta respons samt hur många gånger varje möjligt utfall inträffat. Till exempel ges värdet för false positive, FP, av indikatorfunktionen $1(f(\mathbf{x}_i) = 1, y_i = 0)$ som beskriver hur många observationer som klassificerats som neuropatiska där patienten faktiskt varit frisk.

		Korrekt respons	
		Positiv	Negativ
Predikerad respons	Positiv	True Positive (TP) $1(f(\mathbf{x}_i) = 1, y_i = 1)$	False Positive (FP) $1(f(\mathbf{x}_i) = 1, y_i = 0)$
	Negativ	False Negative (FN) $1(f(\mathbf{x}_i) = 0, y_i = 1)$	True Negative (TN) $1(f(\mathbf{x}_i) = 0, y_i = 0)$

Tabell 3: Schematisk förvirringsmatris.

Låt TPR och FPR stå för *True Positive Rate* samt *False Positive Rate*. TPR beskriver andelen

observationer från neuropaatiska patienter som modellen klassificerat korrekt

$$TPR = \frac{TP}{TP + FN},$$

medan FPR beskriver andelen observationer från friska patienter som modellen klassificerat som neuropatiska

$$FPR = \frac{FP}{FP + TN}.$$

En bra klassificeringsmodell kännetecknas av hög TPR och låg FPR.

Recieving Operator Characteristic-kurvan, ofta förkortad ROC-kurva är en välanvänd metod för att utvärdera tröskelvärdesberoende binär klassificering [12]. Metoden baseras på att se hur TPR förändras för olika nivåer av FPR. För att åstadkomma detta ses TPR och FPR som funktioner av tröskelvärdet p^*;

$$TPR(p^*) = \frac{\mathbf{y}^T \cdot \hat{\mathbf{y}}}{\|\mathbf{y}\|} = \frac{\mathbf{y}^T \cdot \boldsymbol{h}(\boldsymbol{g}(\mathbf{X}), p^*)}{\|\mathbf{y}\|} \qquad \text{samt}$$

$$FPR(p^*) = \frac{(\mathbf{I}_{Nx1} - \mathbf{y})^T \cdot \hat{\mathbf{y}}}{\| (\mathbf{I}_{Nx1} - \mathbf{y}) \|} = \frac{(\mathbf{I}_{Nx1} - \mathbf{y}) \cdot \boldsymbol{h}(\boldsymbol{g}(\mathbf{X}), p^*)}{\| (\mathbf{I}_{Nx1} - \mathbf{y}) \|},$$

där $\boldsymbol{g}(\mathbf{X})$ är resultatet från vår logistiska regressionsmodell och $\mathbf{h}(\mathbf{g}(\mathbf{X}), p^*)$ är klassificeringen gjord med tröskelvärdet p^* givet realiseringarna i $\mathbf{X}$ och motsvarande responsvärden i $\mathbf{y}$.

För något bestämt värde av respektive funktion ger inverserna $TPR^{-1}(p^*)$ och $FPR^{-1}(p^*)$ ett tröskelvärde p^* vilket gör det möjligt att uttrycka den ena funktionen som en funktion av den andra. Genom att anta ett bestämt värde s på FPR kan då TPR uttryckas som

$$ROC(s) = TPR\left(FPR^{-1}(s)\right), \quad s \in [0, 1].$$

ROC-kurvan kan användas för att visualisera hur väl klassificeringen lyckas separera klasserna med olika tröskelvärden p^* [10] och arean under denna kurva, AUC (*Area Under Curve*), kan användas till att estimera Err i ekvation (6). Modeller utvärderas alltså med hjälp av

$$\widehat{\text{Err}} = \hat{L} = 1 - \text{AUC} = 1 - \int_0^1 ROC(s)ds \tag{7}$$

som antar värden mellan 0 och 1. Vid $\widehat{Err} = 0.5$ eller större är klassificeringen inte tillförlitlig alls, eftersom slumpmässiga gissningar kommer ge lika bra eller bättre resultat. Ju närmare $\widehat{Err}$ är 1, desto tillförlitligare är klassificeringen[12]. För att ekvation (7) skulle ge ett väntevärdesriktigt estimat av ekvation (6) användes enbart valideringsmängden för att ta fram $ROC(s)$ samt dess motsvarande AUC.

3.3.3 Akaike information criterion (AIC)

Det finns metoder för att uppskatta modellfelet Err utan att använda valideringsmängd. Det krävs då ett mått på hur mycket information som går förlorad då $g(X)$ används för att approximera Y. En funktion som kan väljas som mått på modellfel är då

$$L(Y, g_{\mathcal{T}}(\chi)) = -2E_{\mathcal{T}}E_a\left[\log P\left(Y_a | g_{\mathcal{T}}(\chi_a)\right)\right] \tag{8}$$

[13], där (χ_a, Y_a) är en observation från stickprovet a, vilket är oberoende av träningsmängden $\mathcal{T}$. Akaike visade 1973 att det maximerade likelihood-värdet var en approximation av ekvation 8 med en bias ungefär lika stor som antalet kovariater k [13]. Detta gav upphov Akaike Information Criterion (AIC)[14] och skrivs:

$$\text{AIC} = 2k - 2\ell.$$

där ℓ är log likelihood-funktionen för modellen som ges i ekvation (3) anpassad på träningsmängden.

För att AIC skall estimera ekvation (8) perfekt krävs ett oändligt stort stickprov. Detta innebär att AIC är en bra approximation om N är stort relativt k. Om ett litet stickprov används ($\frac{N}{k} < 40$)[13] bör istället AIC_c användas. AIC_c är AIC med en andra ordningens korrekturterm för liten stickprovsstorlek och lyder

$$\widehat{\text{Err}} = \text{AIC}_c = 2k - 2\ell + \frac{2k(k+1)}{N-k-1}$$

där k är antalet kovariater för den modell som har flest kovariater[13]. Då $max(\frac{N}{k}) < 27$ för projektet har AIC_c använts för att jämföra modeller med olika antal kovariater.

3.4 Standardisering

Standardisering görs för att träningen av klassificeringsmodeller inte skall påverkas av kovariaters olika skalning. Om kovariater har väldigt olika medelvärden och varians kan detta annars fördröja eller helt förhindra att modellens uppskattning av koefficienter konvergerar. Detta på grund av att många optimeringsalgoritmer, så som Newton-Raphsonalgoritmen[10], utforskar sin kostnadsfunktion stegvis med en konstant steglängd för alla variabler.

Låt $\mathbf{X}_{*,i}$ vara en kolumnvektor med värden för en kovariat i för alla givna observationer $\mathbf{X}$. Normalt sätt utförs standardisering genom att data transformeras med formeln

$$\frac{\mathbf{X}_{*,i} - \overline{\mathbf{X}}_{*,i}}{s(\mathbf{X}_{*,i})},$$

där $\overline{\mathbf{X}}_{*,i}$ är kovariatvärdenas stickprovsmedelvärde och $s(\mathbf{X}_{*,i})$ dess stickprovsstandardavikelsen. Den transformerade datan har då medelvärde 0 och standardavikelse 1.

Om ett dataset har kraftigt avvikande värden, s.k. outerliers, kan dessa komma att påverka standardiseringen då både $\overline{\mathbf{X}}_{*,i}$ och $s(\mathbf{X}_{*,i})$ är känsliga för kraftigt avvikande värden. Eftersom det finns outliers i vår data för ett antal kovariater (se figur 2) gjordes standardisering enligt

$$\frac{\mathbf{X}_{*,i} - m(\mathbf{X}_{*,i})}{MAD(\mathbf{X}_{*,i})}$$

där $m(\mathbf{X}_{*,i})$ är medianen och $MAD(\mathbf{X}_{*,i})$ står för *Median Absolute Deviation* som uttrycks

$$MAD(\mathbf{X}_{*,i}) = median\left(\mathbf{X}_{*,i} - m(\mathbf{X}_{*,i})\right).$$

3.5 Dimensionsreducering

Då antalet kovariater k är många kan det finnas anledning att försöka välja bort vissa, eller på andra sätt minska antalet parametrar som skall estimeras. De två främsta anledningarna är att öka modellens prestation genom att undvika överträning och att göra modellen mer tolkbar [14].

Om det finns ett logistiskt samband mellan $\mathbf{X}$ och $\mathbf{y}$ och antalet observationer $N >> k$ så kommer modellens uppskattade parametrar ha både lågt bias och låg varians. Om däremot N inte är mycket större än k så kommer de skattade parametrarnas varians att öka och modellen överträna, vilket leder till att den inte presterar väl på ny data. Då $N < k$ är klassificeringsproblemet linjärt obestämt och det saknas då en bästa uppskattning vilket även innebär att parametrarnas varians blir oändlig[14]. Antalet observationer är för vår data alltid större än antalet kovariater, men förhållandet kan ändå förbättras så att $N >> k$ genom dimensionsreducering. Detta gäller speciellt då t.ex. bara mätningar från fot används, eftersom N är mindre i ett sådant fall.

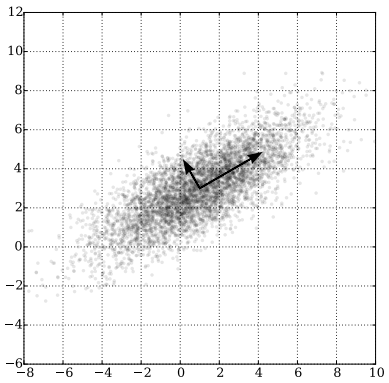
Genom att minska antalet kovariater kan alltså variansen av de skattade parametrarna minskas. Problemet är att om alla kovariater innehåller unik och relevant information för klassificeringen

kommer en reducering av antalet dimensioner öka modellens bias. Detta är vad som kallas för *the bias variance tradeoff* [10]. Utmaningen med dimensionsreducering handlar alltså om att, om möjligt, representera data med färre kovariater på ett sätt som samtidigt behåller det mesta av den relevanta informationen. De metoder vi använt för att reducera antalet dimensioner tas upp i detta kapitel.

3.5.1 Principalkomponentanalys (PCA)

Principalkomponentanalys (PCA) är en metod som ofta används då man har flera korrelerade kovariater, för att representera $\mathbf{X}$ av observationer som ett mindre antal linjära kombinationer av de givna kovariaterna[14]. Dessa linjära kombinationer, även kallade *principalkomponenter* väljs så att så mycket varians som möjligt behålls i $\mathbf{X}$ efter transformationen. Metoden reducerar alltså dimensionen av $\mathbf{X}$ utan att förhoppningsvis tappa relevant information vilket kan resultera i en bättre klassificering. De nya kovariater som returneras av metoden, principalkomponenterna, är även helt okorrelerade vilket också kan ge bättre resultat när klassificeringsmodeller ska anpassas till data $\mathbf{X}$.

Observationerna $\mathbf{x}_i$, $i \in [1, N]$ i $\mathbf{X}$ kan ses som ett punktmoln i $\mathbb{R}^k$. Principalkomponenterna är ett uppsättning ortogonala vektorer där den första komponenten motsvarar den riktning med störst spridning i $\mathbf{X}$, den andra komponenten motsvarar den ortogonala riktningen med näst störst spridning, och så vidare till den sista komponenten som motsvarar den ortogonala riktningen med minst spridning. Ett exempel på beräknade principalkomponenter för punktmoln $\mathbb{R}^2$ kan ses i figur 3.



Figur 3: Punktmoln i $\mathbb{R}^2$ med dess två principalkomponenter. Bild hämtad från Wikipedia. [1]

Kovariansen mellan alla kovariater kan sammanfattas i en $k \times k$-matris, kovariansmatrisen, här betecknad $\mathbf{\Sigma}$. Genom att se en observation som en vektor av stokastiska variabler, $\chi = (K_1, K_2, ..., K_k)^T$, är ett element i kovariansmatrisen

$$\mathbf{\Sigma}_{ij} = \mathrm{Cov}\left(K_i, K_j\right) = E\left[\left(K_i - \mu_{K_i}\right)\left(K_j - \mu_{K_j}\right)\right]$$

[15]. Med de realiseringar av värden på kovariater som finns i $\mathbf{X}$ estimeras elementen med

$$\mathbf{\Sigma}_{ij} = \frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{X}_{ni} - \overline{\mathbf{X}}_{*,i}\right)\left(\mathbf{X}_{nj} - \overline{\mathbf{X}}_{*,j}\right)$$

där $\overline{\mathbf{X}}_{*,i}$ motsvarar medelvärdet för kolumnen i i $\mathbf{X}$ och således ett estimat för $E\left[K_i\right]$. Egenvektorerna för $\mathbf{\Sigma}$, sorterade från största till minsta korresponderande egenvärde, motsvarar principalkomponenterna $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_k$ av $\mathbf{X}$ [16]. $\boldsymbol{z}_1$ är då den principalkomponent där dess riktning är den största variansen för observationerna i $\mathbf{X}$, $\boldsymbol{z}_2$ den näst största och så vidare.

Själva dimensionsreduceringen gjordes med ett underrum av principalkomponenterna. För att först hitta matrisen med samtliga k principalkomponenter, här kallad rotationsmatrisen $\mathbf{Z}$:

$$\mathbf{Z} = \left[z_1, z_2, ..., z_k\right],$$

beräknades den utifrån observationerna i träningsmängden. Samtliga observationer, i både tränings- och valideringsmängd transformerades sedan till matrisen $\mathbf{T}$ enligt

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{Z}.$$

Därefter tränades och validerades modeller med de i första komponenterna för $i \in (1, 2, ..., k)$. Detta innebar k iterationer där en kolumn i $\mathbf{T}$ användes i den första iterationen, två kolumner i den andra, tills alla kolumner användes i den sista iterationen. Det antalet principalkomponenter som gav det minsta modellfelet Err valdes som den bästa varianten.

[1] Licenserad med Creative Commons Attribution 4.0 International `https://creativecommons.org/licenses/by/4.0/`

3.5.2 Stepwise selection

Stepwise selection är ett iterativt tillvägagångssätt för att välja bort kovariater då en klassificeringsmetod tränas flera gånger och för varje gång tas antingen en kovariat bort eller läggs till. Det finns olika sätt att implementera metoden Vid *backwards stepwise selection* är alla kovariater med till en början och succesivt tas den kovariat bort som har minst påverkan på det uppskattade felet i modellen. I *forward stepwise selection* sker processen i omvänd ordning. Klassificeringsmodellen tränas först med alla kombinationer av endast en kovariat och behåller sedan den som har störst påverkan på felet. Detta upprepas för nästa kovariat och modellen lägger succesivt till den av de kvarvarande kovariaterna som har störst påverkan på felet.

Det går teoretiskt att testa alla möjliga kombinationer av kovariater och då skulle det också gå att hitta den optimala kovariatuppsättningen för klassificeringsmetoden med den givna datan, men ofta är det inte beräkningsmässigt praktiskt att söka igenom hela utfallsrummet av kovariatkombinationer. Stepwise selection är ett beräkningsmässigt effektivare sätt att hitta en tillräkligt bra kombination av kovariater.

Enbart backwards stepwise selection användes och implementerades enligt följande:

1. Från början har man den totala mängden av kovariaterna $A = \{v_1, \cdots, v_k\}$.
2. Err_{total} mäts (med AIC eller AUC) med samtliga kovariater.
3. En kovariat v_i väljas bort från kovariatmängden och felet Err i frånvaro av denna kovariaten mäts.
4. Steg 3 upprepas med alla kovariater $i = 1, \cdots, k$. Man sparar de k olika Err.
5. Välj indexet i för det minsta Err_i med $(i = 1, \cdots, k)$ och ta bor det motsvarande vs_i från mängden A, eftersom den tillför modellen med det minsta förbättringen.
6. Mängden A uppdateras till $A = \{v_1, \cdots, v_{i-1}, v_{i+1}, \cdots v_k\}$.
7. Stegen 2-6 upprepas tills det blir en kovariat kvar i mängden A.
8. Den modellen med det minsta Err_i där $i = 1, \cdots, k$ väljs som bästa modellen.

3.6 Implementation av modellval

För att kunna avgöra hur väl logistisk regression kunde separera gruppen neuropatiska från kontrollgruppen, dvs validera logistisk regression som modell, uppskattades modellfelet med hjälp av -AUC av ROC för valideringsmängden baserat på en modellparametrar skattade med träningsmängden. I logistisk regression behövs kovariatrummet först vara reducerat innan parametrar kan skattas, därför föregicks modellvalideringen av modellurval där en dimensionsreduceringsmetod användes för att hitta ett lämpligt kovariatrum.

3.6.1 Uppdelning till träning och validering

Responsen i observationer antas vara beroende av personen som mätningarna gjorts på. I vår data finns det flera observationer av samma personer. Därför skulle inte en slumpmässig uppdelning mellan tränings- och valideringsdata nödvändigtvis kunna betraktas som två oberoende datamängder. När observationerna delades upp till tränings- och valideringsmängd lades därför alla observationer från en person i samma grupp.

Om medelvärdet för responsvariabeln Y är lika i både tränings- och valideringsmängd innehåller de båda mängderna samma andel friska och sjuka vilket tenderar att ge lägre varians i valideringsresultatet och modellanpassningen blir bättre än om andelen friska respektive sjuka skiljer sig kraftigt mellan grupperna [17]. För att åstadkomma detta gjordes en stratifierad uppdelning där observationer från kontrollgruppen och de bekräftat neuropatiska delades upp i valideringsoch träningsmängd separat. Eftersom vi dessutom ville undvika ett beroende mellan tränings- och

valideringsmängd gjordes uppdelningen med avseende på antal personer istället för antal observationer. Konsekvensen av detta blev att stickprovsmedelvärdena av Y i de olika grupperna inte blev fullständigt identiska då antal observationer per person skiljer sig något.

De observationer som användes för validering var från personer i kontrollgruppen och gruppen av bekräftat neuropatiska. Observationer från de obekräftat neuropatiska användes aldrig som valideringsmängd, men i vissa fall i träningsmängden.

3.6.2 Korsvalidering

Ett problem med uppdelningen till träning- och valideringsmängder för denna data är att antalet observationer är starkt begränsat, särskilt för bekräftat neuropatiska personer. Det innebär att risken är stor för att få neuropatiska personer hamnar i valideringsmängden, vilket i sin tur leder till att estimaten av modellfelet får hög varians och blir opålitliga. Därför användes korsvalidering för att dela upp data till träning och validering, vilket möjliggör att alla observationer kan användas till validering[10].

Tillvägagångssättet för korsvalidering är att först fördela alla observationer till P olika delmängder som vi kallar mappar. En observation förekommer endast i en av mapparna, och antalet observationer i de olika mapparna är ungefär samma. Träningen och valideringen itereras sedan P gånger. I varje iteration utgör en av mapparna valideringsmängden, medan resterande mappar utgör träningsmängden. En mapp utgör valideringsmängd endast i en av iterationerna [10]. I figur 4 visas en iteration av korsvalidering med $P = 4$, när andra mappen användes som valideringsmängd och observationer i resterande mappar används som träningsmängd.

Figur 4: De olika mapparna som används för träning och validering, vid den andra iterationen. Varje ruta innefattar en mapp med observationer. Träningsmängden utgörs av alla observationer som finns i gula mappar, medan valideringsmängden enbart innehåller observationer från den andra mappen.

Uppdelningen till mapparna gjordes enligt 3.6.1, där antalet personer i en mapp från en klass motsvarade $1/P$ av de olika personerna från klassen. Eftersom alla personer inte stod för exakt lika många observationer blev storlekarna på de olika mapparna något varierande.

Det har visats att om P väljs till ett värde mellan 2 och 5 får valideringsresultaten en högre varians än om P väljs till 10 [17]. Samtidigt blir variansen högre då P väljs för stort, i extremfallet $P = N$ blir variansen högre än om $P = 5$[18]. Därför är 5 eller 10 vanliga val av P.

Eftersom antalet observationer från de bekräftat neuropatiska är mycket få användes $P = 4$. För vissa fall, t.ex. då bara observationer uppmätta på fot undersöktes, hade ett större värde på P inneburit att valideringsmängden ibland bara innehållit endast en observation från klassen neuropatiska. Precis som i fallet $P = N$ skulle då troligen variansen av resultatet bli stor även för $P \geq 5$.

3.6.3 Yttre korsvalidering

Implementationen med korsvalidering syftade till att uppskatta felet av en modell med parametrar och dimensionsreducering. För att hitta vilken dimensionsreducering som bäst passade träningsmängden behövdes även då modellfel uppskattas. Därför delades implementationen upp till två valideringar: en yttre validering när modellfelet hos modeller som genomgått dimensionsreducering testades, samt en inre validering som returnerade vilken dimensionsreducering som skulle användas i den yttre valideringen.

Den yttre korsvalideringen användes för att undersöka hur bra logistisk regression som modell kunde tillämpas för att skilja klasserna. För varje iteration av korsvalideringen beräknades parametrarna till en modell baserat på ett kovariatrum bestämt av en inre korsvalidering på träningsmängden. Innan modellen tränades standardiserades tränings- och valideringsmängderna enligt 3.4 med värden på median och MAD från träningsmängden. Modellparametrar skattades sedan med hjälp av träningsdata, och validering genomfördes genom att applicera denna tränade modell på den standardiserade valideringsmängden.

Med

$$\boldsymbol{g} = \begin{bmatrix} g_{\kappa(1)} \\ \vdots \\ g_{\kappa(i)} \\ \vdots \\ g_{\kappa(N)} \end{bmatrix}$$

där $\kappa(i)$ motsvarar mappen som observation i tillhör och $g_{\kappa(i)}$ är den modellen som har anpassats av träningsmängden (dvs alla mappar förutom $\kappa(i)$) anges felet för modellerna $\boldsymbol{g}$ som bestäms genom korsvalideringen benämns här som $\mathrm{CV}(\boldsymbol{g})$.

Efter att alla iterationer av den yttre korsvalideringen körts användes alla modeller framtagna genom inre korsvalidering genom

$$\mathrm{CV}(\boldsymbol{g}) = 1 - \mathrm{AUC} \tag{9}$$

AIC_c användes inte här eftersom måttet mer beskriver hur väl modeller passar observationerna snarare än att beskriva hur väl klassificeringen med modellen blev.

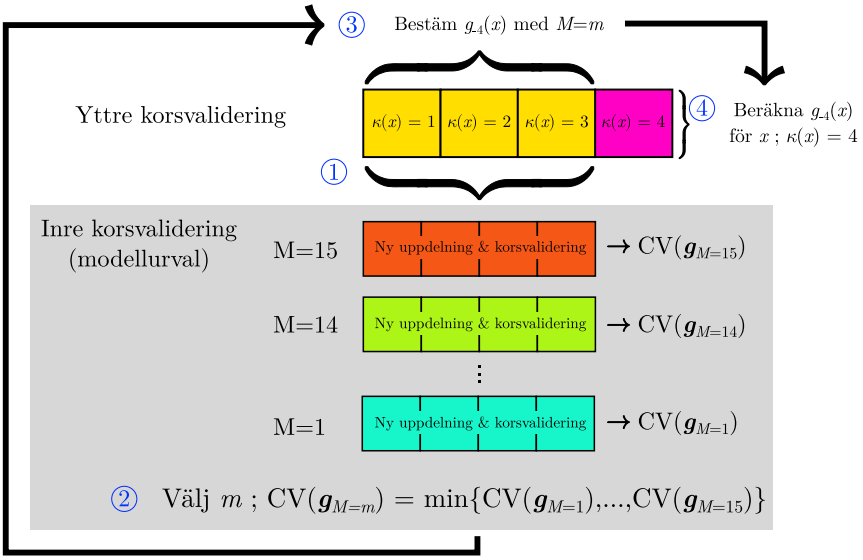
3.6.4 Modellurval i inre validering

För att hitta ett lämpligt kovariatrum baserat på träningsmängden användes inre korsvalideringen där dimensionreduceringsmetoderna Backward Stepwise Selection(3.5.2) eller PCA(3.5.1) användes. Backwise Stepward Selection gav en mängd med kovariater medan PCA gav ett antal principalkomponenter.

I varje varv av någon av dimensionreduceringsmetoderna gjordes en ny uppdelning av träningsmängd till fyra mappar och fyra valideringsiterationer genomfördes. Modellfelet CV uppskattades antingen som i (9) eller med

$$\mathrm{CV}(g) = \mathrm{AIC_c}\,.$$

Här testade endast en modellvariant g, eftersom $\mathrm{AIC_c}$ beräknas utan korsvalidering. Den kovariatmängden respektive de m första principalkomponenterna, beroende på dimensionsreduceringsmetod, som gav minst värde för något av dessa mått användes sedan i yttre korsvalidering. Inre korsvalidering med PCA exemplifieras i figur 5.



Figur 5: Inre korsvalidering då PCA används som dimensionsreducering och observationer $x; \kappa(x) = 4$ används som valideringsmängd. **1.** Träningsmängd, $x; \kappa(x) \neq 4$, används till modellurval. **2.** Det antal komponenter m som ger lägst fel i inre korsvalidering väljs och skickas till yttre korsvalidering. **3.** Parametrar för g_{-4} skattas med hjälp av samma data som använts i inre korsvalidering och antal komponenter m från inre korsvalidering. **4.** Modellfel av g_{-4} uppskattas med hjälp av valideringsmängd.

3.7 Undersökning av variationer i implementation

Metoden för klassificering med logistisk regression testades i flera variationer med avseende på fyra olika aspekter:

1. Vilka observationer som användes beroende på vilken kroppsdel mätningarna är gjorda på: vad, fot eller dessa kombinerade.
2. Med eller utan observationer från neuropatiska personer vars diagnos inte fastställts i träningsmängd.
3. Metod för dimensionsreducering: ingen, PCA eller Backward Stepwise Selection
4. Metod för modelljämförelse i inre korsvalidering när någon dimensionsreduceringsmetod användes: korsvalidering och AUC eller AIC_c

Alla möjliga kombinationer provades, vilket totalt uppgår till 30 stycken.

3.7.1 Observationer grupperat per kroppsdel

Modeller med observationer från tre olika mängder baserat på vilken kroppsdel de uppmätts på testades. De tre observationsmängderna var observationer från enbart vad, observationer från enbart fot eller observationer från båda. För var och en av mängderna användes samtliga observationer i mängden för både modellträning och validering.

3.7.2 Obekräftat neuropatiska i träningsmängden

Samtidigt som det fanns väldigt få observationer från neuropatiska personer som fått en bekräftad läkardiagnos var dessa observationer ofta lika de neuropatiska som inte hade en bekräftad läkardiagnos. De som inte hade en bekräftad diagnos ansågs för osäkra för att uppskatta modellfel med, men eftersom de var så pass lika de bekräftat neuropatiska testades modeller som tränats med de bekräftat neuropatiska i träningsmängden. För dessa modeller blev antalet observationer från neuropatiska i träningsmängden markant fler. För exempelvis fot består då träningsmängden av drygt 30 observationer istället för knappt 7, vilket troligen fångar mer av distributionen.

3.7.3 Undersökning av modellvarianter

För att kunna testa de 30 olika kombinationerna av tillvägagångssätt upprepades den yttre korsvalideringen 30 gånger för varje kombination. Upprepningarna gjorde att varje resultat kunde representeras med en distribution, vilket bättre representerar modellens verkliga prestation [17]. Översiktlig pseudokod av hur resultat genererades syns i figur 6.

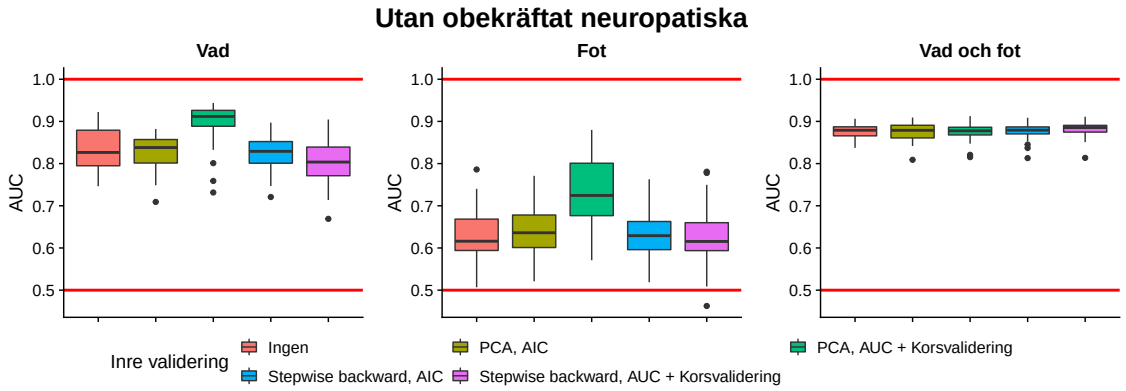
```
för körning i 1:30
   för kroppsdel i [fot, vad, båda]
      för träningsmängd i [med observationer från neuropatiska utan diagnos,
                           utan observationer från neuropatiska utan diagnos]

        korsvalidera modell med kroppsdel & träningsmängd utan dimensionsreducering
        beräkna och spara AUC

        för metod i [PCA, Stepwise backward selection]
           för mått i [AICc, a]
              korsvalidera modell med kroppsdel, träningsmängd, metod & mått
              beräkna och spara AUC
```

Figur 6: Pseudokod för resultatgenerering av de olika kombinationerna av modellvarianter som testades

4 Resultat

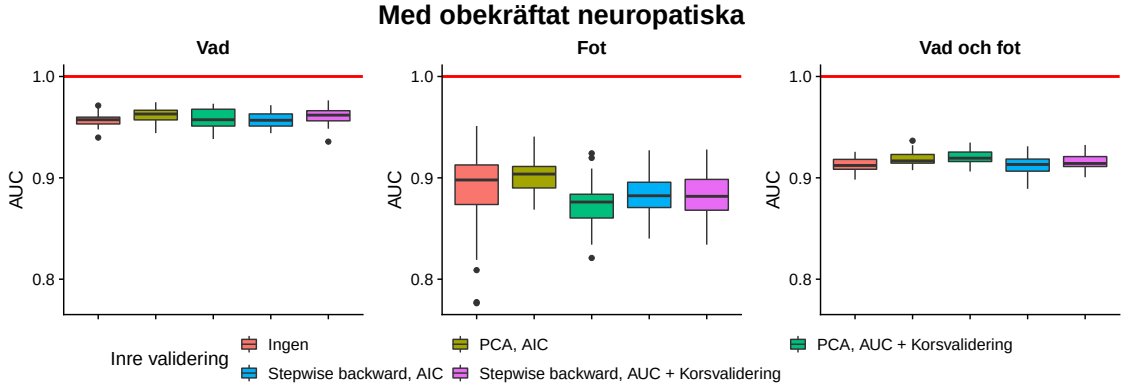
Med hjälp av de upprepade yttre korsvalideringarna som kördes enligt 3.7 undersökte vi först vilken dimensionsreduceringsmetod och uppskattning av modellfel som gav bäst resultat. Vi tittade separat på de modeller som tränades utan respektive med observationer från neuropatiska utan bekräftad diagnos. Överlag var prestationen lika mellan olika dimensionsreduceringar, men PCA med AUC som modellfelsuppskattning var något bättre för de utan obekräftade neuropatiska i träningsmängden. De dimensionsreduceringar och modellfelsuppskattningar som gav bäst för respektive kroppsdel användes för att undersöka hur de olika kroppsdelarna presterade mot varandra och hur resultatet förändrades för olika värden av tröskelvärdet p^* (se ekvation (4)). Här var det tydligt att vad var den kroppsdel som presterade bäst.

Att använda PCA som dimensionsreduceringsmetod och korsvalidering för uppskattning av modellfel gav högst AUC bland de modeller som tränades med observationer från friska och neuropatiska med bekräftad diagnos för vad respektive fots. Detta kan ses i figur 7. Att modellen gav högst AUC innebär att den också gav lägst $\widehat{\text{Err}}$ enligt ekvation (7). PCA med AUC gav ett markant högre medianvärde jämfört med de andra varianterna som testades (inget modellurval, stepwise backward med AICc eller AUC och PCA med AICc) för både vad (0.91) och fot (0.72). För modeller tränade med observationer uppmätta på både vad och fot presterade samtliga varianter av modellurval väldigt lika med en liten varians och medianvärden strax under 0.9.



Figur 7: Lådogram av AUC för olika varianter av modellurval för modeller tränade med friska och neuropatiska med bekräftad diagnos. I vänster delfigur användes enbart observationer från vad, i mittersta delfiguren endast fot och i höger delfigur användes både vad och fot. Varje lådogram visar resultaten av 30 yttre korsvalideringar. De röda stecken visar AUC = 1.0 vilket innebär en perfekt anpassning och AUC = 0.5 vilket motsvarar slumpvis gissning.

Samma jämförelse gjordes för modeller som tränades med även de obekräftat neuropatiska och resultaten blev något annorlunda, vilket åskådliggörs i figur 8. Samtliga varianter gav högre medianvärden av AUC, men prestationen de olika varianterna emellan var mer lika varandra för samtliga kroppsdelar. För vad respektive fot gav PCA med AIC högst medianvärde, 0.96 för vad och 0.90 för fot. För modeller med både vad och fot gav PCA med AUC det högsta medianvärdet 0.92.



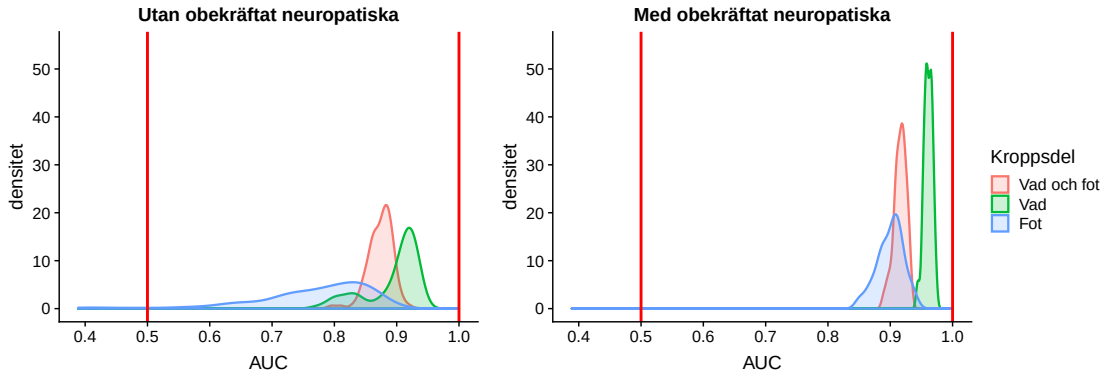
Figur 8: Lådogram av AUC för olika varianter av modellurval för modeller tränade med friska och neuropatiska med bekräftad diagnos. I vänster delfigur användes enbart observationer från vad, i mittersta delfiguren endast fot och i höger delfigur användes både vad oc fot. Varje lådogram visar resultaten av 30 yttre korsvalideringar.

En översikt med resultat från både figur 8 och figur 7 kan ses i figur 11 i Bilaga A.

De varianter som gav bäst resultat för respektive kroppsdel och utan eller med de obekräftade neuropatiska i träningsmängden användes för att visa hur de olika kroppsdelarna presterar, vilket syns i figur 9. För både modeller tränade utan eller med observationer från de obekräftat neuropatiska gav vad högst medelvärde följt av vad och fot tillsammans och till sist av fot.

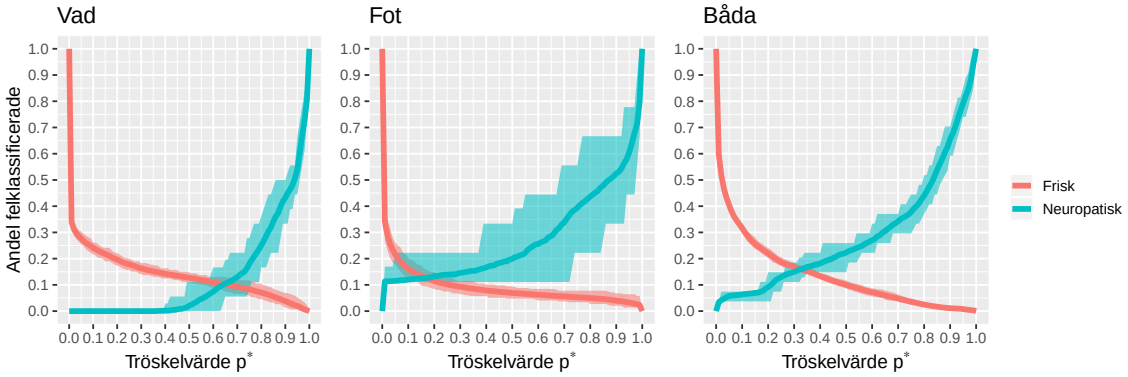
Medelvärden var utan obekräftat neuropatiska 0.89 för vad, 0.77 för fot samt 0.87 för vad och fot. Standardavvikelserna var 0.04 för vad, 0.09 för fot och 0.02 för vad och fot.

För modeller tränade med obekräftade neuropatiska var medelvärden 0.96 för vad, 0.90 för fot och 0.92 för vad och fot. Standardavvikelser var 0.01 för vad, 0.02 för fot och 0.01 för vad och fot.



Figur 9: Uppskattning av sannolikhetstäthet (densitet i figur) av AUC för klassificeringar i grupper om kroppsdelar. Till vänster har enbart friska och neuropatiska med bekräftad diagnos ingått i träningsmängden, medan i den högra har även observationer från neuropatiska utan bekräftad diagnos använts i träningsmängd. Inre validering gjordes med PCA och AUC, undantaget fallen vad respektive fot med obekräftat neuropatiska då PCA och AIC har användes. Varje täthetsuppskattning (en färg i en delfigur) beräknades med gaussisk kernel från AUC-värden från 100 yttre korsvalideringar.

Eftersom modellerna som tränades med obekräftat neuropatiska presterade bättre för samtliga kroppsdelar användes dessa för att se hur resultaten varierade med avseende på tröskelvärdet p^*. Detta tröskelvärde är då gränsen där värdet av sannolikheten från den logistiska regressionsmodellen $g(\mathbf{x})$ omvandlas till en av de två klasserna enligt ekvation (4). Det visade sig att FPR konsekvent gav lika resultat mellan olika körningar medan det hos TPR fanns en större varians. Detta visas i figur 10 där framförallt modellen gav ett brett spann av TPR för olika yttre korsvalideringar med fotobservationer men med samma tröskelvärde. Tröskelvärdet då andelen felklassificerade friska och neuropatiska var lika var för vad ungefär vid $p^* = 0.65$ med ca 10% felklassificering, för fot ungefär $p^* = 0.17$ med ca 13% felklassificering samt för vad och fot vid $p^* = 0.33$ med ca 18% felklassificering.



Figur 10: Klassificeringsresultat för olika tröskelvärden Andelen felklassificerade observationer på y-axeln och modellens tröskelvärde p^*, se (4), på x-axeln. I figuren längst till vänster har enbart mätningar från vad använts, i mittersta figuren har enbart mätningar från fot använts och i figuren till höger har alla observationer använts. I samtliga fall har neuropatiska utan bekräftad diagnos använts som träningsdata. Röd färg kommer från mätningar av friska personer (motsvarar FPR) och turkos färg från neuropatiska (motsvarar $1 - TPR$). De tjocka mörkare linjerna motsvarar medelvärdet från 100 olika körningar. Det ljusare området runtom linjerna indikerar det spann där resultat mellan 2.25-percentilen och 97.5-percentilen befanns.

5 Diskussion

Klassificering med logistisk regression fungerar bra för denna data, speciellt för observationer från vad med obekräftat neuropatiska i träningsmängden. Ett medelvärde på 0.96 på AUC och en standardavvikelse på 0.01 indikerar att modellen klassificerar väldigt nära den verkliga responsen.

5.1 Dimensionsreducering och uppskattning av modellfel

Dimensionsreducering, och då särskilt PCA, förbättrade klassificeringen för modeller tränade utan de obekräftat neuropatiska då enbart vad eller enbart fot användes. Det märktes inte lika tydligt, eller inte alls, när de obekräftade var med i träningsmängden eller då mätningar från både fot och vad användes. Detta är något i linje med idén kring varför dimensionsreducering bör genomföras, eftersom antalet observationer är fler men kovariaterna är lika många.

En tänkbar anledning till att PCA var den dimensionreduceringsmetod som fungerade bäst är att principalkomponenterna är okorrelerade, till skillnad från i Backward stepwise. Från ekvation 3.2 ser vi att skattningen på $\boldsymbol{\beta}$ innehåller termen $\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}$ vilket kan leda till numeriska problem om kovariaterna är starkt korrelerade, eftersom kolumnerna i $\boldsymbol{X}$ kan vara mycket nära linjärt beroende och inversen av matrisen är svår att beräkna. Detta kan vara fallet för vår data, vilket kan ses i korrelationsmatrisen i 1.

Troligtvis håller inte de asymptotiska antaganden som AICc bygger på i fallet med vår data på grund av att antalet observationer är för få. PCA med AICc ger en klart sämre dimensionsreducering än PCA med korsvalidering med AUC för vad respektive fot, vilket syns i figur 7. Eftersom AUC användes som valideringsmått för de slutgiltiga modellerna är det inte konstigt att dimensionsreducering med AUC fungerar bra. Däremot verkar inte motsvarande dimensionsreducering med AICc ge lika bra resultat, vilket tyder på att korsvalidering behövs i både yttre och inre validering.

Utifrån låddiagrammen i figure 7 och 8 som visar resultaten för de olika modellerna verkar det som att mängden träningsdata är vad som har störst påverkan på resultaten eftersom de olika dimensionsreduceringsmetoderna samt korsvalideringen knappt haft någon inverkan då vi använt de icke-bekräftade neuropatiska vid träning.

5.2 Val av kroppsdel för mätning

Resultatet från undersökningen tyder på att mätningar från fot försämrar prestation, medan mätningar från vad minskar modellfelet. Modeller med enbart mätningar från vad gav högst medelvärde av AUC för både träningsmängd utan och med obekräftat neuropatiska i figur 9. Det ska dock noteras att standardavvikelsen med träningsmängd utan obekräftat neuropatiska blev betydligt större än vad den blev med den utökade träningsmängden för modellerna med data från vader. Exempelvis blev tätheten kring AUC=0.8 och strax däröver större för vad än vad och fot. I fallet med de obekräftade neuropatiska fanns inte den effekten då vad tydligt presterade bäst av de tre observationsmängderna.

Modeller från mätningar med enbart fot fick en relativt hög standardavvikelse utan obekräftat neuropatiska i träningsmängden, men den minskade avsevärt för träningsmängd med de obekräftade neuropatiska. Även om de modeller tränade med obekräftat neuropatiska som gav högst AUC för fot blev större än den största för modeller som använde mätningar från både vad och fot blev medelvärdet för fot lägst av de tre observationsmängderna. Dock bör det tilläggas att medelvärdet på 0.90 indikerar att klassificering är fullt möjlig för fot också, och med tanke på den begränsade datamängd som användes går det inte att utesluta att fot ändå skulle vara bättre lämpad mätpunkt för svettest med kamera.

Att modeller med både vad och fot fick den minsta variansen för modeller tränade utan bekräftat neuropatiska kan förklaras med att datamängden är större. Samma mönster uppkom i skillnaden mellan modeller tränade utan och med obekräftat neuropatiska, där variansen minskade för samt-

liga fall när antalet observationer blev större. Även om variansen för vad och fot är mindre än enbart vad är det förväntade modellfelet större.

5.3 Hantering av tröskelvärde

Om modellen ska implementeras för att användas i sjukvården är en viktig frågeställning som behöver besvaras hur viktigt korrekt klassificering av en sjuk person (TP i förvirringsmatrisen, se tabell 3) är kontra korrekt klassificering av frisk person (TN i förvirringsmatrisen). Detta kan justeras enligt önskemål genom värdet för modellens tröskelvärde p^* i ekvation (4). Resultatet från den logistiska regressionsmodellen $g(\mathbf{x})$ är den bedömda sannolikheten att patienten lider av neuropati, och tröskelvärdet justerar för vilka sannolikheter patienter betraktas som friska respektive neuropatiska. Ett tänkbart sammanhang för användningen av modellen är att främst fånga upp många personer som eventuellt lider av sjukdomen och ett annat skulle vara att huvudsakligen filtrera ut personer som med stor säkerhet lider av sjukdomen. För det första sammanhanget skulle ett lågt värde av p^* väljas, medan den senare skulle kräva ett högre värde av p^*.

Andelen felklassificerade friska hamnar inom ett mycket snävt intervall för alla modeller i figur 10, men för neuropatiska är osäkerheten större. För vad är andelen felklassificerade neuropatiska, FNR, nära 0 från $p^* = 0$ ända till $p^* = 0.5$, samtidigt som andelen felklassificerade friska stadigt går ned med ett väldigt förutsägbart mönster. För värden av p^* större än skärningspunkten mellan de två kurvorna går FNR kraftigt upp. Det borde inte vara en önskad egenskap för modellen att ha en liten TPR, dvs andelen rättklassificerade neuropatiska, eftersom idén med testet är att det ska fånga upp neuropatiska i ett tidigt skeende av sjukdomsförloppet. Troligtvis finns ett rimligt val av tröskelvärde för denna modell mellan $p^* = 0$ och skärningspunkten $p^* \approx 0.65$. Det som behöver undersökas mer är hur stor andel friska som klassificerats som neuropatiska, FPR, kan tillåtas vara. Ett alltför stort värde på FPR kommer att leda till ett större behov av fortsatt vård och behandling, alltså mer kostnader, efter att testet genomförts.

5.4 Jämförelse med tidigare forskning

Precis som i Loavenbruck et al. [6] pekar våra resultat på att vaden lämpar sig bättre än fot för att separera friska och neuropatiska i kameratestet.

Deras undersökning använder total volym svett, intensitet av svettkörtlar i bild samt ökning av svettproduktion som kovariater - vilket det finns snarlika motsvarigheter för i vår data - men klassificering gjordes enskilt för varje kovariat utan någon klassificeringsmodell. Med vår metod får vi AUC på 0.96, vilket är bättre än AUC på 0.9 som de åstadkom.

Att vår modell ger ett högre AUC kan delvis bero på att vi använder flera kovariater i vår klassificering och på så vis fångar mer relevant information i vår modell. Loavenbruck et al. använder till exempel inte CI300 och Hazard Mode, eller någon motsvarighet som beskriver svettmönstrenas spatiella struktur.

En annan anledning som verkar spela stor roll är att vi, då vi använder observationerna från de obekräftade neuropatiska patienterna, har mer data i båda klasserna, och av den anledningen kan göra en bättre uppskattning av paramterarna i modellen. I Loavenbrucks klassificering användes observationer från 52 stycken friska och 20 neuropatiska patienter. Det går inte att avgöra hur många observationer som gjorts på varje patient men jämförelsevis innehåller vår data observationer från 120 stycken friska och 18 stycken bekräftat neuropatiska patienter. När vi använder de obekräftade neuropatiska tillkommer ytterligare 47 stycken personer till de 18 bekräftat neuropatiska. Detta visar även på att datamängderna som använts av oss och Loavenbruck et al. inte är exakt samma.

5.5 Framtida utveckling

När vi använt de obekräftat neuropatiska i träningsmängden har vi förmodligen i viss mån tränat med felklassifierad data. Med mer träningsdata från bekräftat neuropatiska är det därför mycket

troligt att modellen kan prestera ännu bättre. Mer data från bekräftat neuropatiska skulle även kunna ge en bättre estimering av modellfelet.

Vår ambition var att utforska hur väl logistisk regression med dess varianter kunde klassificera vår data. Några varianter av logistiska regressionsmetoder som hade varit intressanta att undersöka vidare är t.ex. Ridge-Regression och Lasso-Regression. Det finns även många andra klassificeringsmetoder som fungerar fundamentalt annorlunda och det är inte alls säkert att logistisk regresion är den bästa. En begränsning för mer komplexa metoder är dock att det ofta krävs mer data för att träna modellen.

För fortsatt undersökning kan det vara intressant att utforska mer noggrannt hur väl klassificeringen kan göras med färre kovariater och hur viktigt det är att ha mätningar från olika tidssteg. Om metoden ska användas i praktiken kan det vara en fördel om det räcker att kameran endast tar ett foto istället för att filma. Att variera antalet mappar i korsvalideringen kan också vara intressant att prova men detta tror vi har marginell påverkan på resultatet.

Vi implementerade förutom backward stepwise selection även forward stepwise selection men bedömde att resultaten var så pass lika att det endast var intressant att redovisa resultat för en av metoderna. Att uttömande prova alla kombinationer av kovariater är även en metod som inte bedömdes relevant då det finns orimligt många kombinationer för att de skall gå att testa alla.

6 Slutsatser

Det är tydligt att det går att klassificera huruvida en patient lider av perifer neuropati eller inte beroende på dess svettmönster. Det bästa resultatet från undersökningen kom från att använda de mätningar som gjordes på vaden. För observationer från vaden gavs med 171 observationer, i vårt fall 153 friska och 18 bekräftat neuropatiska och 45 obekräftat neuropatiska i träningsmängden en klassificering där AUC av ROC beräknas till 0.96 med en standardavvikelse på 0.01. Detta resultat innebär en mer korrekt klassificering än den som presenteras av Loavenbruck et al.[6]

Vid insamlande av mer data bör mätningar göras på patienters vader eftersom dessa mätningar visat sig skilja mer mellan friska och neuropatiska. Fler mätningar bör göras på bekräftat neuropatiska patienter men i brist på sådana mätningar har det varit gynnsamt att använda observationer från icke-bekräftat neuropatiska patienter tillsammans med bekräftat neuropatiska vid modellträning.

För att använda klassificeringsmodellen i praktiken behöver tröskelvärdet p^* bestämmas och vilket värde på p^* som är optimalt beror på hur man värderar att felaktigt klassificera en frisk som neuropatisk i förhållande till att felaktigt klassificera en neuropatisk som frisk. För att i praktiken göra en bra klassificering borde därför valet av p^* göras tillsammans med läkare eller neurologer som har kännedom av följderna för de två olika typerna av felklassificering. Med hjälp av resultaten från undersökningen kan en klar bild av hur valet av $p*$ påverkar risken för de olika typerna av felklassificering. Med denna till hälp borde en sakkunnig kunna göra ett bra val för modellen.

Oavsett vilket tröskelvärde som i praktiken är bäst så kan vi konstatera att vår modell i genomsnitt kan klassificera mellan 90-100% av de identifierade neuropatiska korrekt utan att andelen korrekt klassificerade friska understiger 85%. Då en av de vanligare diagnostiseringsmetoden i dagsläget, *Qualitative sudomotor axom reflex* (QSART) endast lyckas identifiera 75% av patienter som lider av neuropati[4] finns det motiv till att byta metod.

Referenser

[1] P. Saunders David K., "Neuralgia, neuritis, and neuropathy.", *Magill's Medical Guide (Online Edition)*, 2017. URL: http://proxy.lib.chalmers.se/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=86194344&site=eds-live&scope=site (hämtad 2019-05-16).

[2] A. J. Loavenbruck, J. S. Hodges, V. Provitera, M. Nolano, G. Wendelshafer-Crabb och W. R. Kennedy, "A device to measure secretion of individual sweat glands for diagnosis of peripheral neuropathy", *Journal of the Peripheral Nervous System*, årg. 22, nr 2, s. 139–148, juni 2017, ISSN: 1529-8027. DOI: 10.1111/jns.12212. URL: https://doi.org/10.1111/jns.12212 (hämtad 2019-05-16).

[3] C. H. Illigens Ben M. W. and Gibbons, "Sweat testing to evaluate autonomic function", *Clinical Autonomic Research*, årg. 19, nr 2, s. 79, juli 2008. DOI: 10.1007/s10286-008-0506-8. URL: https://doi.org/10.1007/s10286-008-0506-8 (hämtad 2019-05-16).

[4] V. A. Low, P. Sandroni, R. D. Fealey och P. A. Low, "Detection of small-fiber neuropathy by sudomotor testing", *Muscle & Nerve*, årg. 34, nr 1, s. 57–61, juli 2006, ISSN: 1097-4598. DOI: 10.1002/mus.20551. URL: https://doi.org/10.1002/mus.20551 (hämtad 2019-05-16).

[5] V. Provitera, M. Nolano, G. Caporaso, A. Stancanelli, L. Santoro och W. R. Kennedy, "Evaluation of sudomotor function in diabetes using the dynamic sweat test", *Neurology*, årg. 74, nr 1, s. 50–56, 2010, ISSN: 0028-3878. DOI: 10.1212/WNL.0b013e3181c7da4b. eprint: https://n.neurology.org/content/74/1/50.full.pdf. URL: https://n.neurology.org/content/74/1/50 (hämtad 2019-05-16).

[6] A. Loavenbruck, N. Sit, V. Provitera och W. Kennedy, "High-resolution axon reflex sweat testing for diagnosis of neuropathy", *Clinical Autonomic Research*, årg. 29, nr 1, s. 55–62, juli 2018. DOI: 10.1007/s10286-018-0546-7. URL: https://doi.org/10.1007/s10286-018-0546-7 (hämtad 2019-05-16).

[7] C. M. Bishop, *Pattern recognition and machine learning, 5th Edition*, ser. Information science and statistics. Springer, 2007, ISBN: 9780387310732. URL: http://www.worldcat.org/oclc/71008143 (hämtad 2019-05-16).

[8] T. T. Maalouf M Homouz D, "Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods", *Computational Intelligence*, årg. 134, s. 161–174, 2018.

[9] J. A. Rice, *Mathematical Statistics and Data Analysis.* Third. Belmont, CA: Duxbury Press., 2006.

[10] T. Hastie, R. Tibshirani och J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2. utg. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/ (hämtad 2019-05-16).

[11] J. A. Hanley och B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.", *Radiology*, årg. 143, nr 1, s. 29–36, 1982, PMID: 7063747. DOI: 10.1148/radiology.143.1.7063747. eprint: https://doi.org/10.1148/radiology.143.1.7063747. URL: https://doi.org/10.1148/radiology.143.1.7063747 (hämtad 2019-05-16).

[12] T. Gneiting och P. Vogel, "Receiver Operating Characteristic (ROC) Curves", *arXiv e-prints*, sept. 2018. arXiv: 1809.04808 [stat.ME]. (hämtad 2019-05-16).

[13] K. P. Burnham och D. R. Anderson, "Multimodel inference understanding aic and bic in model selection", *Sociological methods & research*, årg. 33, nr 2, s. 261–304, 2004.

[14] G. James, D. Witten, T. Hastie och R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014, ISBN: 9781461471370.

[15] K. I. Park, *Fundamentals of Probability and Stochastic Processes with Applications to Communications*, 1st. Springer Publishing Company, Incorporated, 2017, ISBN: 3319680749, 9783319680743.

[16] I. Jolliffe, *Principal Component Analysis.* Springer Verlag, 1986.

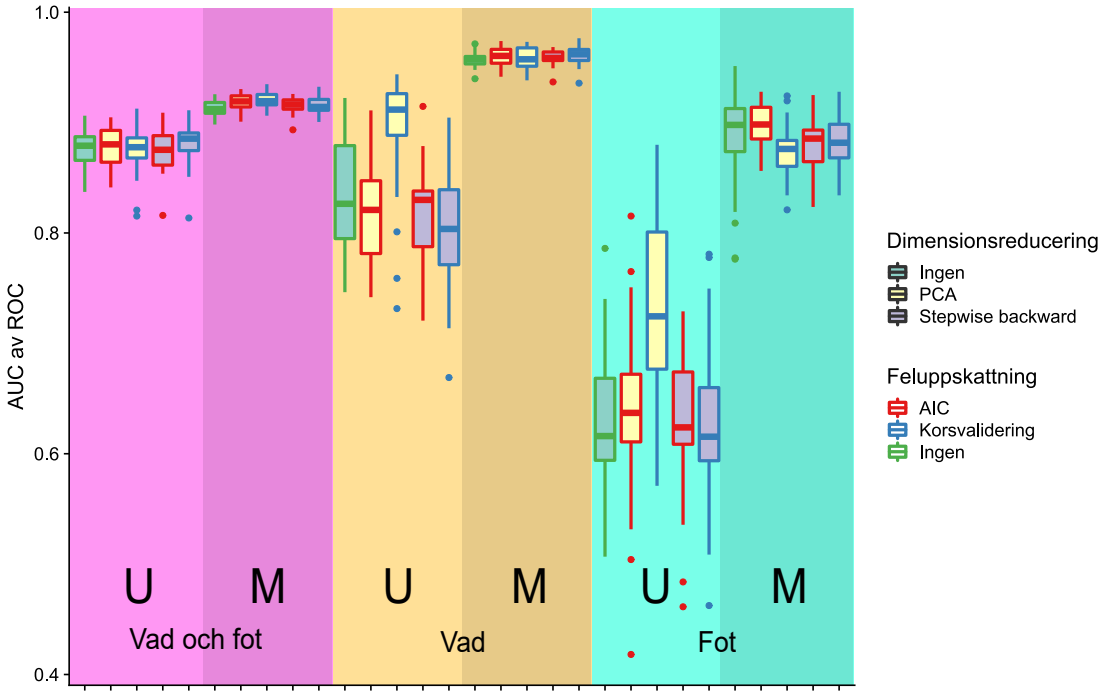
[17] R. Kohavi m. fl., "A study of cross-validation and bootstrap for accuracy estimation and model selection", i *Ijcai*, Stanford, CA, vol. 14, 1995, s. 1137–1145.

[18] L. Breiman och P. Spector, "Submodel selection and evaluation in regression. the x-random case", *International Statistical Review / Revue Internationale de Statistique*, årg. 60, dec. 1992. DOI: 10.2307/1403680.

[19] J. Illian, P. Penttinen, H. Stoyan och D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns*, ser. Statistics in Practice. Wiley, 2008, ISBN: 9780470725153. URL: https://books.google.se/books?id=%5C_U6BER2stYsC (hämtad 2019-05-16).

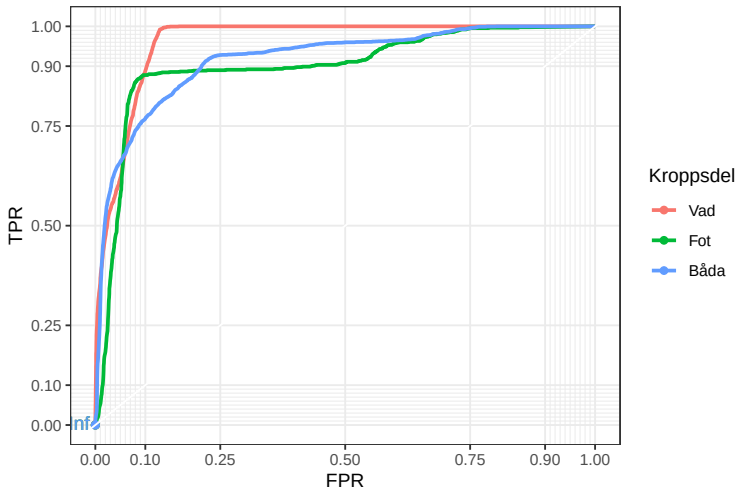
[20] P. Diggle, *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2013, ISBN: 9781466560246. URL: https://books.google.se/books?id=5FHSBQAAQBAJ (hämtad 2019-05-16).

Bilaga A Jämförelse mellan alla modellvariationer



Figur 11: AUC från ROC av 30 modellevalueringar med vardera kombination av: mätningar från både vad och fot (lila bakgrund till vänster), mätningar från endast vad (beige bakgrund i mitten) eller mätningar från endast fot (turkos del till höger); utan obekräftade neuropatiska i träningsdata (ljusare del till vänster av vardera läge, markerat med U) eller med obekräftade neuropatiska i träningsdata (mörkare del till höger av vardera läge, markerat med M); ingen dimensionsreduceringsmetod (grön fyllning av låda), PCA som dimensionsreduceringsmetod (gul fyllning av låda) eller Stepwise backward som dimensionsreduceringsmetod (blå fyllning av låda); ingen dimensionsreducering feluppskattningsmetod (grön kant på låda), AIC som feluppskattningsmetod (röd kant på låda) eller inre korsvalidering med modell som ger högst värde på AUC som feluppskattningsmetod (blå kant på låda)

Bilaga B ROC-kurvor för olika kroppsdelar



Figur 12: ROC av klassificering med logistisk regression. Röda kurva är mätningar från enbart vad, grön enbart fot och blå mätningar från både vad och fot. Yttre korsvalidering genomfördes 100 gånger för vardera kroppsdel och samtliga klassificeringsresultat från dessa utgjorde data för dessa kurvor. Träningsdata bestod av mätningar från friska, bekräftade neuropatiska och obekräftade neuropatiska. PCA med inre korsvalidering användes som dimensionsreducering.

Bilaga C Matematiska förklaring av kovariater

Av de kovariater som används för att klassificera patienter är CI300 och Hazard Mode baserade på mer avancerad matematik än de övriga. Nedan följer en beskrivning på hur dessa två har beräknats.

C.0.1 CI300

Klusterindexet CI300 utgår från K-funktionen (även kallad Ripley's K-funktion), som är ett vanligt mått för punktmönster inom spatial statistik. Låt λ vara intensiteten, det genomsnittliga antalet punkter per areaenhet i punktmönstret och $N(r, o)$ vara antalet punker inom radie r från punken o. $\lambda K(r)$ betecknar då det förväntade antalet andra punkter inom avståndet r från en godtycklig existerande punkt o[19][20]. Genom att dividera med intensiteten λ ges värdet på K-funktionen

$$\lambda K(r) = \mathbb{E}[N(r, o) \setminus \{o\}] \qquad \Longrightarrow \qquad K(r) = \lambda^{-1}\mathbb{E}[N(r, o) \setminus \{o\}].$$

Om n är det totala antalet punkter och $n_i(r) = N(r, x_i) \setminus \{x_i\}$, d.v.s. antalet andra punkter inom avståndet r från en given punkt x_i kan en enkel estimering av $\lambda K(r)$ göras genom att beräkna medelvärdet av $n_i(r)$ för alla punkter,

$$\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n} n_i(r)$$

[19][20]. Med detta medelvärde ges sedan en uppskattning av $K(r)$ med

$$\hat{K}(r) = \hat{\lambda}^{-1}\bar{n}(r).$$

I praktiken behöver kantkorrigering göras för att ta hänsyn till att det inte finns information om punkter utanför bilden. Detta har gjorts då $K(r)$ uppskattats för vår data och avståndet r har då mätts i antal pixlar.

K-funktionen i sin tur skrivs ofta om till den så kallade L-funktionen som brukar definieras som antingen

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \qquad \text{eller} \qquad L^*(r) = \sqrt{\frac{K(r)}{\pi}} - r.$$

L-funktionen innehåller samma information som K-funktionen men förekommer oftare då K-funktionen har visat sig ha en varians som ökar med radien r vilken stabiliseras för i L-funktionen. För homogena Poissonprocesser gäller även att $K(r) \propto r^2$ medan $L(r) \propto r$ samt att

$$L(r) = r \qquad \text{och} \qquad L^*(r) = 0 \qquad \text{för } r \geq 0$$

[19][20]. Ett värde på $L(r) > r$ och $L^*(r) > 0$ innebär att punkerna är klustrade och tenderar att ligga tätt inpå varnadra medan $L(r) < r$ och $L^*(r) < 0$ innebär att punkter tenderar att ligga långt ifrån varandra. Klusterindexet CI300 i vår data är definerat som

$$CI300 = \int_0^{300} L^*(r)dr$$

C.0.2 Hazard Mode

Hazard mode eller hazard rate är ett klassiskt mått inom överlevnadsanalys (Eng: *Survival Analysis*). Den spatiala statistikens motsvarighet *nearest-neighbour pair hazard rate*

$$h(r) = \frac{f(r)}{1 - F(r)}$$

kan tolkas som sannolikheten att en godtycklig punkts närmaste grannpunkt ligger på ett avståndet mellan r och $r + dr$ givet att vi redan sökt av det cirkulära området med radie r och inte stött på grannpunkten ännu [19]. $F(r)$ är den kumulativa fördelningsfunktionen för avståndet d till den närmsta grannpunkten o

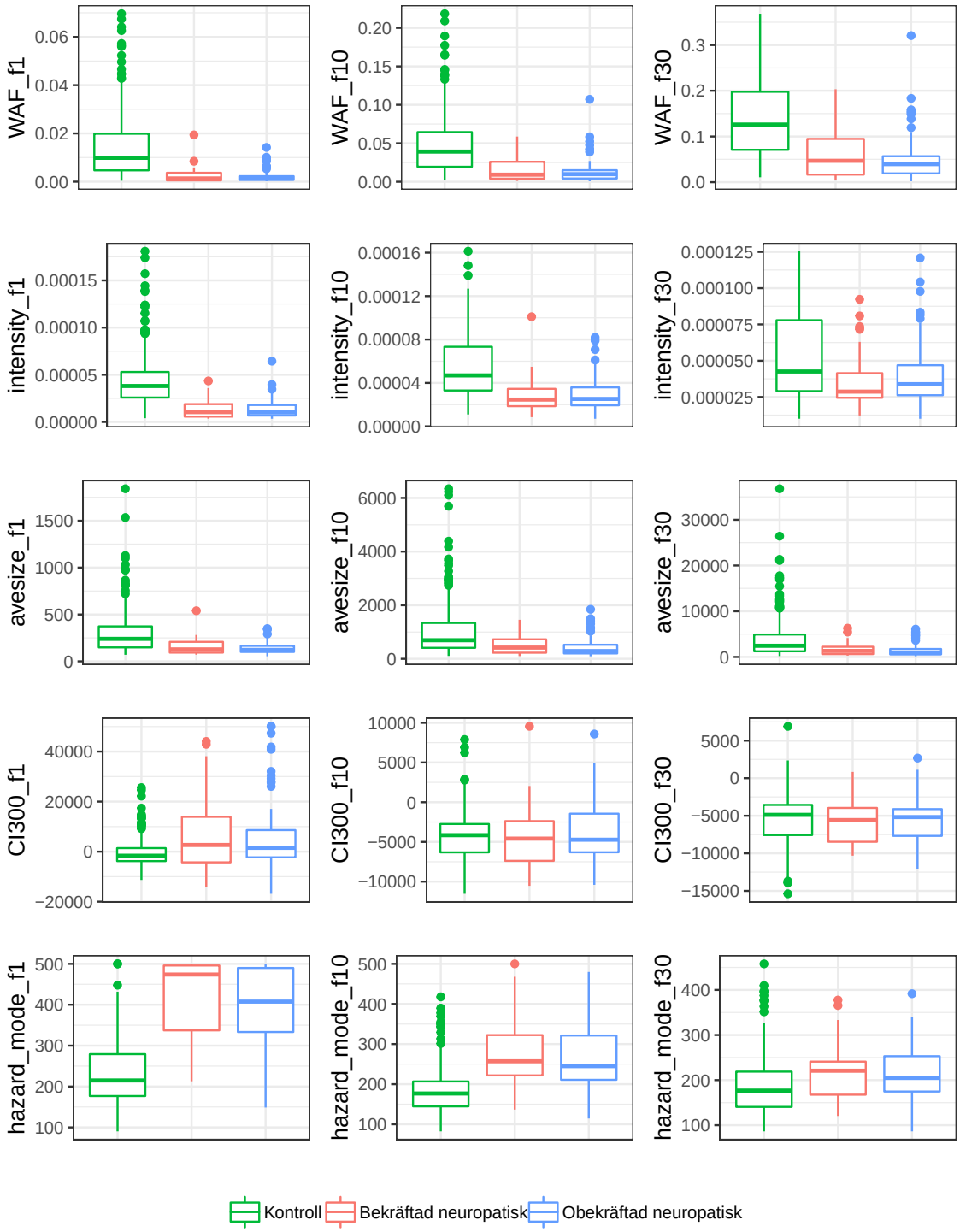
$$F(r) = P(0 \leq d(o) \leq r)$$

och $f(r) = F'(r)$ är motsvarande täthetsfunktion [19].

Värdena som givits i datan är för de avstånd r då hazard mode är som störst. Alltså för de avstånd det är mest troligt att den närmsta grannpunkten befinner sig på, i varje bild, givet att den närmsta grannen inte hittats på ett mindre avstånd.

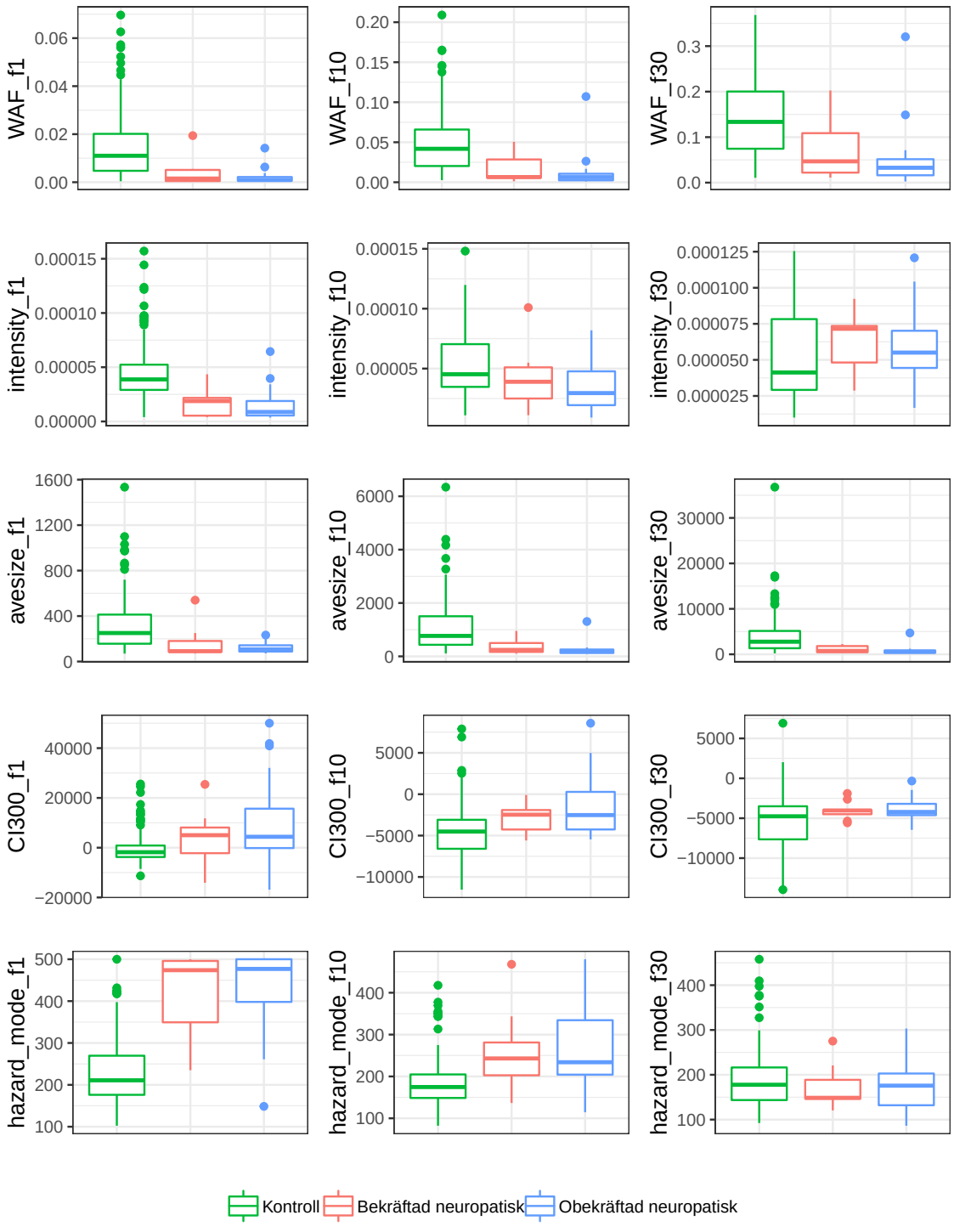
Bilaga D Spridningen av kovariaterna i de olika grupperna

D.1 Data från både fot och vad



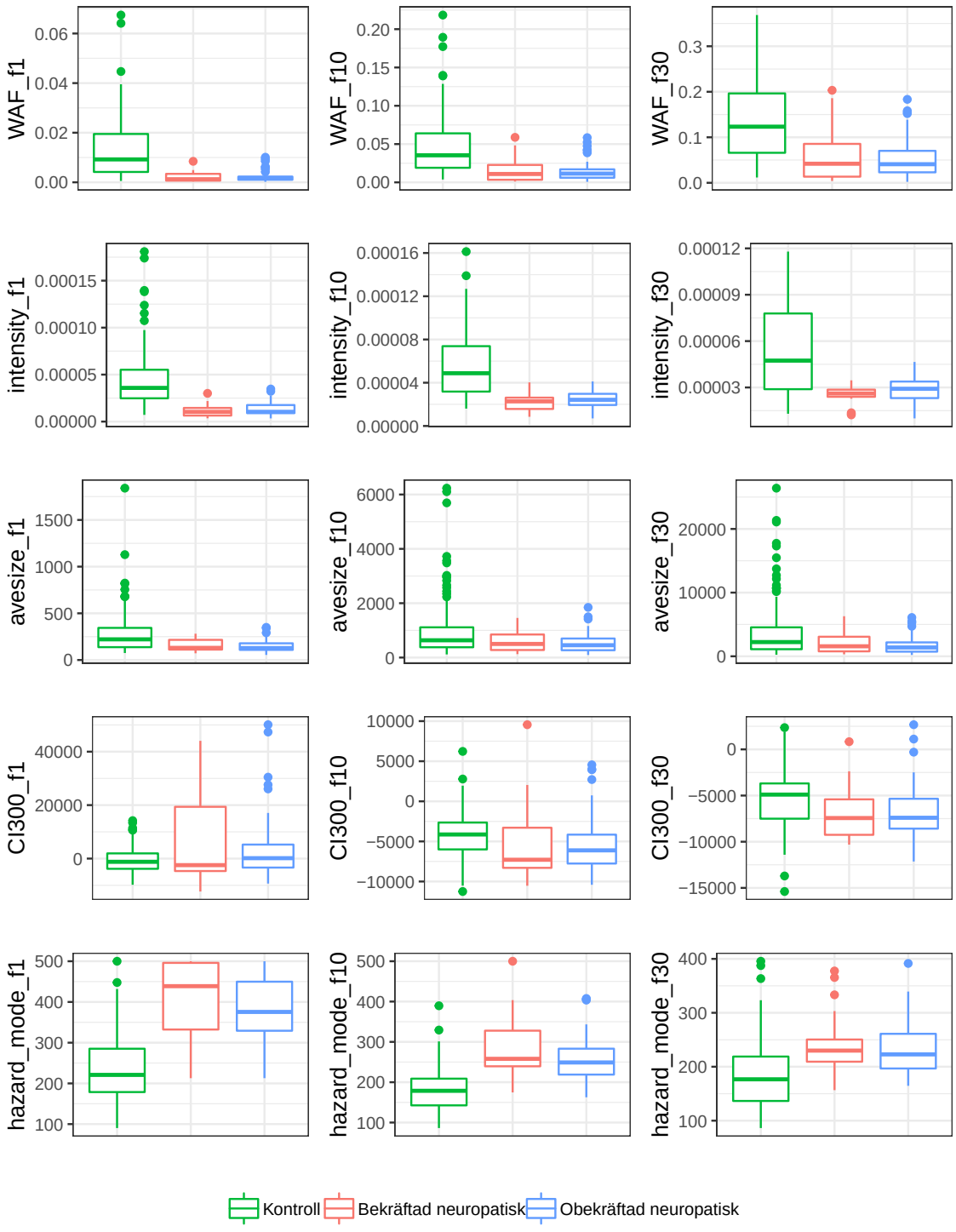
Figur 13: Låddigram som visar hur alla kovariaterna sprider sig för de olika patienter beroende på deras sjukdomstillstånd baserat på observationer från vad och fot. Varje rad av figuren visar spridningen av en kovariat i tre olika tid punkter.

D.2 Data från bara fot



Figur 14: Låddigram som visar hur alla kovariaterna sprider sig för de olika patienter beroende på deras sjukdomstillstånd baserat på bara observationer från fot. Varje rad av figuren visar spridningen av en kovariat i tre olika tid punkter.

D.3 Data från bara fot



Figur 15: Låddigram som visar hur alla kovariaterna sprider sig för de olika patienter beroende på deras sjukdomstillstånd baserat på bara observationer från vad. Varje rad av figuren visar spridningen av en kovariat i tre olika tid punkter.