CHALMERS

GÖTEBORGS UNIVERSITET

Prediktion av hårfärg och ögonfärg från genetiska markörer inom forensisk verksamhet

Prediction of Hair and Eye Color Using Genetic Markers in Forensic Work

Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Stella Brenden Linna
Sophie Dahlgren
David Huynh Thuan Duong
Felix Årebo Nettby

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2025

Prediktion av hårfärg och ögonfärg från genetiska markörer inom forensisk verksamhet

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers
David Huynh Thuan Duong Felix Årebo Nettby

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk fysik vid Chalmers
Stella Brenden Linna Sophie Dahlgren

Handledare: Petter Mostad

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2025

Förord

Vi vill börja med att tacka vår handledare Petter Mostad för hans engagemang och fantastiska stöd under hela processen med vårt kandidatarbete. Vi vill även tacka Rättsgenetik vid nationellt forensiskt centrum i Linköping och Andreas Tillmar för den data som användes i arbetet.

Under projektet har loggbok förts över de enskilda gruppmedlemmarnas prestationer. Denna loggbok bestod av en veckovis "dagbok" som inkluderade tidsloggar för varje gruppmedlem, där det noterades vad varje medlem jobbat med under veckan som gått. Nedan presenteras en tabell över de huvudsakliga författarna för varje avsnitt i kandidatrapporten. Alla gruppmedlemmar bidrog sedan till korrekturläsning och renskrivning av den färdiga rapporten. Under utvecklingsfasen av modellerna bidrog alla gruppmedlemmar med att skriva kod. Den prediktionsmodell som i slutändan användes skrevs av David, som vidareutvecklade modellen och producerade resultatet. Gruppen anser att alla medlemmar har bidragit likvärdigt till arbetet.

Bidragsrapport		
Avsnitt	Rubrik	Författare
	Förord	Stella
	Populärvetenskaplig presentation	Stella
	Sammandrag och abstract	Stella, Sophie
1	Inledning	Stella
1.1	Syfte	Felix, Sophie
2	Teori	Stella
2.1	Genetik	Stella
2.1.1	Genotyp och fenotyp	Stella
2.1.2	Genetisk analys inom forensik	Stella
2.1.3	Användning av DNA-teknik för att förutsäga ögon- och hårfärg	Stella
2.2	Statistisk modellering	Stella
2.2.1	Bayesiansk statistik	Stella, David
2.2.2	Logistisk regression	Stella, David
2.2.3	Trolighetsfunktion med logistisk regression	Stella, David
2.2.4	Multinomial logistisk regression	Stella, David
2.2.5	Markov Chain Monte Carlo	Stella
2.2.6	Den multinormala fördelningen	David
2.2.7	Centrala gränsvärdessatsen	Felix
2.2.8	Mätning av modellprestanda	David, Felix
3	Metod	Felix
3.1	Avgränsningar	Sophie, Felix
3.2	Datainsamling och visualisering av data	Sophie
3.3	Modellutveckling	Felix
3.4	Modellvalidering	Felix
4	Resultat	Alla
5	Diskussion	Alla
5.1	Tolkning av resultat	Alla
5.2	Begränsande faktorer och förbättringsmöjligheter	Alla
6	Samhälleliga och etiska aspekter	Felix, Stella
7	Slutsatser	Felix, Sophie
Bilaga A	Datavisualisering	Sophie, Felix
Bilaga B	Figurer och tabeller	Sophie, David
Bilaga C	Källkod	David, Felix

Figurer och tabeller i huvuddelen	
Nummer	Skapare
Tabell 1, 2, 4 och 6 Tabell 3 och 5	David Sophie
Fig: 1, 2, 5, 6 och 7 Fig: 3, 4	Sophie David

Figurer i bilagan	
Nummer	Skapare
Fig: 8 och 9	Sophie, Felix
Fig: 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 och 27	Sophie
Fig: 10, 11, 12, 13, 14, 15, 28, 29, 30, 31, 32, 33 och 34	David

Populärvetenskaplig presentation

Skulle det vara möjligt att endast utifrån DNA-spår på en brottsplats kunna skapa en perfekt fantombild av en gärningsman? För många utredare inom rättsväsendet låter detta som en utopi, en möjlighet att sätta ett ansikte på en gärningsman även då vittnen saknas. I nuläget är detta en ny teknik som är under utveckling, och det finns redan amerikanska företag som påstår sig behärska att producera sådana fantombilder. För att skapa en fantombild utan vittnen måste det finnas teknik som kan förutsäga visuella egenskaper hos en individ utifrån specifika DNA-sekvenser. Det är här matematiken kommer in, eftersom statistiska modeller kan byggas upp för att göra precis sådana förutsägelser. I det här arbetet har statistiska modeller använts för att, utifrån DNA-sekvenser, förutsäga vilken ögonfärg eller hårfärg en individ har. Dessutom har sannolikheten för att modellen förutspår rätt resultat beräknats, så att det är möjligt att avgöra huruvida resultaten är tillförlitliga att använda i praktiken.

DNA är den kod som bestämmer allt kring en människas uppbyggnad. I DNA-spiralen finns små sekvenser som kan vara olika uppbyggda mellan olika individer, och det är dessa som bidrar till att vi människor har unika egenskaper. Dessa sekvenser kan kallas för genetiska markörer. Det finns vissa specifika genetiska markörer som styr till exempel vilken ögonfärg en person har, och andra genetiska markörer som styr hårfärgen. Genom att använda sig av befintlig data över vilka genetiska markörer som bidrar till en viss egenskap kan en statistisk modell byggas upp, som sedan kan användas för att göra förutsägelser av exempelvis vilken ögonfärg en viss sekvens av genetiska markörer kan ge.

Modellen som har tagits fram bygger på logistisk regression, vilket är en statistisk modell som används för att modellera sannolikheter. Den enklaste logistiska regressionsmodellen har två möjliga utfall, till exempel då ögonfärg studeras kan den förutsäga att en individ antingen har blå ögon eller inte. Detta kallas för binär logistisk regression. Dessutom ger modellen sannolikheten för hur säker den är på att förutsägelsen stämmer. Modellen byggs upp från data över en mängd testpersoner, där det är givet vilken ögonfärg personen har och vilka genetiska markörer dessa kan vara kopplade till. Den logistiska modellen använder sedan informationen i datamängden för att hitta ett samband mellan blå ögonfärg och de genetiska markörerna. Därefter kan modellen ta emot genetisk information från en individ med okänd ögonfärg och beräkna sannolikheten för att denna individ har blå ögonfärg, eller inte blå ögonfärg.

Den logistiska modellen kan dessutom utvecklas för att hantera fler möjliga utfall än två, som till exempel då modellen ska förutspå de tre ögonfärgerna blå, brun och grön. En sådan modell kallas för multinomial logistisk regressionsmodell. Precis som tidigare lär modellen sig samband mellan genetiska markörer och ögonfärg, för att sedan göra en förutsägelse av vilken ögonfärg som en okänd individ har. Skillnaden är att modellen bestämmer vilken ögonfärg som förutsägs genom att beräkna sannolikheterna för att de genetiska markörerna ska ge blå, brun eller gröna ögon som utfall och använder sedan det mest sannolika utfallet som resultat. På det tekniska planet fungerar det även lite annorlunda, eftersom en multinomial logistisk regressionsmodell jämför varje kategori (till exempel brun och grön) med en referenskategori (blå). Sedan utför modellen en logistisk regression för varje jämförelse för att sedan slå ihop resultatet av dessa.

I arbetet har data använts som insamlades av Rättsgenetik vid nationellt forensiskt centrum i Linköping. Datamängden bestod av information över 85 testpersoners ögon- och hårfärg, samt data över individernas genetiska markörer. Genom denna data har logistiska modeller tagits fram för att göra förutsägelser av vilken ögon- och hårfärg en okänd person har utifrån dennes genetiska markörer. Modellen testades genom att jämföra resultatet av förutsägelsen med de faktiska ögon- och hårfärgerna för personen.

Förhoppningen med modellen är att den ska ge så pass tillförlitliga resultat att den i praktiken skulle vara möjlig att använda för att avgöra visuella aspekter som ögon- och hårfärg utifrån DNA. Det innebär att sannolikheten för att den gör rätt förutsägelser behöver vara väldigt hög, eller att det är tydligt till vilken grad modellen går att lita på genom att sannolikheterna tydligt presenteras. Detta är en teknik som är under utveckling i Sverige, med stor potential för användning i rättsväsendet. I framtiden kan därför liknande tekniker bli ett viktigt verktyg för brottsutredningar, och ett sätt att bidra till att bringa klarhet i fler ouppklarade fall.

Sammandrag

Ett utvecklingsområde inom forensiska verksamheter är DNA-fenotyping, vilket är en teknik för att utifrån DNA kunna predicera visuella egenskaper för en individ. I detta arbete undersöktes hur väl statistiska modeller kan utföra prediktion av ögon- och hårfärg utifrån genetiska markörer i DNA. De modeller som framtogs var logistiska regressionsmodeller som använde en MCMC-metod med en Metropolis-Hastings-algoritm för att uppskatta posteriorifördelningen. Två binära logistiska regressionsmodeller med två olika priorifördelningar jämfördes, en icke-proper likformig priorifunktion och en multinormal priorifunktion. På samma sätt jämfördes två multinomiala logistiska regressionsmodeller med samma två priorifördelningar.

Den datamängd som användes i arbetet omfattade ögon- och hårfärg för 85 individer och bestod av observerade färger samt genetiska markörer kopplade till färgerna. De ögonfärger som undersöktes var brun, blå och intermediär, medan de hårfärger som undersöktes var brun, blond, röd och svart.

Modellernas prestanda utvärderades genom ROC-grafer (eng: Receiver Operating Characteristic) och tillhörande AUC-värden (eng: Area Under the Curve). Resultatet uppvisade att modellerna överlag hade låga AUC-värden och därmed presterade dåligt. Ingen av modellerna lyckades uppnå ett totalt AUC-värde på över 0,75. Däremot presterade modellerna för prediktion av ögonfärg generellt bättre än modellerna för hårfärg. Det upptäcktes emellertid att modellen för multinomial hårfärg gav bra resultat när det gällde att predicera röd hårfärg med ett AUC-värde på 0,94. En stor begränsning för modellerna som kan ha påverkat resultatet är den begränsade datamängden. Finns det inte tillräckliga skillnader i den givna datamängden över genetiska markörer mellan de olika klasserna av färger får modellen svårt att kunna göra säkra prediktioner. Eftersom alla modeller har stora begränsningar är de i nuläget inte användbara för användning i praktiken, men genom fortsatt forskning skulle liknande statistiska modeller i framtiden kunna användas för mer tillförlitliga prediktioner.

Abstract

An emerging area within forensic sciences is DNA phenotyping, which is a technique used to predict an individual's attributes from their DNA. This thesis investigated how well statistical models could predict eye and hair color from genetic markers. The models developed were logistic regression models that utilized an MCMC method using the Metropolis-Hastings algorithm to estimate the posterior distribution. Two binary logistic regression models with different prior distributions were compared: one with an improper uniform prior and another with a multinormal prior. Similarly, two multinomial logistic regression models with the same two prior distributions were also evaluated.

The dataset used in this study included eye and hair color information from 85 individuals, consisting of observed phenotypes as well as their associated genetic markers. The eye colors investigated were blue, brown and intermediate, while the hair colors included brown, blonde, red and black.

The models' performance was evaluated using ROC curves (Reciever Operating Characteristic) and the corresponding AUC values (Area Under the Curve). The results showed that the models generally had low AUC values, and therefore performed unsatisfactorily. None of the models achieved a total AUC value over 0.75. However, the models that predicted eye color generally performed better than those predicting hair color. Interestingly, the multinomial hair color model was able to predict red hair with high accuracy, achieving an AUC value of 0.94. A key limitation of the models was the small dataset. If there are insufficient differences in the genetic marker data between the different color classes, the model struggles to make reliable predictions. Given these limitations, the models are currently not suitable for practical use. Nonetheless, with continued research, future statistical models of this kind could potentially provide more reliable predictions.

Innehåll

1 Inledning

Sedan 1989 har DNA-analys använts som ett verktyg inom forensiska verksamheter i Sverige. Ett område inom detta fält som fortfarande är under utveckling är Forensisk DNA fenotypning. Fenotyping innebär att delar av DNA används för att göra fenotypiska prediktioner, alltså förutsägelser av visuella egenskaper hos en individ. Detta kan vara exempelvis hårfärg och ögonfärg och kan vara värdefull information då ögonvittnen eller övervakningskameror saknas under en brottsutredning [1].

Denna typ av prediktion kan baseras på regressionsmodeller från bayesiansk statistik för att analysera sambandet mellan genetiska markörer i DNA och fenotypiska egenskaper. Genom statistiska metoder kan även osäkerheter i modellens prediktioner kvantifieras. En perfekt modell skulle kunna ta fram en färdig fantombild av en brottsmisstänkt mer effektivt och med mindre resurser än vad som krävs i nuläget. Genom DNA-fenotyping skulle alltså brottsutredningar kunna effektiviseras och därmed öka sannolikheten för att brottsmål uppklaras, vilket är av intresse för rättsväsendet. Samtidigt måste prediktionerna från modellen alltid vägas mot annan typ av bevisning för att undvika felaktiga slutsatser.

1.1 Syfte

Syftet med arbetet är att utveckla en prediktionsmodell som utifrån genetiska markörer kan förutspå karakteristiska egenskaper hos individer. I projektet utvecklas modellen med målet att förutspå genetiska ögon- och hårfärger. Modellen kommer att utvecklas med hjälp av bayesiansk statistik, och dess effektivitet och träffsäkerhet kommer att utvärderas för att bedöma om modellen kan tillämpas i praktiska sammanhang. Arbetet fokuserar på tre centrala aspekter: datamaterialets tillförlitlighet, modellens konstruktion samt utvärdering av dess prestanda.

2 Teori

I detta kapitel presenteras den teoretiska bakgrunden som ligger till grund för prediktionsmodellen. Till en början introduceras viktiga begrepp inom genetisk analys och vidare ges en introduktion till de metoder inom bayesiansk statistik och de regressionsmodeller som är relevanta för arbetet.

2.1 Genetik

Nedan introduceras centrala begrepp som används inom genetisk analys, såsom genotyp och fenotyp. Vidare presenteras hur DNA används inom forensiska undersökningar och den nya tekniken DNA-fenotyping.

2.1.1 Genotyp och fenotyp

Två begrepp som ofta används inom genetiken är genotyp och fenotyp. Genotyp beskriver en individs totala uppsättning gener, den arvsmassa (DNA) som individen ärvt från sina föräldrar. Fenotyp beskriver de fysiologiska egenskaper en individ har, såsom utseende och andra fysiska beteenden. Fenotypen för en individ formas utifrån genotypen men kan även till viss del påverkas av miljö [2]. På kemisk nivå är DNA uppbyggt av fyra sorters kvävebaser; adenin, guanin, cytosin och tymin. Dessa förkortas med A, G, C och T [3, s.54]. DNA har strukturen av en dubbel-helix, där de två strängarna med kvävebaser kopplas samman med vätebindningar. Dessa kopplingar bildar baspar av kvävebaser där A i den ena strängen alltid är kopplad samman med T i den andra, och på samma sätt är G alltid sammankopplad med C [3, s.57-59].

Ett sätt att beskriva skillnader i fenotyp för olika individer är genom analys av genomet. Polymorfi är en variation i DNA-sekvensen mellan två individer som kan förklara varför individerna har olika egenskaper. Den vanligaste formen av polymorfi är SNP, enbaspolymorfi (eng: Single Nucleotide Polymorphism), vilket är enskilda positioner i DNA-sekvensen som kan variera mellan olika individer. Detta sker genom att en enda kvävebas i sekvensen är utbytt mot en annan kvävebas [3, s. 686]. Det uppskattas att det i mänskligt genom finns ett SNP per tusen baspar, och mer än tre

miljoner enbaspolymorfier har hittills kartlagts. Omkring hälften av dessa kan vara direkt kopplade till fenotypiska egenskaper för en individ [4, s.16-17].

2.1.2 Genetisk analys inom forensik

DNA finns i flera typer av biologiska material, såsom blod, kroppsvätskor eller hudceller från fingeravtryck och svett [5, s.45]. Målet med analys av sådan typ av bevisning är bestämning av identitet för den person som lämnat DNA på en brottsplats. Historiskt har det varit möjligt att utföra analys av blodgrupp eller andra genetiska markörer såsom proteiner, men dessa kan skapa en profil som stämmer för ett flertal personer. Genom DNA-analys av biologisk bevisning kan istället en enskild person identifieras, eftersom alla individer (bortsett från enäggstvillingar) har en unik uppsättning gener [6, s.63]. Den kemiska stabiliteten i DNA är användbar för forensiker eftersom det är möjligt att analysera biologisk bevisning lång tid efter att brottet begåtts [5, s.45]. Sedan 1985 har det varit möjligt att göra DNA-analys på en liten mängd insamlat material genom PCR-tekniken (eng: Polymerase Chain Reaction). Tack vare PCR kan DNA från en enda cell kopieras till flera miljoner DNA-segment som sedan kan analyseras. PCR kan även genomföras under kort tid, och oftast uppnås en stor mängd DNA efter mindre än 24 timmar [6, s.64-67].

Ett relativt nytt område inom forensik är Forensisk DNA-fenotyping (eng: Forensic DNA Phenotyping, FDP). Detta kan användas när klassisk DNA-profilering inte kan hjälpa en brottsutredning då det inte finns någon misstänkt att matcha profilen mot, eller det inte går att hitta någon matchning med existerande DNA-databaser. Genom FDP kan prediktioner göras av synliga fenotypiska egenskaper från insamlat DNA, vilket kan smalna av potentiella misstänkta i ett brottsmål. FDP kan även användas för identifikation av till exempel försvunna personer, genom att skapa fantombilder utifrån DNA-spår [7].

2.1.3 Användning av DNA-teknik för att förutsäga ögon- och hårfärg

Forensisk DNA-fenotyping är en komplicerad teknik eftersom många gener har inverkan på de flesta synliga karaktärsdragen för en människa. I nuläget är de genetiska markörerna som styr kön, hårfärg och ögonfärg de mest tillförlitliga. Det pågår även forskning kring kartläggning av geners inblandning i bland annat ansiktsform, längd och ålder, men mer kunskap behövs kring både geners påverkan och miljö. Än så länge går det endast att förutsäga fenotypiska drag utifrån DNA med en viss sannolikhet, men trots detta kan FDP bidra till att leda brottsutredningar framåt [8].

Det finns däremot flera begränsande faktorer till prediktion av bland annat ögon- och hårfärg inom forensiskt arbete. UV-ljus, sjukdomar, droger och åldrande kan ha en påverkan på melaninets syntes. Till exempel kan barn med blont hår bli allt mer brunhåriga när de blir äldre. Ett ytterligare problem är att grön ögonfärg är svår att förutsäga, då det i dagsläget saknas väl kartlagda genetiska markörer för denna egenskap. Även yttre faktorer, såsom färgning av hår, kan påverka hur användbara fenotypiska prediktioner är i praktiken [8].

Hur tillförlitliga existerande metoder för fenotypiska prediktioner är har tidigare studerats i Sverige. I denna analyserades säkerheten för prediktion av 111 svenska individer genom användning av systemet ForenSeq och instrumentet MiSeq FGx (Verogen). De förutspådda ögon- och hårfärgerna som genom systemen gav störst sannolikhet jämfördes med de observerade färgerna. Resultatet av studien uppvisade att 80% av ögonfärgerna förutspåddes korrekt, men att systemet misslyckades med att förutsäga grön ögonfärg. För hårfärg lyckades systemet göra 58% korrekta prediktioner. När sedan en sannolikhetströskel på 0,7 infördes ökade korrekt predicerade ögonfärger till 85%, medan andelen korrekta hårfärger inte påverkades nämnvärt [9].

2.2 Statistisk modellering

För att göra prediktioner av fenotypiska egenskaper utifrån ett givet DNA kan statistiska modeller användas. En specifik modell som bygger på bayesiansk statistik är den logistiska regressionsmodellen. Den teoretiska bakgrunden till den logistiska regressionsmodellen för både binära och multipla utfall presenteras i detta kapitel. Dessutom presenteras MCMC-metoden (eng: Markov

Chain Monte Carlo) som använder Metropolis-Hastings-algoritmen vilket är en central komponent i metoden.

2.2.1 Bayesiansk statistik

Bayes sats är kärnan inom bayesiansk statistik och beskriver ett sätt att beräkna betingade sannolikheter. För en modell där observerad data beskrivs av den stokastiska variabeln y och där θ är en vektor av modellparametrar ger Bayes sats ett samband för posteriorifördelningen $\pi(\theta|y)$, vilket är den betingade sannolikheten för θ givet y, enligt

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \propto \pi(y|\theta)\pi(\theta). \tag{1}$$

Sannolikheten för observerad data givet parametrarna $\pi(y|\theta)$ kallas trolighetsfunktionen (eng: likelihood function) och $\pi(\theta)$ är priorifördelningen som beskriver tidigare kunskap kring θ. Nämnaren kan ses som en normaliseringskonstant, vilket innebär att ekvationen (1) kan skrivas $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$ [10, s.9].

Den betingade sannolikheten är definierad som

$$\pi(y|x) = \frac{\pi(x,y)}{\pi(x)}, \tag{2}$$

där $\pi(x,y)$ är den gemensamma sannolikhetsfördelningen för de stokastiska variablerna x och y. Marginalfördelningen $\pi(x)$ definieras som [10, s.20]

$$\pi(x) = \int \pi(x,y)dy, \tag{3}$$

där

$$\int \pi(x)dx = 1.$$

Från ekvation (2) och (3) kan sedan sannolikheten $\pi(y_{\text{new}}|y)$ beskrivas, där y_{new} är nya observationer som ska förutsägas och y är tidigare observerad data. Detta beskrivs av

$$\pi(y_{\text{new}}|y) = \int \pi(y_{\text{new}},\theta|y)d\theta = \int \pi(y_{\text{new}}|\theta)\pi(\theta|y)d\theta, \tag{4}$$

vilket är en integral som i praktiken är svår att beräkna exakt, och istället används ofta numeriska metoder såsom MCMC för att göra en uppskattning av denna.

Ett vanligt användningsområde för att modellera sannolikheter inom bayesiansk inferens är logistisk regression, vilken möjliggör direkt modellering av sannolikheten för ett givet utfall som en funktion av en uppsättning oberoende variabler. I följande avsnitt introduceras den logistiska regressionsmodellen för binärt utfall, som sedan generaliseras till en modell för multinomial logistisk regression där fler än två utfall kan hanteras.

2.2.2 Logistisk regression

Den logistiska funktionen

$$f(x) = \frac{e^x}{1+e^x}$$

lämpar sig väl för modellering av sannolikheter eftersom dess definitionsmängd är $(-\infty, \infty)$ och dess värdemängd ligger i intervallet (0,1). Den är utformad för att representera sannolikheter, vilka är ett tal mellan 0 och 1. Definitionsmängden innebär att funktionen kan ha alla reella värden som argument och samtidigt säkerställa att den ger ett definierat värde som utfall. Den logistiska funktionen kan därför generaliseras till att modellera sannolikheten för ett utfall baserat på en linjärkombination av flera oberoende variabler. Detta görs genom att definiera en vektor av data $\bar{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}^T$ där x_i är data som ska analyseras. Låt sedan $\theta = \begin{bmatrix} \alpha & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}$

vara en vektor av okända parametrar som ska uppskattas, där α är en konstantterm och β_i är regressionskoefficienter för $i \in \{1,2,\cdots,k\}$. Definiera sedan $z(\bar{X}|\theta) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Från detta kan den generaliserade logistiska funktionen sammanfattas som

$$f(\bar{X},\theta) = \frac{e^{z(\bar{X}|\theta)}}{1 + e^{z(\bar{X}|\theta)}},$$

vilken används för att modellera sannolikheten för ett binärt utfall [11, kap. 1]. De okända parametrarna behöver uppskattas utifrån observerad data, vilket kan göras genom konstruktion av en trolighetsfunktion.

2.2.3 Trolighetsfunktion med logistisk regression

Den logistiska regressionsmodellen bygger på att uppskatta de okända parametrarna i modellen, genom att observera den givna uppsättningen data. Den logistiska funktionen som introducerades i tidigare avsnitt har två möjliga utfall D som definieras $D = 0$ samt $D = 1$. Sannolikheterna för dessa utfall ges av den logistiska funktionen genom

$$\pi(D = 1|\theta, \bar{X}) = \frac{e^{z(\bar{X}|\theta)}}{1 + e^{z(\bar{X}|\theta)}}$$

och

$$\pi(D = 0|\theta, \bar{X}) = 1 - \pi(D = 1|\theta, \bar{X}).$$

Summan av sannolikheterna för de två utfallen är 1, vilket säkerställer att hela utfallsrummet täcks. Utifrån detta kan nu sannolikheten för att observera en given uppsättning data beräknas genom trolighetsfunktionen

$$\mathcal{L}(\theta|\bar{X}) = \prod_{i=1}^{n} \pi(D = 1|\theta, x_i)^{z_i} \pi(D = 0|\theta, x_i)^{1-z_i}, \tag{5}$$

där n är antalet datapunkter och varje observation x_i har ett tillhörande utfall $z_i \in \{0,1\}$, där $z_i = 0$ om datapunkt i har utfallet $D = 0$ och $z_i = 1$ om datapunkt i har utfallet $D = 1$. Trolighetsfunktionen är alltså en produkt av de individuella sannolikheterna för varje observation [12, kap. 1] och representerar således sannolikheten att observera den givna datamängden givet de valda parametrarna.

R använder sig av IEEE 754 binär flyttalsaritmetik. Denna begränsar den numeriska noggrannheten till ungefär 16 decimaler [13, s.753]. Vid beräkning av trolighetsfunktionen kan sannolikheterna bli väldigt små, särskilt när många observationer multipliceras. För att undvika numeriska fel är det därmed mer lämpligt att istället beräkna logaritmen av trolighetsfunktionen [14, s.30] genom

$$\ln(\mathcal{L}(\theta|\bar{X})) = \sum_{i=1}^{n} \left[z_i \ln\left(\pi(D = 1|\theta, x_i)\right) + (1 - z_i) \ln\left(\pi(D = 0|\theta, x_i)\right) \right]. \tag{6}$$

För att faktiskt göra en uppskattning av de okända parametrarna i den logistiska regressionsmodellen används en bayesiansk skattning med utgångspunkt i ML-metoden (eng: Maximum-Likelihood Estimation), vilket ger en skattning av parametrarna. En ML-skattning är de värdena på parametrarna som maximerar värdet på trolighetsfunktionen i ekvation (5). Med andra ord så maximerar ML-skattningen sannolikheten att observera datamängden. I praktiken är det lättare genom att hitta värdet som maximerar ekvation (6) [13, s.161].

2.2.4 Multinomial logistisk regression

I tidigare avsnitt har den logistiska regressionsmodellen för binära utfall presenterats. Denna modell kan generaliseras till att hantera fler än två diskreta utfall, vilket kallas för multinomial logistisk regression [12, s.35]. För att utföra multinomial regression utses en av utfallsvariablerna som en

referenskategori, vilken de andra variablerna jämförs med. Valet av referenskategori har inte någon påverkan på resultatet och kan därmed väljas godtyckligt [11, s.435]. I detta avsnitt kommer referenskategorin betecknas kategori 0.

Modellen för multinomial logistisk regression fungerar i princip på samma sätt som logistisk regression för det binära fallet, med skillnaden att det är en vektor av sannolikheter som beräknas. För en uppsättning av k oberoende variabler och s utfall inklusive en referenskategori, kan vektorn av data $\bar{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \end{bmatrix}^T$ definieras. Vektorn av konstanttermer ges av $\bar{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{bmatrix}^T$, och regressionskoefficienterna $\bar{\beta}$ beskrivs av en matris med dimension $(s-1) \times k$ enligt

$$\bar{\beta} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \beta_{21} & \ddots & & \beta_{2k} \\ \vdots & & \ddots & \vdots \\ \beta_{(s-1)1} & \beta_{(s-1)2} & \cdots & \beta_{(s-1)k} \end{bmatrix}.$$

Parametrarna $\theta = [\bar{\alpha}, \bar{\beta}_1^T, \ldots, \bar{\beta}_{s-1}^T]$ är på samma sätt som tidigare okända parametrar som ska uppskattas, nu för $i \in \{1,2,\cdots,s-1\}, j \in \{1,2,\cdots,k\}$. För att förenkla notationen framöver definieras sedan $\bar{h} = \bar{\beta}\bar{X} + \bar{\alpha} = \begin{bmatrix} h_1(\bar{X}) & h_2(\bar{X}) & \cdots & h_{s-1}(\bar{X}) \end{bmatrix}^T$.

Därefter definieras sannolikhetsfördelningarna för det stokastiska utfallet D givet data och modellparametrarna enligt

$$\mathbb{P}(D|\bar{X}) = \begin{bmatrix} \pi(D=0|\bar{X},\bar{\alpha},\bar{\beta}) \\ \pi(D=1|\bar{X},\bar{\alpha},\bar{\beta}) \\ \vdots \\ \pi(D=s-1|\bar{X},\bar{\alpha},\bar{\beta}) \end{bmatrix} = \frac{1}{1+\sum_{i=1}^{s-1}\exp\left(h_i(\bar{X})\right)} \begin{bmatrix} 1 \\ \exp(h_1(\bar{X})) \\ \vdots \\ \exp(h_{s-1}(\bar{X})) \end{bmatrix}.$$

Här ges sannolikheten för att observationen tillhör respektive kategori $i = 0, 1, \ldots, s-1$, där alla sannolikheter är positiva och summeras till 1. Givet n oberoende observationer kan den fullständiga trolighetsfunktionen för modellen uttryckas som

$$\mathcal{L}(\bar{\alpha}, \bar{\beta}|\bar{X}) = \prod_{m=1}^{n}\prod_{i=0}^{s-1} \pi(D=i|\bar{X}_m,\bar{\alpha},\bar{\beta})^{z_{im}}, \tag{7}$$

där $\bar{X}_m$ är de observerade variablerna för datapunkt m och indikatorvariabeln z_{im} definieras

$$z_{im} = \begin{cases} 1, & \text{om } m \text{ tillhör kategori } i \\ 0, & \text{annars.} \end{cases}$$

De okända parametrarna $\bar{\alpha}$ och $\bar{\beta}$ uppskattas därefter genom att maximera $\mathcal{L}(\bar{\alpha}, \bar{\beta}|\bar{X})$ [11, kap.12]. På samma sätt som tidigare går det även att maximera $\ln(\mathcal{L}(\bar{\alpha}, \bar{\beta}|\bar{X}))$ [14, s.30].

2.2.5 Markov Chain Monte Carlo

I de fall där posteriorifördelningen inte kan beräknas analytiskt kan MCMC-metoder användas för att uppskatta denna, genom att använda markovkedjor för att ta fram stickprov från fördelningen. Givet en sannolikhetsfördelning π är målet med MCMC att simulera en slumpvariabel θ som har just denna fördelning. Markovkedjan som ska skapas är en sekvens slumpvariabler $\theta_0, \theta_1, ..., \theta_n$ där varje parameter är beroende av endast den tidigare parametern i kedjan. Kedjan genereras tills dess att den konvergerar till en stationär fördelning, vilken är en god uppskattning av π. När kedjan har konvergerat används de genererade värdena som stickprov för sannolikhetsfördelningen π [15].

En vanlig metod för att konstruera markovkedjan är Metropolis-Hastings-algoritmen. Algoritmen konstruerar en reversibel markovkedja $\theta_0, ..., \theta_n$ som har stationär fördelning (målfördelning) $\pi(\theta)$, där π är en diskret sannolikhetsfördelning [15]. Algoritmen fortgår enligt följande:

1. Simulera en startpunkt θ_1 för algoritmen. Låt sedan θ_t vara det nuvarande värdet i kedjan vid en viss tid t.

2. Generera ett förslagsvärde θ^* till kedjan från förslagsfördelningen $q(\theta^*|\theta_t)$, som beskriver sannolikheten för att θ^* föreslagits givet att det nuvarande värdet är θ_t.

3. Beräkna sedan acceptanssannolikheten

$$\alpha = \min\left\{\frac{\pi(\theta^*)q(\theta_t|\theta^*)}{\pi(\theta_t)q(\theta^*|\theta_t)}, 1\right\} \tag{8}$$

   där $\pi(\theta^*)$ är målfördelningen vid θ^*, vilket mäter hur sannolikt det nya värdet är enligt den sökta fördelningen. Hela kvoten beskriver därmed hur väl det föreslagna värdet stämmer överens med målfördelningen, jämfört med det tidigare värdet i markovkedjan.

4. Generera ett slumpmässigt tal $U \sim U(0,1)$ som är likformigt fördelat mellan (0,1). Om $U \leq \alpha$ accepteras det föreslagna steget och då sätts $\theta_{t+1} = \theta^*$, alltså det föreslagna värdet läggs till i kedjan. Annars avslås det föreslagna steget, och då sätts istället $\theta_{t+1} = \theta_t$.

Steg 2-5 i algoritmen upprepas sedan fram till att kedjan som skapats konvergerat till den önskade sannolikhetsfördelningen [10, kap. 9]. Förslagsfördelningen q som används för att utforska potentiella nya steg i markovkedjan kan i stort sett vara godtycklig, så länge det är möjligt att enkelt dra ett stickprov från denna. Är förslagsfördelningen symmetrisk, så att $q(\theta_t|\theta^*) = q(\theta^*|\theta_t)$, kan kvoten i uttrycket för acceptanssannolikheten α förenklas till att endast vara en kvot mellan målfördelningarna. Detta gäller bland annat då förslagsfördelningen är normalfördelad [10].

Med en sekvens av stickprov $\theta_1, \theta_2, \ldots, \theta_N$ för posteriorifördelningen som genererats genom MCMC kan därefter ekvation (4) uppskattas numeriskt enligt

$$\pi(y_{\text{new}}|\theta) = \int \pi(y_{\text{new}}|\theta)\pi(\theta|y) \approx \frac{1}{N}\sum_{i=1}^{N}\pi(y_{\text{new}}|\theta_i), \tag{9}$$

där $\pi(y_{\text{new}}|\theta)$ är den prediktiva fördelningen för nya observationer y_{new} [16, s.262].

2.2.6 Den multivariata normalfördelningen

Normalfördelningen kan generaliseras till flerdimensionella vektorer, denna kallas för den multivariata normalfördelningen, multinormala fördelningen eller multinormalfördelningen. En multinormalfördelad $n-$dimensionell vektor $\bar{X} = \begin{bmatrix} X_1 & X_2 & \ldots & X_n \end{bmatrix}$ kan beskrivas med $\bar{\mu} = \mathbb{E}(\bar{X})$ och en kovariansmatris

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1n} \\ \lambda_{21} & \ddots & & \lambda_{2n} \\ \vdots & & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \ldots & \lambda_{nn} \end{bmatrix},$$

där $i,j \in \{1,2,\cdots,n\}$ och $\lambda_{ij} = \lambda_{ji} = \text{Cov}(X_i, X_j)$ om $i \neq j$ och $\lambda_{ii} = \text{Var}(X_i)$ annars. Detta kan betecknas som $\bar{X} \sim \mathcal{N}(\bar{\mu}, \Lambda)$. En konsekvens av denna generalisering är att alla komponenter X_i av $\bar{X}$ är normalfördelade [17].

2.2.7 Centrala gränsvärdessatsen

Centrala gräsvärdessatsen säger att summan av oberoende och likafördelade slumpvariabler approximerar en normalfördelning. Mer precist, låt $\theta_1,\theta_2,\ldots,\theta_n$ vara oberoende och likfördelade slumpvariabler med väntevärde μ och varians σ^2. För stora n gäller då [18],

$$\theta_1 + \cdots + \theta_n \approx \mathcal{N}(n\mu, n\sigma^2).$$

Satsen kan även generaliseras till det flerdimensionella fallet, där summan istället approximeras av en multinormalfördelning [19].

2.2.8 Mätning av modellprestanda med AUC-värden

En vanlig teknik för att utvärdera en prediktionsmodells prestanda är korsvalidering. Vid korsvalidering delas datamängden upp i två delar: den träningsdata som används för att bygga modellen, och den testdata som används för att bedöma modellens förutsägelseförmåga. Det är viktigt att testdata inte används under träningen av modellen, då detta kan leda till att modellens förutsägelseförmåga överanpassas. En särskild typ av korsvalidering är LOOCV (eng: leave one out cross validation), där varje enskild datapunkt används som testdata, medan resterande datapunkter används som träningsdata. I varje iteration tränas modellen på träningsdata och utvärderas på den aktuella testdata. Proceduren upprepas tills varje datapunkt har använts som testdata en gång. LOOCV möjliggör beräkning av teststatistik som återspeglar modellens förutsägelseförmåga, där medelvärdet av dessa utvärderingar används som en uppskattning av modellens generella prestanda. LOOCV är särskilt lämplig för fall med små datamängder [20].

För att kvantifiera modellens prestanda kan en konfusionsmatris (eng: confusion matrix) användas. Varje rad i denna matris representerar antalet gånger modellen förutspådde en viss klassifikation, medan varje kolumn visar det faktiska antalet individer i respektive klass. En klassifikation kan delas upp i en positiv klass och en negativ klass. Exempelvis kan en positiv klass vara blåögdhet och en negativ klass icke-blåögdhet. Detta ger upphov till begreppen sanna positiva förutsägelser, falska positiva förutsägelser, falska negativa förutsägelser och sanna negativa förutsägelser. Dessa kommer hädanefter benämnas TP, FP, FN respektive TN, efter engelskans true och false, positive och negative. I tabell 1 presenteras en generaliserad konfunsionsmatris som använder sig av dessa benämningar.

Tabell 1: Tabellen visar ett exempel på en konfusionsmatris, där TP står för true positive, FP för false positive, FN för false negative och TN för true negative.

		Förutspådda klasser	
		Positiv förutsägelse	Negativ förutsägelse
Observerade klasser	Positiva klasser	TP	FN
	Negativa klasser	FP	TN

Med detta kan måtten sensitivitet, specificitet och noggrannhet definieras. Sensitiviteten, som är ett mått på andelen korrekta förutsägelser på den positiva klassen, definieras som $\frac{TP}{TP+FN}$. Specificiteten ges av $\frac{TN}{FP+TN}$ och mäter istället andelen korrekta förutsägelser på den negativa klassen. Slutligen beräknas noggrannheten till $\frac{TP+TN}{TP+FP+FN+TN}$ och tolkas som ett mått på andelen korrekta gissningar på den totala datamängden.

Modellens prestanda kan sedan visualiseras med hjälp av ROC-kurvor (eng: Receiver Operating Characteristic) och arean under dessa [21]. Denna area kallas för AUC (eng: Area Under Curve). En ROC-kurva plottas på området $[0,1] \times [0,1]$, där y-axeln representerar modellens sensitivitet och x-axeln visar $1 -$ specitiviteten $= \frac{FP}{FP+TN}$. Kurvan plottas genom att betrakta sannolikheterna för den positiva klassen för respektive individ. En gräns varieras sedan stegvis, där individerna vars sannolikhet överskrider gränsen klassificeras som positiva och övriga negativa. Till exempel innebär en gräns på 0% att modellen tilldelar den positiva klassen till alla individer oavsett sannolikheterna den tilldelat dem, och en gräns på 100% innebär att modellen endast klassificerar en individ som medlem av den positiva klassen om den är 100% säker på att individen tillhör den klassen. Kurvan skapas sedan genom att öka gränsen från 0% till 100% och plotta sensitiviteten och $1-$specitiviteten vid varje värde.

AUC-värdet är ett lämpligt mått på en modells prestanda då det är invariant mot fördelningen av klasserna i datamängden [21]. En annan fördel med AUC är att den kan tolkas som sannolikheten att modellen tilldelar en högre sannolikhet för en positiv förutsägelse till en slumpvald individ från den positiva klassen, jämfört med en slumpvald individ från den negativa klassen. Detta innebär att AUC får ett värde mellan 0,5 och 1, eftersom modellens prediktioner kan inverteras och byta plats på positiva och negativa utfall om värdet är mindre än 0,5.

Värt att notera är att metoden endast fungerar för binära utfallsrum. Det finns flera generaliseringar av ROC-grafer och AUC [21]. En av dessa generaliseringar är att producera s stycken ROC-grafer, där s är antalet klasser. För varje klass låts den valda klassen vara den positiva klassen och resterande klasser vara den negativa klassen. AUC-värdet kan sedan beräknas för varje klass. Notera att detta kan bli lägre än 0,5 eftersom det nu inte går att vända om modellen på samma sätt som för två kategorier. Det totala AUC-värdet för modellen väljs sedan som ett viktat medelvärde av alla AUC-värden enligt

$$\mathrm{AUC}_{\mathrm{total}} = \sum_{i=1}^{s} \frac{c(i)}{N} \cdot \mathrm{AUC}_i, \tag{10}$$

där $c(i)$ är antalet individer av klass i i datamängden, N antalet individer i hela datamängden och AUC_i är värdet med klass i som det positiva utfallet [22]. Fördelen med denna generalisering är att den är enkel att beräkna och visualisera. Dock sker den på bekostnad av AUC-värdenas invarians av klassfördelningar [21].

Generellt gäller det att det högsta AUC-värdet är det bästa. Det tolkas som att modeller med AUC-värde $\geq 0{,}8$ har god prestanda, och AUC $\geq 0{,}9$ har utmärkt prestanda. Även AUC $\geq 0{,}7$ kan tolkas som en acceptabel prestanda för modellen, medan AUC $\leq 0{,}7$ kan tolkas som att modellens prestanda är låg eller otillräcklig [23].

3 Metod

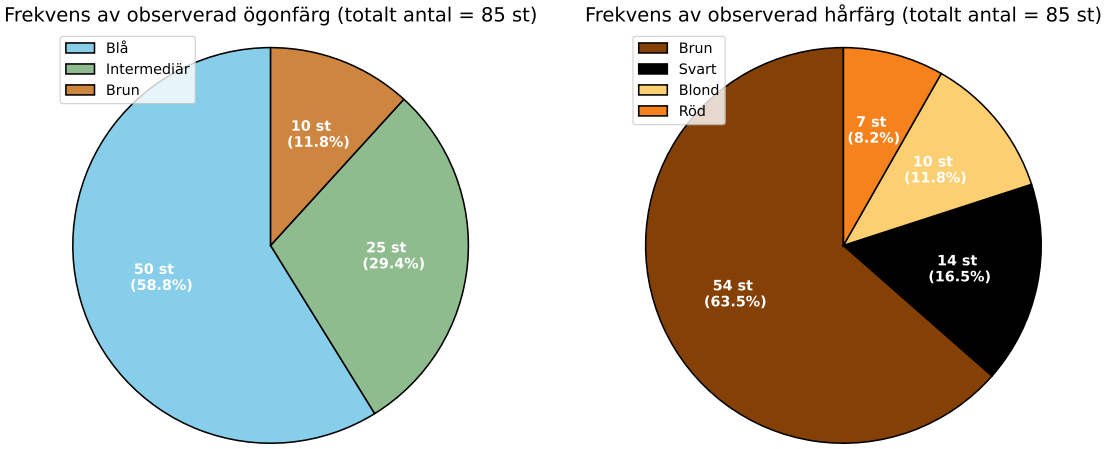
Eftersom en prediktionsmodell kan utvecklas på flera olika sätt behövde metoden avgränsas genom flera val. I följande delkapitel beskrivs hur de centrala komponenterna i modellen är sammanfogade samt de metodval som har gjorts.

3.1 Avgränsningar

Utförandet av projektet begränsades främst av tidsramen, men även av tillgången till data för de fenotypiska egenskaper som modellen predicerade. Av denna anledning valdes det att endast utveckla modeller för att predicera hår- och ögonfärg, medan andra fenotypiska egenskaper uteslöts. Tidsbegränsningen påverkade även de metodval som undersöktes, vilket blev valet av regressionsmodell samt priorifunktionen. Den typ av modell som arbetet fokuserade på var logistisk regression, där både en binär och en multinomial variant utvecklades och analyserades. Priorifunktionen går att väljas på flera sätt, men denna rapport begränsade sig till att undersöka två alternativ: en icke-proper likformig fördelning och en multinormalfördelning. Den icke-propra likformiga priorifunktionen är vanligt förekommande när det saknas tidigare information om parametrarnas fördelning [24, s.24]. Användningen av en multinormalfördelning motiverades av centrala gränsvärdessatsen, då modellparametrarna antogs gå mot en multinormalfördelning.

3.2 Datainsamling och databehandling

Datamaterialet var uppsamlat och givet av Rättsgenetik vid nationellt forensiskt centrum i Linköping. Den var uppdelad i två delar, en för ögonfärg och en för hårfärg. Varje del innehöll information kopplad till 85 individers ögon- respektive hårfärg och deras tillhörande genotyp. Fördelningarna av de observerade hår- och ögonfärgerna presenteras i figur 1. Ögonfärgerna delades upp i kategorierna blå, intermediär och brun. De intermediära ögonfärgerna avsåg de som inte tydligt kunde klassificeras som blå eller brun. Hårfärgerna klassificerades som brun, svart, blond och röd. Dessa uppdelningar hade med störst sannolikhet gjorts genom visuell bedömning av en människa, vilket kan introducera viss variation i datamängden genom subjektiv bedömning.



(a) (b)

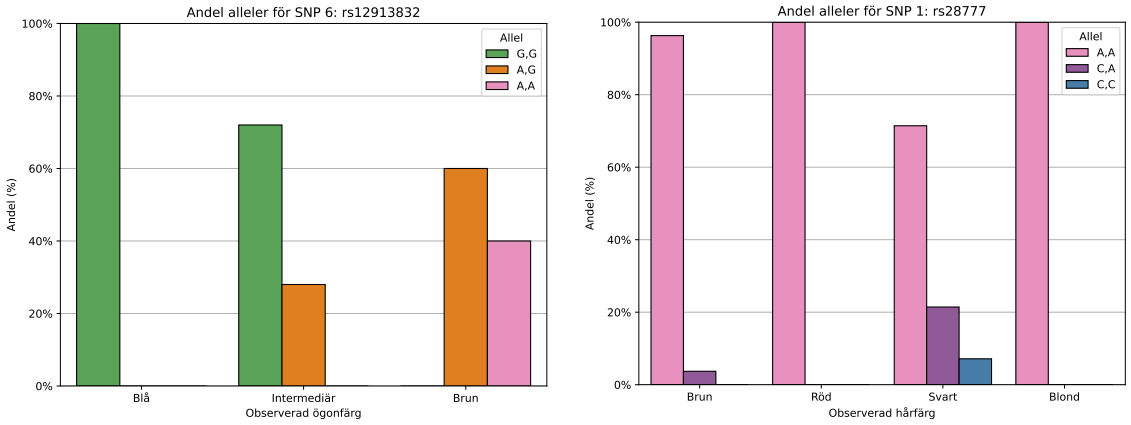
Figur 1: Frekvensfördelningen av observerad ögonfärg (a) och hårfärg (b), angivet i både antal och andel i procent. Det totala antalet observerade individer är 85 för respektive datamängd.

I den givna datamängden fanns ett antal SNP:er för ögonfärgerna och hårfärgerna. Namnen på dessa presenteras i tabell 2.

Tabell 2: Givna SNP:er i datamängden som användes i denna studie.

SNP:er för ögonfärg		**SNP:er för hårfärg**				
rs12203592	rs1393350	rs28777	rs12203592	rs4959270	rs1805005	rs1805006
rs12896399	rs1800407	rs683	rs1042602	rs12821256	rs11547464	rs1805007
rs16891982	rs12913832	rs312262906_N29insA	rs1800407	rs2402130	rs1110400	rs1805008
		rs1805009	rs2378249	rs2228479	rs12913832	rs885479
		rs201326893_Y152OCH	rs16891982			

För att visualisera datamängden och få en bättre förståelse för hur varje SNP påverkade förekomsten av de olika fenotyperna visualiserades de i ett histogram över frekvensen av alla alleler för alla SNP:er. För data över ögonfärg observerades det att i SNP rs12913832 hade alla med blå ögonfärg allelen GG, medan endast personerna med allelen AA hade brun ögonfärg enligt figur 2a. Detta innebar att om en individ hade AA i den specifika SNP:n hade den bruna ögon enligt den klassificering av ögonfärg som användes. Liknande slutsatser kunde även dras för flera genetiska markörer för hårfärgsdatasetet. Exempelvis var den svarta hårfärgen den enda som innehöll allelen CC i SNP rs28777 enligt figur 2b. Resterande genetiska markörer som uppvisade liknande beteende presenteras i bilaga A.



(a) (b)

Figur 2: Andelen, i procent, av de olika allelerna i rs12913832 (a) och rs28777 (b) för ögonfärg respektive hårfärg.

För hårfärgsdatamängden kunde det även observeras att vissa SNP:er endast innehöll samma allel för alla individer. Dessa presenteras i tabell 3. Då dessa inte kan användas för att särskilja individer påverkar de inte prediktionsförmågan hos modellen, och därför togs de bort.

Tabell 3: SNP:er med endast en allel för alla observationer i hårfärgsdatasetet.

SNP	**Observerad allel**
rs312262906_N29insA	CC
rs1805006	CC
rs201326893_Y152OCH	CC
rs11547464	GG
rs1110400	TT

För att möjliggöra analysen behövde datamängden tilldelas numeriska värden. Därför tilldelades klassifikationerna och varje allel inom varje SNP ett numeriskt heltal. Till exempel valdes blå till 0, intermediär till 1, brun till 2 för den multinomiala ögonfärgsmodellen och för SNP rs12203592 tilldelades allelerna CC och CT värdena 1 respektive 2. På detta sätt tilldelades en modellparameter β_i för varje SNP.

3.3 Modellutveckling

För denna undersökning användes programmeringsspråket R. Utvecklingen av de binära och de multinomiala logistiska regressionsmodellerna var nästan identisk. De skiljde sig endast åt i hur trolighetsfunktionerna definierades: för den binära modellen beskrevs trolighetsfunktionen av ekvation (5), medan den beskrevs av ekvation (7) för den multinomiala modellen.

De utvecklade modellerna grundades i en bayesiansk statistikram, som introducerades i teoridelen av rapporten. Syftet med modellerna var att, givet observerad data y, beräkna sannolikheten $\pi(y_{\text{new}}|y)$ för nya prediktioner y_{new}. Denna sannolikhet approximerades numeriskt med hjälp av MCMC-metoder enligt ekvation (9). För att möjliggöra detta introducerades stokastiska parametrar θ.

ML-skattningen för θ valdes som starten av markovkedjan. Denna extrempunkt beräknades numeriskt genom R-funktionen *nlm*, där samtliga parametrar initialiserades med startvärde 1. Därefter simulerades en markovkedja med 10000 iterationer, förutom för de binära hårfärgsmodellerna där

markovkedjan simulerades med 50000 iterationer. Varje nytt förslag på parametervektorn θ^* genererades enligt $\theta^* = \theta_t + \epsilon$, där θ_t var det senaste elementet i markovkedjan och ϵ var en slumpmässig vandring med multinormalfördelningen $\mathcal{N}(0, \sigma^2 I)$, där I var identitetsmatrisen. Värdet på σ anpassades individuellt för varje modell: för ögonfärgerna sattes $\sigma = 0{,}7$ för binära regressionsmodeller och $\sigma = 1$ för multinomiala modeller. För hårfärgerna valdes $\sigma = 0{,}01$ för den multinomiala modellen med likformig priorifunktion, medan $\sigma = 0{,}1$ användes i övriga modeller. Valet av standardavvikelserna baserades på behovet av en balans mellan spridning och acceptans i kedjan. Om σ var för stor blev hoppet mellan gamla och nya parametrar för stort, vilket ledde till låg acceptans i markovkedjan, medan för små σ skulle det innebära att kedjan konvergerade mycket långsamt.

För att avgöra om det föreslagna steget θ^* skulle accepteras och läggas till i markovkedjan beräknades acceptanssannolikheten enligt ekvation (8). Eftersom förslagsfunktionen $q(\theta^*|\theta_t) = \mathcal{N}(\theta_t, \sigma^2)$ hade en symmetrisk densitet kring väntevärdet θ_t innebar det att $q(\theta^*|\theta_t) = q(\theta_t|\theta^*)$, vilket förenklade uttrycket för acceptanssannolikheten enligt teoriavsnitt 2.2.5. Posteriorifördelningarna för både θ_t och θ^* beräknades med hjälp av ekvation (1), där trolighetsfunktionen i sin tur beräknades via ekvation (5) för de binära modellerna respektive ekvation (7) för de multinomiala modellerna.

Det fanns många olika priorifunktioner att välja bland. I detta arbete valdes det att utveckla modeller både med en icke-proper likformig priorifunktion och med en multinormal priorifunktion. Vid icke-proper likformig priori gällde det att $\pi(\theta_t) = \pi(\theta^*)$, vilket förenklade acceptanssannolikheten ytterligare. För den multinormalfördelade priorifunktionerna användes R-funktionen *dmnormt* från biblioteket *mnormt*, som beräknade sannolikheten för en multinormalfördelning med väntevärde μ och en kovariansmatris Λ. Här valdes ML-värdet θ_{start} som väntevärde och en diagonal kovariansmatris med värde d_i längs diagonalen. Detta val baserades på det naiva antagandet att generna för olika SNP:er var oberoende, vilket innebar att kovarians mellan dem var 0. Värdet för diagonalelementen d_i valdes utifrån observationer av konvergensen av parametrarna för modellerna med de icke-propra likformiga priorifunktionerna, se bilaga B.5. För modellerna baserade på hårfärgsdatamängden sattes samtliga diagonalelement till 1, och för den binära modellen för ögonfärgsdatamängden valdes $d_i = 10$ för SNP rs12913832 och konstanttermen medan resterande diagonalelement valdes till $d_i = 1$. För den multinomiala ögonfärgsmodellen sattes samtliga diagonalelement till $d_i = 10$.

Av numeriska stabilitetsskäl beräknades logaritmen av acceptanssannolikheten. Detta påverkade inte beslutet så länge som logaritmen även togs av det slumpmässiga talet U i steg 4 av Metropolis-Hastings-algoritmen. Då accepterades θ^* om $\log(U) \leq \log(\alpha)$, vilket var korrekt eftersom logaritmen är kontinuerlig och strikt monoton då U,$\alpha > 0$. Iterationen avslutades med att antingen lägga till θ^* på slutet av kedjan om den accepterades, eller att lägga till det gamla θ_t på kedjan. Därefter påbörjades en ny iteration av MCMC-algoritmen.

Med den senare halvan av markovkedjan kunde uppskattningen av sannolikheten $\pi(y_{\text{new}}|y)$ beräknas enligt ekvation (9). Anledningen till att den första halvan av kedjan inte användes var för att ta bort den delen av kedjan då parametrarna inte hade konvergerat till den stationära fördelningen. Därefter valdes den klassificering som maximerade sannolikheten $\pi(y_{\text{new}}|y)$ som modellens prediktion.

3.4 Modellvalidering

Korsvalidering av typen LOOCV, som beskrivs i teoriavsnitt 2.2.8, användes för samtliga modeller för att beräkna teststatistik för varje enskild datapunkt. Dessa teststatistiker användes för att konstruera konfusionsmatriser som tabulerade modellernas prestanda. Teststatistiken användes även för att skapa ROC-kurvor samt beräkna AUC-värden, vilket genomfördes med hjälp av R-funktionen *roc* och *auc* från biblioteket *pROC*. LOOCV var lämpligt att använda eftersom datamängderna var förhållandevis små.

4 Resultat

Avsnittet nedan börjar med att presentera en sammanställning av de viktigaste resultaten i tre tabeller: modellernas AUC-värden, sensitivitet och specificitet. Därefter uppvisas resultatet mer

utförligt för två modeller: den binära ögonfärgsmodellen med en icke-proper likformig priorifunktion och den multinomiala hårfärgsmodellen med en multinormal priorifunktion. Detta då dessa modeller är de som uppvisar bäst respektive sämst prestanda. Resultaten som presenteras för de två modellerna är deras ROC-kurvor, konfusionsmatriser och sannolikhetsfördelningarna över prediktionerna. Resultaten för de övriga modellernas ROC-kurvor, konfusionsmatriser och sannolikhetsfördelningar presenteras i bilaga B.1, B.2 respektive B.4. Påverkan av en sannolikhetströskel på 70% för samtliga modeller visualiseras i bilaga B.3. Konvergensen för parametrarna för de olika modellerna illustreras i bilaga B.5. Det kan observeras att alla parametrar konvergerade. Undantag görs för den multinomiala modellen baserad på hårfärgsdata eftersom den modellen använder 54 modellparametrar vilket är opraktiskt att analysera.

4.1 AUC-värden för alla modeller

Modellernas AUC-värden sammanställs i tabell 4 där varje kolumn representerar de fall då den givna färgen valdes till det positiva utfallet. Det totala AUC-värdet beräknades med ekvation (10).

Tabell 4: Sammanställning av AUC-värden för de olika modellerna med olika priorifunktioner.

Modell	Priorifunktion	Blå	Intermediär	Brun	Blond	Röd	Svart	**Total**
Binär ögonfärg	Icke-proper likformig	0,732	-	-	-	-	-	**0,732**
	Multinormal	0,7154	-	-	-	-	-	**0,7154**
Multinomial ögonfärg	Icke-proper likformig	0,7469	0,5847	0,5747	-	-	-	**0,6789**
	Multinormal	0,7194	0,5713	0,5573	-	-	-	**0,6568**
Binär hårfärg	Icke-proper likformig	-	-	0,5812	-	-	-	**0,5812**
	Multinormal	-	-	0,5795	-	-	-	**0,5795**
Multinomial hårfärg	Icke-proper likformig	-	-	0,5125	0,5840	0,9469	0,6529	**0,5798**
	Multinormal	-	-	0,5119	0,5800	0,9469	0,6479	**0,5781**

4.2 Sensitivitets- och specificitetstabeller

Tabell 5 visar modellernas sensitivitet och noggrannhet. Tabell 6 visar modellernas specificitet. Vid beräkning av specificiteterna ansågs samtliga negativa klasser som samma klass. Exempelvis sågs en gissning på brun ögonfärg på en individ med intermediär ögonfärg som en sann negativ förutsägelse eftersom både färgerna var negativa klasser när blå ögonfärg valdes som positiv klass.

Tabell 5: Sammanställning av sensitiviteter för olika modeller med olika priorifunktioner. Vid användande av en sannolikhetströskel togs de individerna där prediktionsmodellen gav en lägre sannolikhet än 70% ut ur totalen. Noggrannheten visas i kolumnen längst till höger.

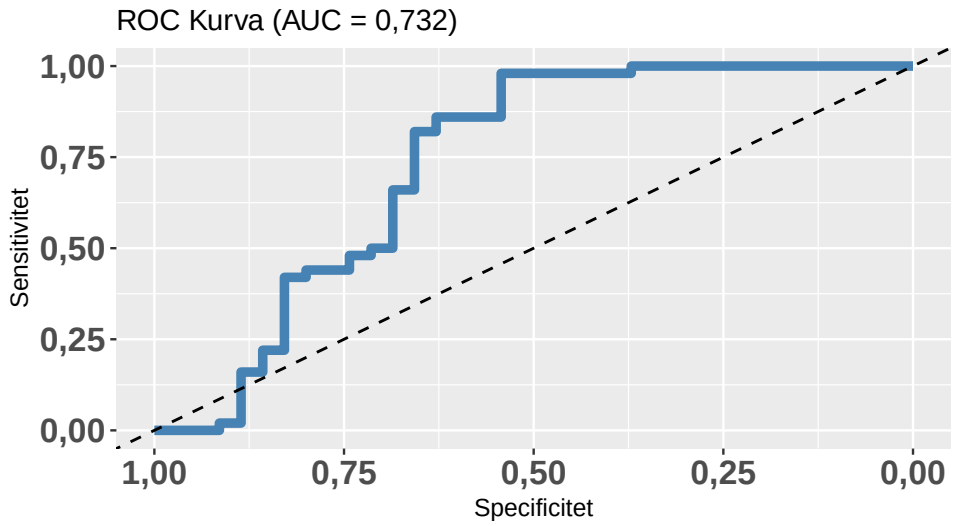
Modell	Priorifunktion	Blå	Intermediär	Brun	Blond	Röd	Svart	Icke-blå	Icke-brun	**Noggrannhet**
Multinomial ögonfärg	Icke-proper likformig	94%	28%	70%	-	-	-	-	-	**71,76%**
	Multinormal	94%	28%	70%	-	-	-	-	-	**71,76%**
Multinomial ögonfärg (70% **tröskel**)	Icke-proper likformig	97,56%	27,78%	66,67%	-	-	-	-	-	**75%**
	Multinormal	97,30%	31,58%	70%	-	-	-	-	-	**74,24%**
Binär ögonfärg	Icke-proper likformig	96%	-	-	-	-	-	54,29%	-	**78,82%**
	Multinormal	98%	-	-	-	-	-	54,29%	-	**80%**
Binär ögonfärg (70% **tröskel**)	Icke-proper likformig	97,30%	-	-	-	-	-	58,62%	-	**80,30%**
	Multinormal	97,50%	-	-	-	-	-	58,62%	-	**81,16%**
Multinomial hårfärg	Icke-proper likformig	-	-	85,19%	0%	57,14%	35,71%	-	-	**64,71%**
	Multinormal	-	-	85,19%	0%	57,14%	35,71%	-	-	**64,71%**
Multinomial hårfärg (70% **tröskel**)	Icke-proper likformig	-	-	76,47%	0%	60%	37,50%	-	-	**59,26%**
	Multinormal	-	-	77,14%	0%	60%	28,57%	-	-	**59,26%**
Binär hårfärg	Icke-proper likformig	-	-	75,93%	-	-	-	-	22,58%	**56,47%**
	Multinormal	-	-	77,78%	-	-	-	-	19,35%	**56,47%**
Binär hårfärg (70% **tröskel**)	Icke-proper likformig	-	-	78,57%	-	-	-	-	14,29%	**51,02%**
	Multinormal	-	-	83,87%	-	-	-	-	10,00%	**54,90%**

Tabell 6: Sammanställning av specificiteter för olika modeller med olika priorifunktioner. Vid användande av en sannolikhetströskel togs de individerna där prediktionsmodellen gav en lägre sannolikhet än 70% ut ur totalen.

Modell	Priorifunktion	Blå	Intermediär	Brun	Blond	Röd	Svart	Icke-blå	Icke-brun
Multinomial ögonfärg	Icke-proper likformig	54,29%	90,00%	97,33%	-	-	-	-	-
	Multinormal	51,43%	90,00%	98,67%	-	-	-	-	-
Multinomial ögonfärg (70% tröskel)	Icke-proper likformig	59,26%	92,00%	96,61%	-	-	-	-	-
	Multinormal	58,62%	91,49%	98,21%	-	-	-	-	-
Binär ögonfärg	Icke-proper likformig	54,29%	-	-	-	-	-	96,00%	-
	Multinormal	54,29%	-	-	-	-	-	98,00%	-
Binär ögonfärg (70% **tröskel)**	Icke-proper likformig	58,62%	-	-	-	-	-	97,30%	-
	Multinormal	58,62%	-	-	-	-	-	97,30%	-
Multinomial hårfärg	Icke-proper likformig	-	-	32,26%	97,33%	97,44%	92,96%	-	-
	Multinormal	-	-	32,26%	97,33%	97,44%	92,96%	-	-
Multinomial hårfärg (70% **tröskel)**	Icke-proper likformig	-	-	30,00%	97,87%	95,92%	89,13%	-	-
	Multinormal	-	-	28,32%	97,87%	95,92%	89,36%	-	-
Binär hårfärg	Icke-proper likformig	-	-	22,58%	-	-	-	-	75,93%
	Multinormal	-	-	19,35%	-	-	-	-	77,78%
Binär hårfärg (70% **tröskel)**	Icke-proper likformig	-	-	14,29%	-	-	-	-	78,57%
	Multinormal	-	-	10,00%	-	-	-	-	83,87%

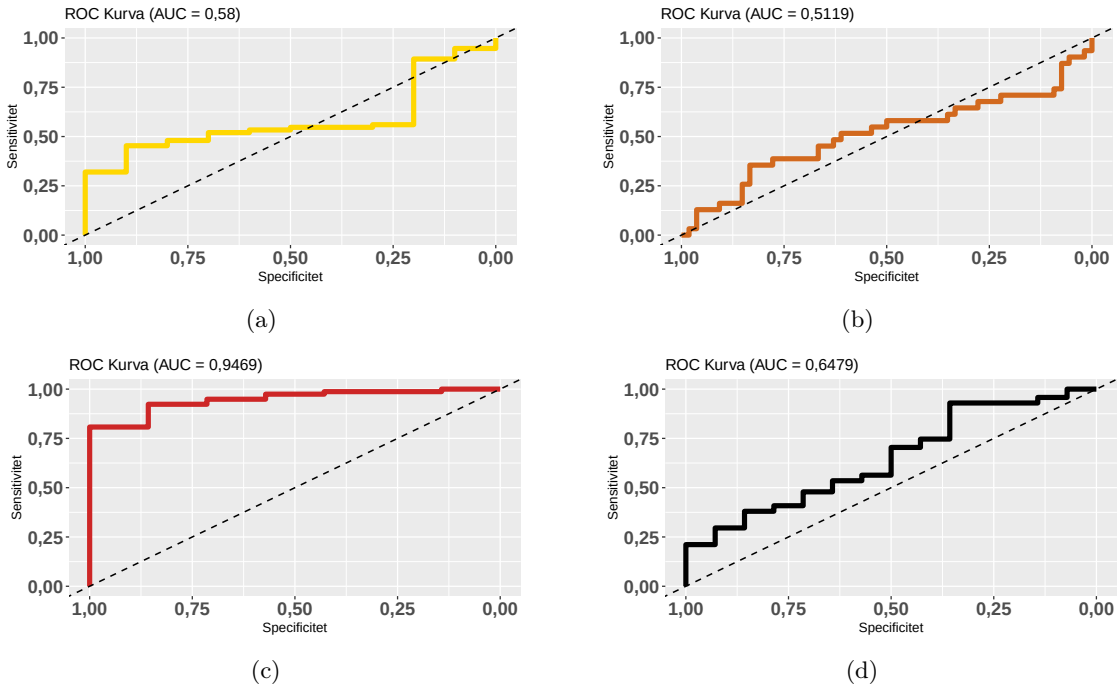
4.3 ROC-kurvor för de utvalda modellerna

Figur 3 visar ROC-kurvan för den binära ögonfärgsmodellen med en icke-proper likformig prior, tillsammans med det tillhörande AUC-värdet på 0,732.



Figur 3: ROC-kurva och tillhörande AUC för binär ögonfärg vid användning av en icke-proper likfördelad priorifunktion. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

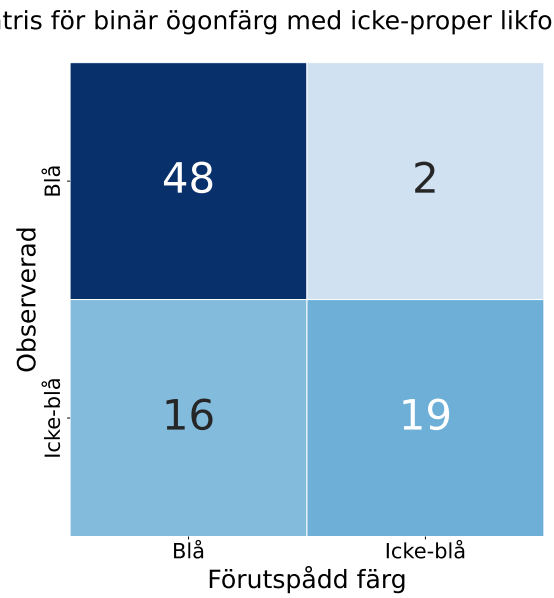
Det totala AUC-värdet för multinomiala hårfärgsmodellen med multinormal priorifunktion beräknades till 0,5781 enligt ekvation (10). ROC-kurvor och AUC-värden för respektive hårfärg presenteras i figur 4, där AUC uppgick till 0,5800 för blond, 0,5119 för brun, 0,9469 för röd och 0,6479 för svart hårfärg.



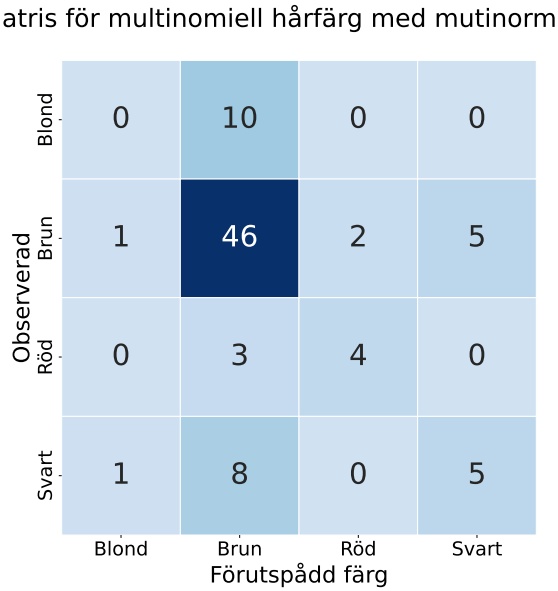
Figur 4: ROC-kurvor för den multinomiala modellen för hårfärg vid användande av multinormal prior. Kurvan (a) tillhör blond hårfärg, (b) brun, (c) röd och (d) svart som positivt utfall. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

4.4 Konfusionsmatriser för de utvalda modellerna

Konfusionsmatrisen för den binära ögonfärgsmodellen med en icke-proper likformig priorifunktion visas i figur 5. Denna predicerar förekomsten eller avsaknaden av blå ögonfärg hos individerna. Figur 6 presenterar konfusionsmatrisen för den multinomiala hårfärgsmodellen med en multinormal priorifunktion.



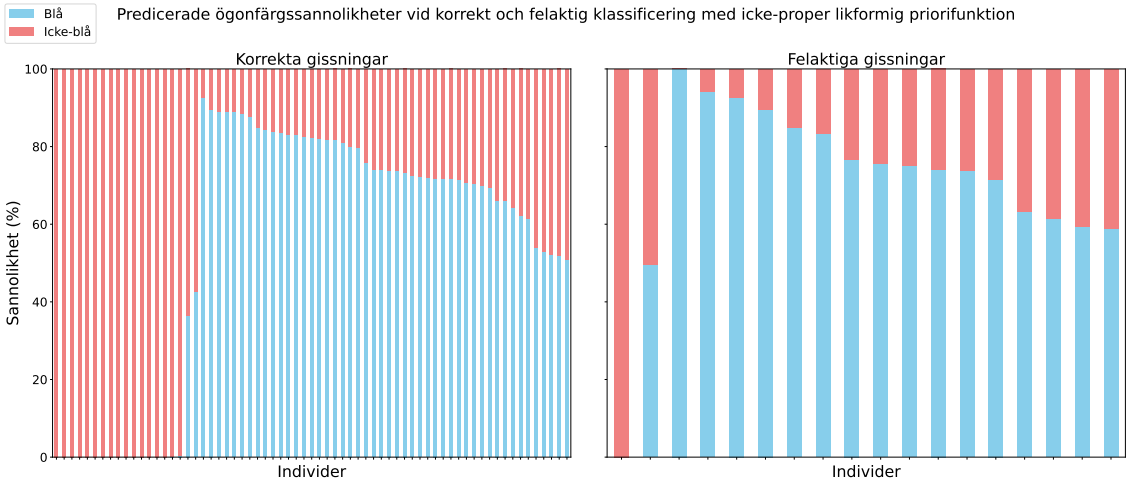
Figur 5: Konfusionsmatrisen för den binära ögonfärgsmodellen med en icke-proper likformig priorifunktion. Matrisen visar antal korrekta och felaktiga klassificeringar, där diagonalen visar träffar och övriga rutor felklassificeringar.



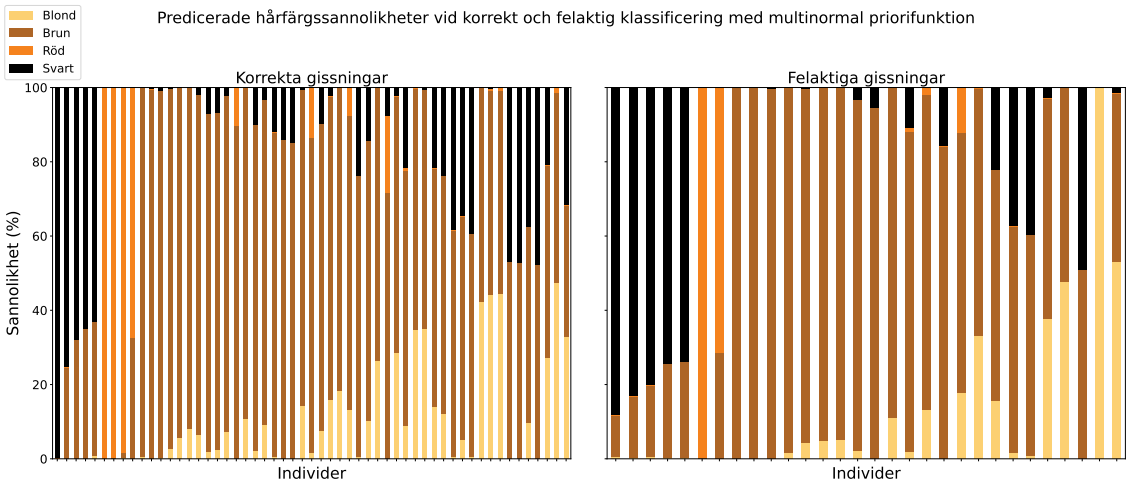
Figur 6: Konfusionsmatrisen för den multinomiala hårfärgsmodellen med en multinormal priorifunktion. Matrisen visar antal korrekta och felaktiga klassificeringar, där diagonalen visar träffar och övriga rutor felklassificeringar.

4.5 Sannolikhetsfördelningar för de utvalda modellerna

Figur 7a visar resultatet över de predicerade sannolikheterna för den binära ögonfärgsmodellen med en icke-proper likformig prioritetsfördelning. Varje stapel innehåller information om procentsatserna för samtliga klassificeringar för varje prediktion i LOOCV. Figuren är uppdelad i korrekta och felaktiga prediktioner där staplarna är sorterade utefter klassificeringen av prediktionen. Sannolikhetsfördelningen för den multinomiala hårfärgsmodellen med en multinormal priorifunktion uppvisas i figur 7b.



(a)



(b)

Figur 7: Sannolikhetsfördelningarna för binär ögonfärgsmodellen och multinomial hårfärgsmodell. Figur (a) visar sannolikheterna för prediktionsmodellen för binär ögonfärg med icke-proper likformig priorifunktion. Figur (b) visar sannolikheterna för prediktionsmodellen för multinomial hårfärg med en multinormal priorifunktion. De vänstra figurerna illustrerar de korrekta prediktionerna medan de högra figurerna illustrerar de felaktiga prediktionerna.

5 Diskussion

I detta arbete har det studerats hur ögon- och hårfärg kan förutspås från genetiska markörer i DNA utifrån de tre centrala aspekterna: datamaterialets tillförlitlighet, modellernas konstruktion samt dess prestanda. Två huvudsakliga prediktionsmodeller baserade på logistisk regression har framtagits, en binär och en multinomial. Modellerna utvecklades ytterligare genom att två olika priorifunktioner testades. Dessa var en icke-proper likformig priorifunktion samt en multinormal priorifunktion. Sammanlagt utvecklades därmed fyra olika prediktionsmodeller för varje datamängd.

I följande avsnitt tolkas och jämförs resultatet från prediktionsmodellerna med hänsyn till hur väl de presterar, samt mot bakgrund av tidigare forskning inom ämnet. Därtill diskuteras begränsande faktorer som kan ha haft inverkan på resultatet, likaså förslag på möjliga förbättringar för framtida studier.

5.1 Tolkning av resultat

Resultaten visar överlag att modellen uppnår låga AUC-värden, vilket tyder på en otillfredsställande prediktiv förmåga. Dessutom kan det observeras i tabell 4 att ett användande av en multinormal priorifunktion gav lägre AUC-värden jämfört med de modeller som använde en icke-proper likformig priorifunktion. De totala AUC-värdena varierar mellan 0,5781 och 0,732, vilket motsvarar en prestanda som sträcker sig från låg till, i bästa fall, acceptabel. Modellen för binär ögonfärg, med ett totalt AUC-värde på 0,7154 och 0,732 för en multinormal respektive likformig prior, är den enda modellen som uppnår en acceptabel prestanda enligt intervallen i kapitel 2.2.8. Det tyder på att det finns egenskaper bland den givna DNA-data som modellen kan använda för att förutspå ögonfärger, men att dessa inte är tillräckliga för att få säkra svar från modellerna. Däremot lyckades modellen för multinomial hårfärg få ett AUC-värde på 0,9469 för röd hårfärg, vilket kan tyda på att modellerna har en förmåga att särskilja rödhåriga från övriga individer. Emellertid kan det innebära att de rödhåriga individerna i datamängden har varit lättare att särskilja från resterande individer, men att detta nödvändigtvis inte är representativt för alla rödhåriga.

Genom att använda LOOCV för att utvärdera modellernas prestanda framgår det att samtliga modeller tenderar att predicera den klassifikation med flest datapunkter. För ögonfärgsdatamängden är det den blåa färgen med en andel på 58,8% och för hårfärgsdatamängden är det den bruna färgen med en andel på 63,5%, vilket visas i figur 1. Att klassifikationerna med flest datapunkter överrepresenteras i prediktionerna återspeglas i konfusionmatriserna i bilaga B.2. De binära ögonfärgsmodellerna predicerar i genomsnitt blå i 75,9% av fallen, medan de multinomiala ögonfärgsmodellerna gör detta i 74,7%. De binära och multinomiala hårfärgsmodellerna förutspår brun hårfärg i 77,6% respektive 78,8% av fallen i genomsnitt. Detta kan vara en indikation på att modellerna har en bias för majoritetsklassen, vilket sannolikt beror på den obalanserade klassfördelningen i datamängderna. Det kan vara ett tecken på att minoritetsklasserna innehåller för få observationer för att modellen ska kunna identifiera tillräckliga mönster och därmed göra tillförlitliga prediktioner. Undersökning med en större datamängd skulle därför behöva göras.

Tabell 5 och 6 illustrerar modellernas sensitivitet och specificitet. Eftersom sensitivitet och specificitet är mått på andelen korrekta förutsägelser av positiva respektive negativa klasser, kan modeller med låga AUC-värden fortfarande användas beroende på användarens mål. En modell med hög sensitivitet och låg specificitet tenderar att ofta klassificera observationer tillhörande den positiva klassen, samtidigt som den sällan korrekt identifierar de negativa klasserna. Denna egenskap skulle kunna användas för att minska misstankarna kring förekomsten av den givna positiva färgen när modellen förutspår att individen tillhör den negativa klassen. Med liknande resonemang kan en hög specificitet och låg sensitivitet användas för att styrka befintliga misstankar om förekomsten av den givna färgen. Exempelvis klassificerar den multinomiala hårfärgsmodellen brun hårfärg med hög sensitivitet och låg specificitet, vilket innebär att individer som förutspås som icke-brunhåriga troligen är det. Den multinomiala hårfärgsmodellen klassificerar även röd hårfärg med hög specificitet och låg sensitivitet, vilket innebär att individer som förutspås som rödhåriga troligen är rödhåriga. Det omvända gäller däremot inte, vilket innebär att en prediktion som utesluter röd hårfärg inte bör påverka misstanken om att en individ är rödhårig.

Genom att sannolikhetströskeln på 70% läggs till kan flera prediktioner med låg säkerhet sållas bort, vilket illustreras i figurerna i bilaga B.3. Det visar sig dock i figur 22 och 23 att modellerna för hårfärg ger så pass osäkra prediktioner att en väldigt stor del av dessa inte överstiger sannolikhetströskeln. När det gäller andelen korrekta prediktioner visar tabell 5 att sannolikhetströskeln ger en liten förbättring av exaktheten för ögonfärgsmodellerna. Däremot visar tabellen att andelen korrekta prediktioner för hårfärgsmodellerna minskar med sannolikhetströskeln. Detta tyder på att fler korrekta prediktioner har låg sannolikhet, medan de felaktiga prediktionerna i större grad har högre sannolikhet. I kombination med de låga AUC-värdena för hårfärgsmodellerna uppvisar detta modellernas låga prestanda. Observera även att sannolikhetsfördelningarna i figur 7 för både korrekta och felaktiga prediktioner har en liknande fördelning av sannolikheter, vilket även gäller för de resterande modellerna som presenteras i bilaga B.4. En sannolikhetströskel bidrar alltså inte till en förbättring av prediktionsförmågan för modellerna, eftersom tröskeln tar bort samma andel korrekta och inkorrekta förutsägelser.

Tidigare forskning från Nationellt forensiskt centrum av noggrannheten för prediktionsmodeller av hår- och ögonfärg resulterade i att 80% av ögonfärgerna förutspåddes korrekt, medan 58% av hårfärgerna förutspåddes korrekt enligt teoriavsnitt 2.1.3. Denna studie visade även att sannolikhetströskeln på 0,7 ökade noggrannheten för ögonfärgerna till 85% men inte påverkade andelen korrekt förutspådda hårfärger [9]. Studien använde en datamängd av liknande storlek, vilket gör den relevant för jämförelse med detta arbete. I tabell 5 presenteras de beräknade noggrannheterna för respektive modell, med och utan sannolikhetströskel. Det uppvisas att noggrannheten för de multinomiala ögonfärgsmodellerna ger liknande resultat som i den tidigare studien, med 71,76% noggrannhet för både icke-proper likformig och multinormal priorifunktion. Dessutom orsakar sannolikhetströskeln en liknande förbättring av noggrannheten som i den tidigare studien, med 75% respektive 74,24% noggrannhet. Modellerna för multinomial hårfärg uppvisade högre noggrannhet jämfört med den tidigare studien, med 64,71% för båda priorifunktionerna. Sannolikhetströskeln medför däremot lägre noggrannhet för alla hårfärgsmodeller. En slutsats som kan dras från detta är att de framtagna modellerna ändå uppvisar förhållandevis höga noggrannheter i jämförelse med tidigare modeller som redan används inom prediktion av fenotypiska egenskaper, åtminstone när det gäller en liknande datamängd.

5.2 Begränsande faktorer och förbättringsmöjligheter

En problematik med modellerna är hur färgerna klassificeras. Möjligtvis klassificerades de observerade färgerna genom en subjektiv bedömning. Därför kan en ögonfärg som genetiskt förväntas vara blå ha klassificerats som en intermediär färg, om den subjektiva bedömningen av färgen har varit otydlig. Sedan tidigare är det känt att grön ögonfärg är en svår ögonfärg att förutspå [9]. Detta uppenbaras även i detta projekt då sensitiviteten för intermediär ögonfärg var låg enligt tabell 5. En intressant observation är att de multinomiala ögonfärgsmodellerna aldrig predicerar brun ögonfärg på blåögda individer eller blå ögonfärg på brunögda individer, vilket illustreras i figur 21 i bilaga B.3. Detta är ett tecken på att modellerna lyckas differentiera de olika ögonfärgerna men att subjektiviteten av den intermediära kategorin och de observerade ögonfärgerna introducerar en felkälla.

Det är sedan tidigare känt att hårfärger kan bli mörkare med ålder [8], vilket kan vara ett problem för klassificeringen av hårfärger. Subjektiviteten i kategorisering av data kan därmed innebära att en observerad brunhårig individ bär blonda gener. För att undvika sådana felkällor skulle en större studie behövas kring hur ögon- och hårfärger uppfattas i samband med de genetiska förutsättningarna. Det finns även risk för att mörkandet av hårfärg inte är inkodat i DNA. Mer forskning om gener för hårmörkande och modeller som även tar hänsyn till individers ålder kan möjligtvis ge bättre resultat.

Ett ytterligare problem med modellerna är valet av att representera förekomsten av olika alleler med heltal. Den numeriska representationen introducerar en artificiell ordning på allelerna som inte överensstämmer med dess verkliga karaktär, eftersom det finns en inbördes ordning hos de reella talen. Detta kan påverka modellernas förmåga att anpassa parametrarna till datamängden. Framtida modeller kan kräva en bättre förbehandling av data för att motverka problemet, exempelvis genom att representera varje möjlig allel för varje SNP med en separat parameter. Dessutom skulle detta potentiellt förbättra modellens förmåga att identifiera mönster i allelfördelningen för varje SNP. Modellen skulle, i enlighet med figur 2a, exempelvis få möjlighet att lära sig att endast individer som har allelparet AA för SNP rs12913832 är brunögda. Då den föreslagna modellen skulle använda en separat parameter för varje allel skulle den potentiellt ha bättre kapacitet att lära sig denna information.

I arbetet baserades modellerna i en bayesiansk statistisk ram. En undersökning av andra förslagsfunktioner och deras påverkan på prediktionsförmågan skulle även kunna utföras genom andra val på standardavvikelsen eller förslagsfunktionen. Alternativa modeller kan även utvecklas inom en frekventistisk ram med konfidensintervall och signifikansnivå. I teorin är det även möjligt att använda neutrala nätverk som en prediktionsmodell, men i praktiken skulle detta kräva en betydligt större datamängd.

Priorifördelningarnas påverkan på modellernas prestanda kan även undersökas genom att variera

valet av prioriförderlningar. Dessutom kan hyperparametrar, såsom kovariansmatrisen för multinormalfördelningarna, justeras och därmed kan högre AUC-värden åstadkommas. För större datamängder skulle en adaptiv multinormal priorifunktion kunna användas. Denna skulle implementeras genom att först skapa en markovkedja för parametrarna genom att använda MCMC med en icke-proper likformig priorifunktion, för att sedan numeriskt beräkna kovariansmatrisen för parameterkedjan. Kovariansmatrisen skulle sedan användas för att skapa en multinormal priorifunktion.

6 Samhälleliga och etiska aspekter

Den utvecklade prediktionsmodellen har en inverkan på flera samhälleliga och etiska aspekter. Prediktionsmodellen är utvecklad för att huvudsakligen användas av rättsväsendet för att identifiera gärningsmän utifrån DNA-spår på brottsplatser. Detta kan bidra till effektivare och träffsäkrare brottsutredningar, vilket kan stärka rättssäkerheten och öka förtroendet för rättsväsendet. Trots denna samhällsnytta så finns det samtidigt en risk för felträffar, vilket kan negativt påverka enskilda individer genom att oskyldiga felaktigt pekas ut.

Det är av denna anledning viktigt, från ett samhälleligt perspektiv, att avgöra i vilket sammanhang som prediktionsmodellerna ska användas och vara medveten om modellens begränsningar. Eftersom prediktionsmodellen endast tar fram sannolikheter för fenotypiska egenskaper utifrån de genetiska markörerna, så finns det alltid en viss osäkerhet i prediktionen. Därmed är det viktigt att prediktionen inte ensamt kan användas som bevisning i en rättegång, utan att övrig bevisning alltid måste vägas in för att faktiskt kunna fälla en misstänkt för ett brott.

För att prediktionsmodellen i praktiken ska vara användbar för rättsväsendet behöver modellen ha hög träffsäkerhet för korrekta prediktioner. Träffsäkerheten kräver att den data som modellen bygger på är tillförlitlig när det gäller korrekthet, storlek av datamängd och bias. Bias i den ursprungliga datamängden kan leda till att individer i redan utsatta grupper i större grad blir utpekade av modellen, då modellen kan bli partisk för att göra vissa prediktioner. Detta kan därmed öka diskriminering och kränkningar för redan utsatta grupper i samhället. På lång sikt kan felaktiga utpekningar från undermåliga prediktionsmodeller påverka samhällets syn på rättssystemets trovärdighet.

Ett annat etiskt dilemma är integritetsfrågan. Användning av prediktionsmodellen kan möjligtvis innebära en inskränkning av individers skydd av deras genetiska information. Detta kan skapa en känsla av övervakning, vilket riskerar att minska tilliten för myndigheter och andra institutioner som använder modellen. Däremot kan det diskuteras om visuella egenskaper som utseende egentligen kan ses som privat data. Utseende är inte endast känt för individen i fråga, utan även för alla som sett personen.

En ytterligare central fråga är vem som får tillgång till prediktionsmodellen. Även om användandet av modellen begränsas till auktoriserade aktörer, så kommer med stor sannolikhet ytterligare aktörer kunna få tillgång till dessa eller liknande modeller. Detta innebär att det finns en risk för att privatpersoner, företag eller utländska myndigheter missbrukar användandet av modellen för att främja deras intressen. Om prediktionsmodellen vidare utvecklas till en perfekt modell som kan konstruera fantombilder från genetiska markörer, så kan missbruket exempelvis vara att företag samlar och säljer mer personlig data om privatpersoner, vilket ökar inskränkningarna i individers rätt till integritet och autonomi. Om modellerna skulle missbrukas av länder som inte följer Sveriges lagar och värderingar finns det även en risk för att den skulle kunna användas som ett verktyg för att diskriminera och förfölja redan utsatta minoritetsgrupper i samhället.

7 Slutsatser

I detta projekt utvecklades en statistisk modell för att predicera hår- och ögonfärg på individer baserat på en datamängd med uppsättning genetiska markörer, bestående av 85 individers genotyp och observerade fenotyper. Modellen byggdes med hjälp av bayesiansk statistik och logistisk regression, både i binär och multinomial form. För dessa användes två typer av priorifunktioner: en

icke-proper likformig priorifunktion och en multinormal priorifunktion. Dessa modellers prestanda jämfördes med hjälp av ROC-kurvor samt AUC-värden.

Samtliga modeller fick låga totala AUC-värden. Detta förklaras av flera faktorer, däribland hur parametrarna tilldelades, storleken av datamängden samt hur den klassificerades. Att tilldela en separat parameter för varje SNP medför att modellen introducerar ett artificiellt numeriskt samband, vilket i sin tur försvårar upptäckten av mönster i hur allelerna är fördelade inom varje SNP. Den begränsade storleken på datamängden innebär att modellen inte lär sig relevanta mönster som skiljer klasserna åt. Den ospecifika klassificeringen av datamängden, speciellt färgen intermediär, kan även ha haft inverkan i det slutgiltiga resultatet. Detta då variationen i allelerna kan ha varit för stor för modellen att upptäcka. Trots resultaten antyds det att ett samband mellan fenotypen och den genetiska uppsättningen existerar, även om modellerna i sin helhet inte är lämpliga för användning inom forensisk verksamhet.

Referenser

[1] Statens medicinsk-etiska råd (SMER), *Kort om DNA och brottsutredning*, 2021. URL: https://smer.se/wp-content/uploads/2021/04/smer_dna_brott_tga.pdf, Hämtad: 30 januari 2025.

[2] N. Hagen, "I gränslandet mellan genotyp och fenotyp. Motsägelser i samband med prediktiv genetisk testning", *Socialmedicinsk tidskrift*, årg. 88, nr 3, s. 266–272, 2011. URL: https://socialmedicinsktidskrift.se/index.php/smt/article/download/788/636/0, Hämtad 31 januari 2025.

[3] D. P. Clark, *Molecular Biology*, 1. utg. Elsevier Science & Technology Books, juni 2005, ISBN: 9780123785893. URL: https://shop.elsevier.com/books/molecular-biology/clark/978-0-12-378589-3.

[4] A. A. Komar, *Single Nucleotide Polymorphisms: Methods and Protocols* (Methods in Molecular Biology), 2. utg. Totowa, NJ: Humana Press, 2009, vol. 212. DOI: https://doi.org/10.1007/978-1-60327-411-1.

[5] S. Bader, *A Guide to Forensic DNA Profiling*, 1. utg. John Wiley & Sons, Incorporated, mars 2016, ISBN: 9781118751527.

[6] M. A. Farley, *Forensic DNA Technology*, 1. utg. Boca Raton, FL, USA: Taylor & Francis Group, 2017. DOI: https://doi.org/10.1201/9781351072120.

[7] P. R. Haddrill, "Developments in forensic DNA analysis", *Emerging Topics in Life Sciences*, årg. 5, nr 3, s. 381–393, 2021, ISSN: 2397-8562. DOI: https://doi.org/10.1042/ETLS20200304.

[8] M. Wallin, "Användning av ny DNA-teknik vid brottsbekämpning för att förutsäga människors ögon-, hår- och hudfärg", *Bioscience Explained*, årg. 9, nr 1, 2016, Göteborgs universitet, Institutionen för biologi och miljövetenskap. URL: https://gup.ub.gu.se/publication/277670.

[9] K. Junker, A. Staadig, M. Sidstedt, A. Tillmar och J. Hedman, "Phenotype prediction accuracy – A Swedish perspective", *Forensic Science International: Genetics Supplement Series*, årg. 7, nr 1, s. 384–386, dec. 2019. DOI: https://doi.org/10.1016/j.fsigss.2019.10.022.

[10] P. M. Lee, *Bayesian Statistics: An Introduction* (New York Academy of Sciences Series), 1. utg. Hoboken, NJ: John Wiley & Sons, Incorporated, 2012, ISBN: 9781118332573.

[11] K. David G och M. Klein, *Logistic Regression, A Self Learning Text* (Statistics for Biology and Health), 3. utg. Springer, 2010, ISBN: 978-1-4419-1741-6. DOI: https://doi.org/10.1007/978-1-4419-1742-3.

[12] D. W. H. Jr., S. Lemeshow och R. X. Sturdivant, *Applied Logistic Regression* (Wiley Series in Probability and Statistics). John Wiley & Sons, Inc., 2013, ISBN: 9780470582473. DOI: https://doi.org/10.1002/9781118548387.

[13] H. Richard M och B. Holland, *Statistical Analysis and Data Display, An Intermediate Course with Examples in R* (Springer Texts in Statistics). Springer, 2015, Appendix G, ISBN: 978-1-4939-2122-5. DOI: https://doi.org/10.1007/978-1-4939-2122-5.

[14] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, English, 1. utg. Oxford University Press, Incorporated, 2001, ISBN: 9780199671229. DOI: https://doi.org/10.1093/oso/9780198507659.001.0001.

[15] R. P. Dobrow, *Introduction to stochastic processes with R*. Nashville, TN: John Wiley & Sons, febr. 2016. DOI: 10.1002/9781118740712.

[16] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari och D. B. Rubin, *Bayesian Data Analysis*, 3. utg. Chapman och Hall/CRC, 2025. URL: https://sites.stat.columbia.edu/gelman/book/.

[17] A. Gut, *An Intermediate Course in Probability* (Springer Texts in Statistics), 2. utg. Springer New York, NY, 2009. DOI: https://doi.org/10.1007/978-1-4419-0162-0. URL: https://link.springer.com/book/10.1007/978-1-4419-0162-0.

[18] S. M. Ross, "Distributions of Sampling Statistics", i *Introduction to Probability and Statistics for Engineers and Scientists*, 5. utg., Författarens affiliering: University of Southern California, Los Angeles, USA., Oxford: Academic Press, 2014, kap. 6, s. 207–233, ISBN: 978-0-12-394811-3. DOI: 10.1016/B978-0-12-394811-3.50006-X, Hämtad: 4 maj 2025.

[19] J. Tacq, "Multivariate Normal Distribution", i *International Encyclopedia of Education (Third Edition)*, 3. utg., Oxford: Elsevier, 2010, s. 332–338, ISBN: 978-0-08-044894-7. DOI: https://doi.org/10.1016/B978-0-08-044894-7.01351-8, Hämtad: 4 maj 2025.

[20] M. McDonough, *Cross-validation*, Encyclopedia Britannica Academic. Senast reviderad av Erik Gregersen, 12 september 2023. URL: https://www.britannica.com/technology/cross-validation-computer-science, Hämtad: 26 april 2025.

[21] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, årg. 27, nr 8, s. 861–874, 2006, ROC Analysis in Pattern Recognition, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[22] F. Provost och P. Domingos, "Well-trained PETs: Improving probability estimation trees", *Raport instytutowy IS-00-04, Stern School of Business, New York University*, årg. 1, okt. 2000. URL: https://pages.stern.nyu.edu/~fprovost/Papers/pet-wp.pdf, Hämtad: 9 maj 2025.

[23] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians", *Korean Journal of Anesthesiology*, årg. 75, nr 1, s. 25–36, jan. 2022. DOI: https://doi.org/10.4097/kja.21209.

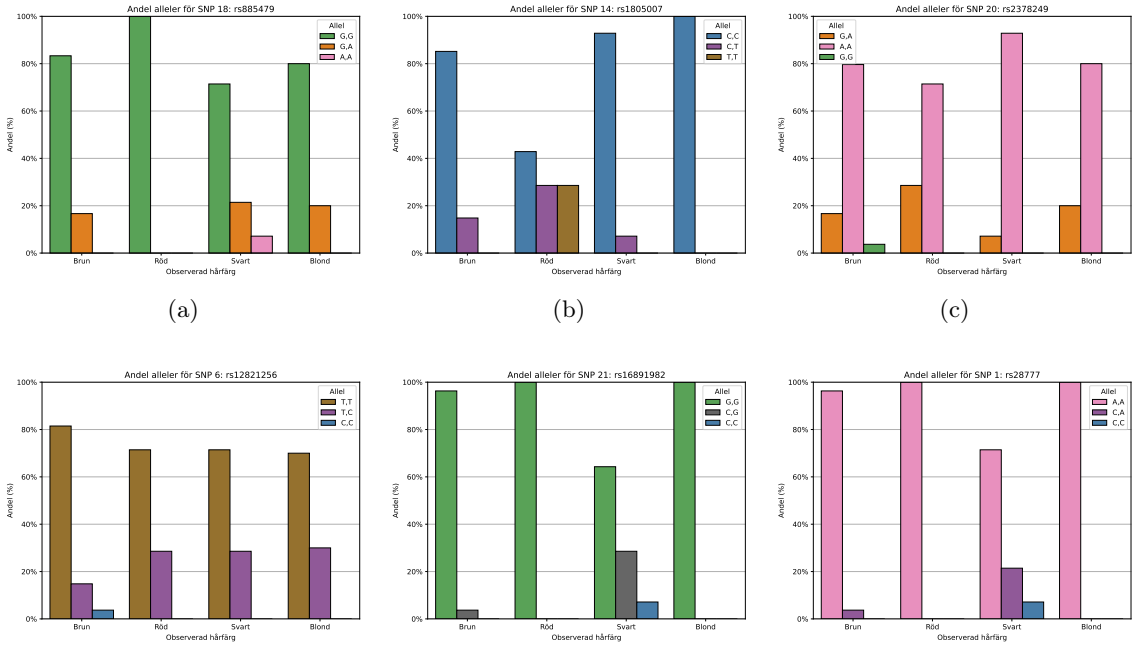
[24] D. Rios Insua, F. Ruggeri och M. P. Wiper, *Bayesian Analysis of Stochastic Process Models* (Wiley Series in Probability and Statistics). John Wiley & Sons, Ltd, 2012. DOI: https://doi.org/10.1002/9780470975916.

8 AI-användning

AI-användning har begränsats i arbetet. AI användes för att ge förslag på hur tabellerna i förordet kan utformas, men användes inte för att generera hela tabellerna. På detaljnivå har AI i vissa fall använts för att ge förslag på omformuleringar av redan färdigskrivna meningar samt grammatik- och stavningskontroll, men detta har gjorts väldigt sparsamt. AI har inte använts för att generera hela text- eller kodstycken. AI användes inte under arbetet med att utforma prediktionsmodellen. För att generera plottar har AI använts som en "sökmotor", för att till exempel söka upp hur specifika paket fungerar. Den AI som har använts har varit GPT-4o mini och den inbyggda GPT modellen för overleaf (OpenAI).

A Datavisualisering

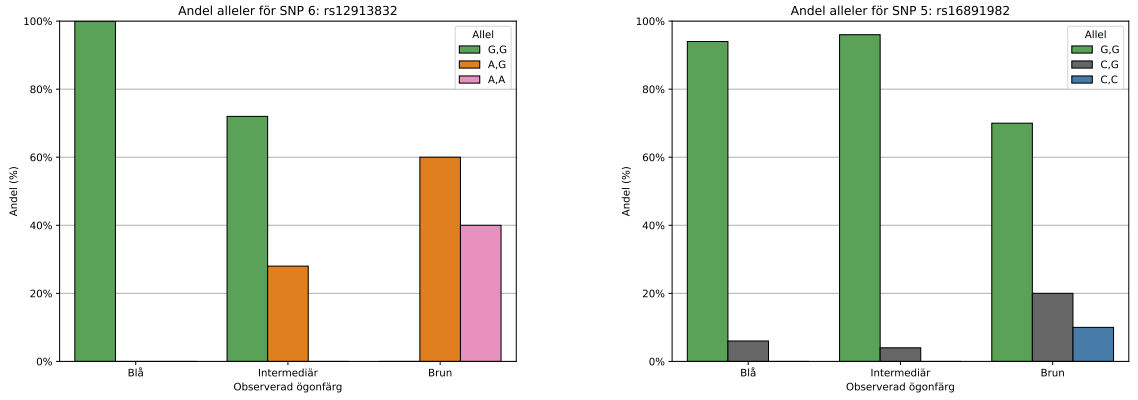
Figur 8 och 9 är de SNP:er som uppvisar samma beteende som de i figur 2, där en allel är unik för en observation.



(a) (b) (c)

(d) (e) (f)

Figur 8: Alla SNP:er där en allel är unik för en hårfärgsobservation. Figurerna visar andelen alleler, i procent, för varje observation. I (a) ses AA endast hos individer med svart hår, (b) visar att TT är unikt för rött hår, (c) och (d) visar att GG respektive CC förekommer enbart vid brun hårfärg, medan CC i både (e) och (f) är specifik för svart hår.



(a) (b)

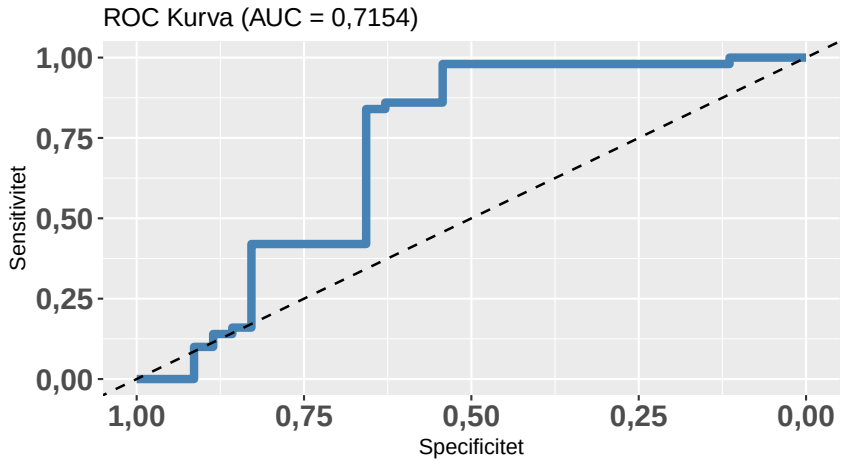
Figur 9: Alla SNP:er där en allel är unik för en ögonobservation. Figurerna visar andelen alleler, i procent, för varje observation. I (a) och (b) förkommer endast allelen AA respektive CC endast för brun ögonfärg.

B Figurer och tabeller

Figurerna nedan illustrerar resultaten som modellerna skapade.

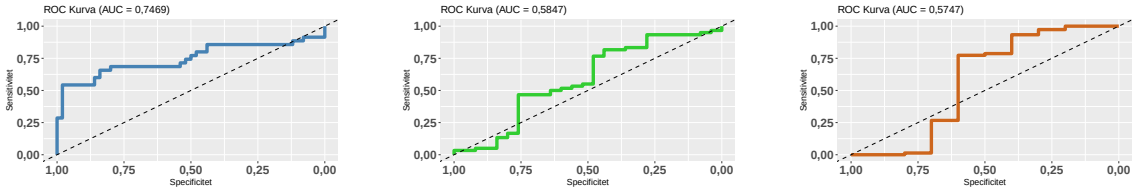
B.1 ROC-kurvor

ROC-kurvan för binär ögonfärg med multinormal priorifunktion med tillhörande AUC-värde presenteras i figur 10, där AUC beräknades till 0,7154.



Figur 10: ROC-kurva och tillhörande AUC för binär ögonfärg, vid användning av en multinormal priorifunktion. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

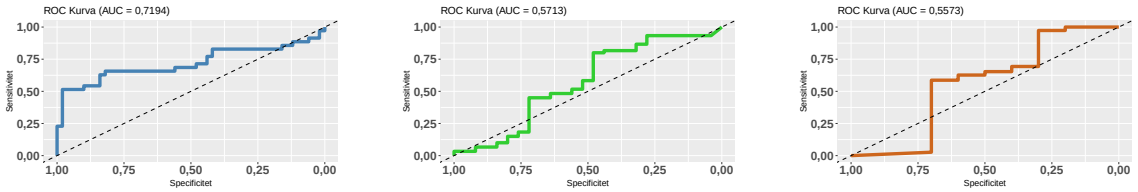
ROC-kurvorna för modellen för multinomial ögonfärg med likformig priorifunktion återges i figur 11, där AUC-värdena för blå, intermediär och brun ögonfärg var 0,7469, 0,5847 respektive 0,5747. Det sammanvägda AUC-värdet beräknades enligt ekvation (10) och uppgick till 0,6789.



(a) (b) (c)

Figur 11: ROC-kurvor för den multinomiala modellen för ögonfärg vid användande av icke-proper likformig prior. Kurvan (a) tillhör blå ögonfärg, (b) intermediär och (c) brun som positivt utfall. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

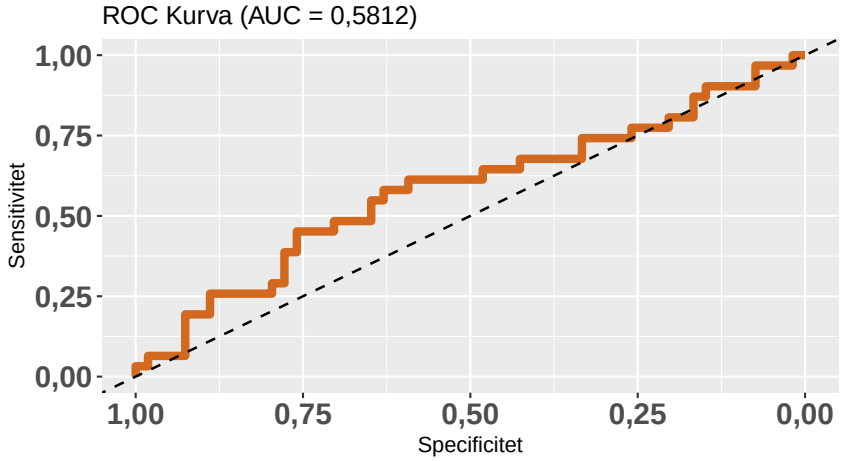
ROC-kurvorna för multinomial ögonfärg med multinormal priorifunktion illustreras i figur 12. Figur 12a, 12b och 12c representerar ROC-kurvorna för respektive ögonfärg. AUC för blå, intermediär och brun ögonfärg som positivt utfall var 0,7194, 0,5713 och 0,5573, vilket ger det totala AUC-värdet på 0,6568 enligt ekvation (10).



(a) (b) (c)

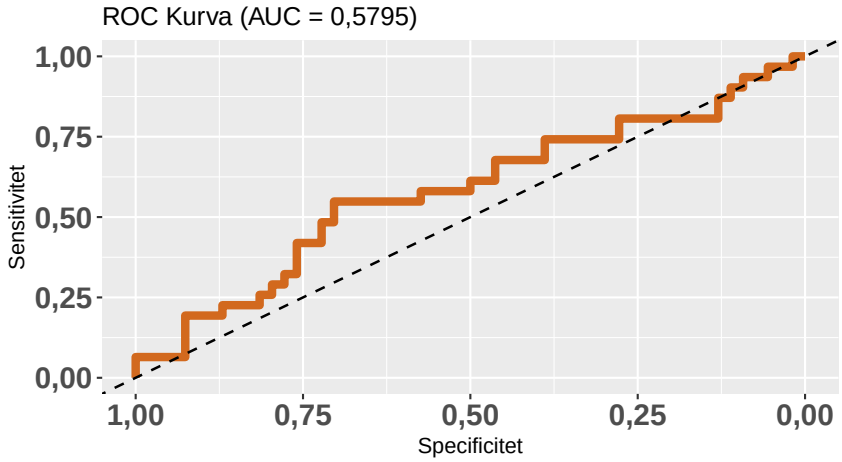
Figur 12: ROC-kurvor för den multinomiala modellen för ögonfärg vid användande av multinormal likformig prior. Kurvan (a) tillhör blå ögonfärg, (b) intermediär och (c) brun som positivt utfall. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

AUC-värdet för modellen för binär hårfärg med likformig priorifunktion beräknades till 0,5812 enligt figur 33.



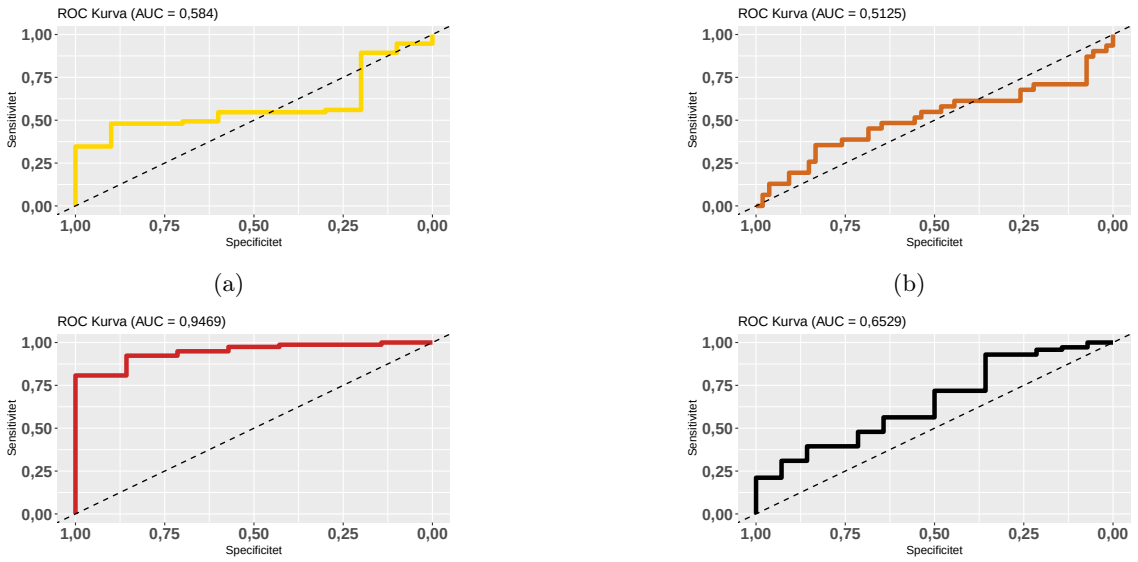
Figur 13: ROC-kurva och tillhörande AUC för binär hårfärg, vid användning av en icke-proper likformig priorifunktion. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

För modellen för binär hårfärg med en multinormal priorifunktion beräknades AUC-värdet till 0,5795, vilket redovisas i figur 14.



Figur 14: ROC-kurva och tillhörande AUC för binär hårfärg, vid användning av en multinomial priorifunktion. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

För modellen för multinomial hårfärg med icke-proper likformig priorifunktion beräknades det totala AUC-värdet till 0,5798 med hjälp av ekvation (10). Varje fenotyp hade AUC-värdena 0,5840, 0,5125, 0,9469 och 0,6529 för blond, brun, röd respektive svart. Detta redovisas i figur 15.



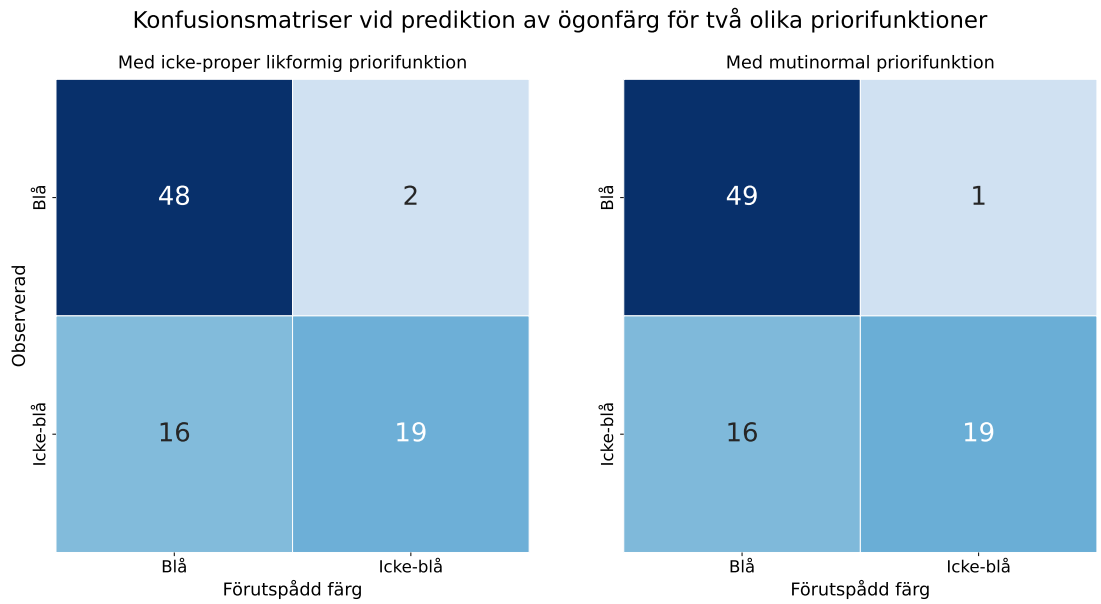
(a) (b)

(c) (d)

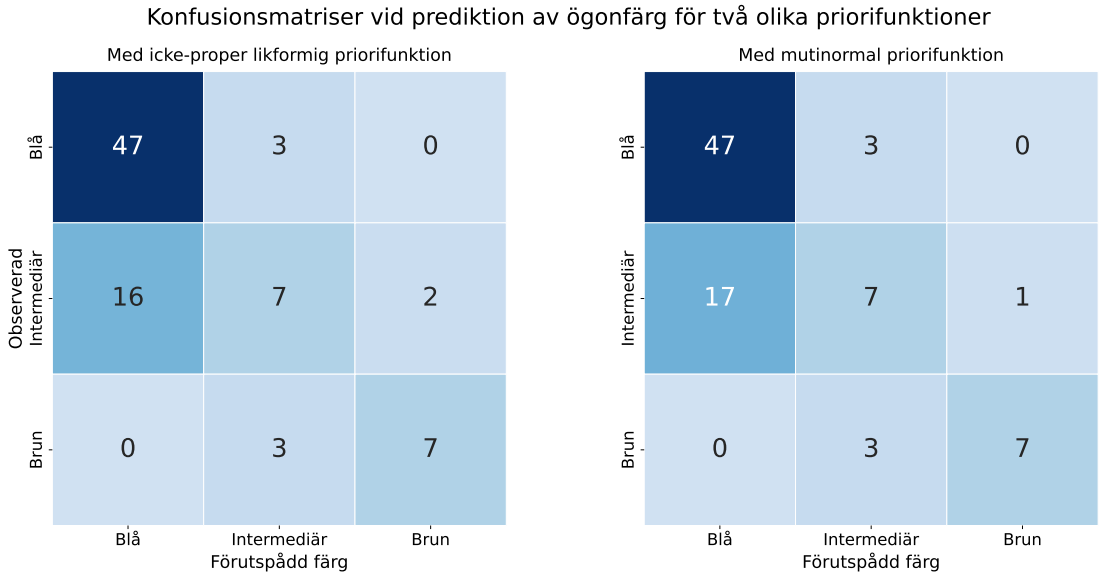
Figur 15: ROC-kurvor för den multinomiala modellen för hårfärg vid användande av icke-proper likformig prior. Kurvan (a) tillhör blond hårfärg, (b) brun, (c) röd och (d) svart som positivt utfall. Den streckade diagonalen representerar en slumpmässig klassificering och används i detta fall som referens.

B.2 Konfusionsmatriser

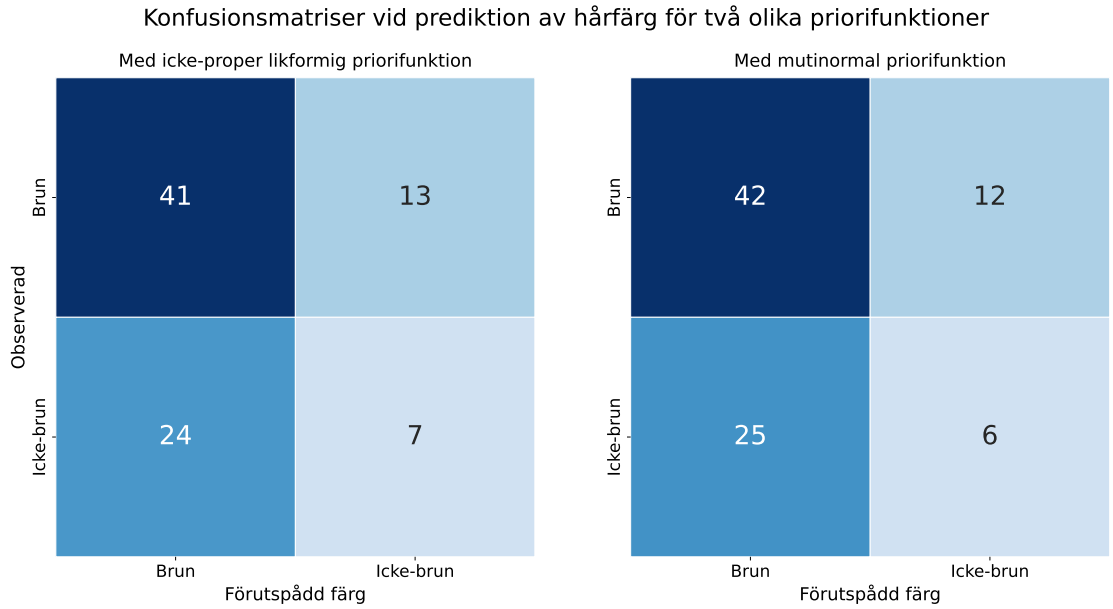
Nedan presenteras konfusionsmatriserna för alla modeller med båda priorifunktionerna. Figur 16 visar konfusionsmatriserna för binär ögonfärg och figur 17 för den multinomiala modellen som förutspådde förekomsten av blå, intermediär eller brun ögonfärg. I figurer 18 och 19 visas konfusionsmatriserna för de binära respektive multinomiala hårfärgsmodellerna.



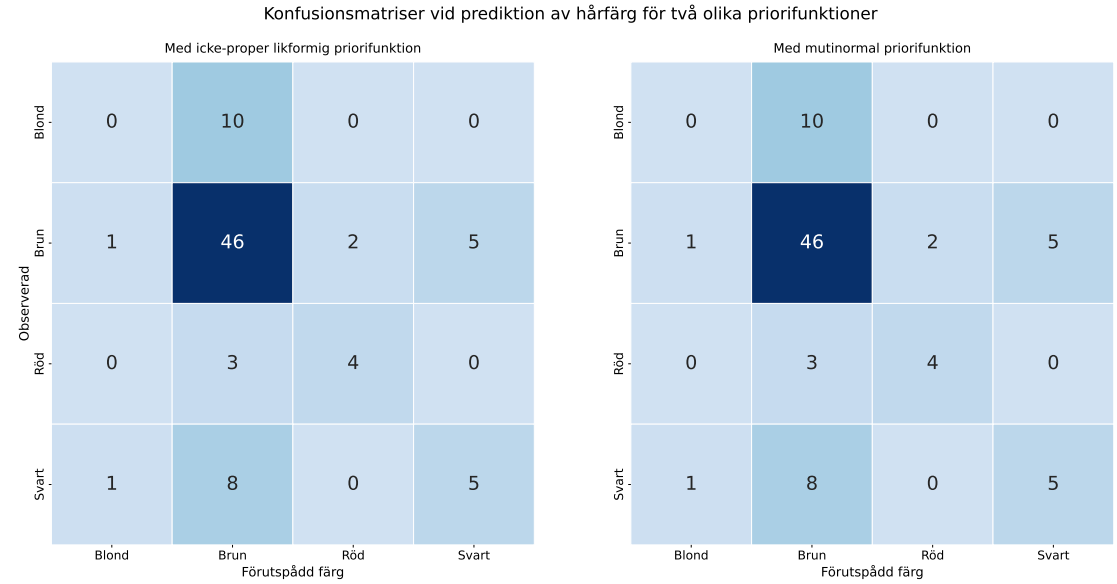
Figur 16: Konfusionsmatris för binär prediktion av ögonfärg. Prediktionen har gjorts med två olika priorifunktioner.



Figur 17: Konfusionsmatris för multinomial prediktion av ögonfärg. Prediktionen har gjorts med två olika priorifunktioner.



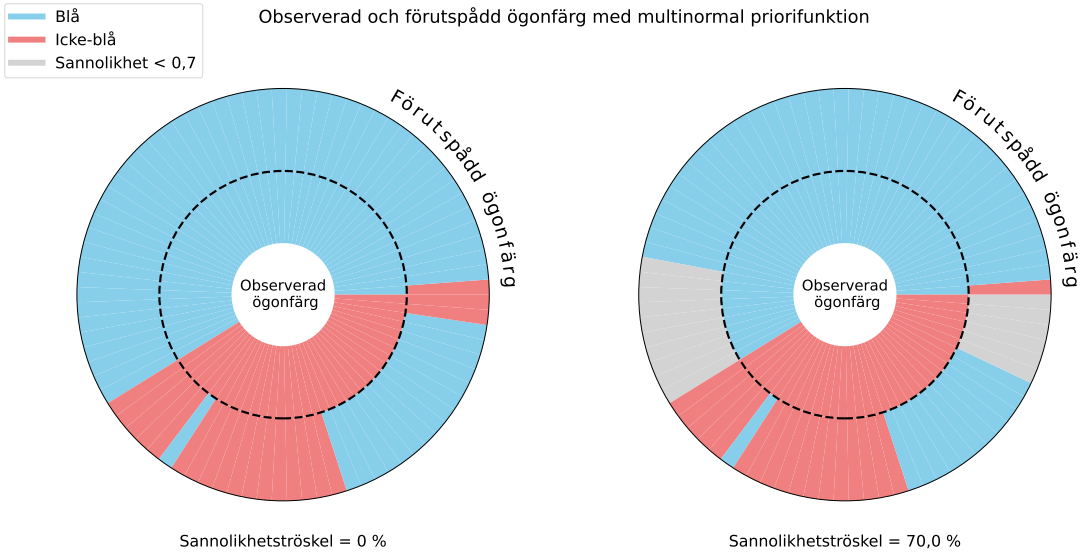
Figur 18: Konfusionsmatris för binär prediktion av hårfärg. Prediktionen har gjorts med två olika priorifunktioner.



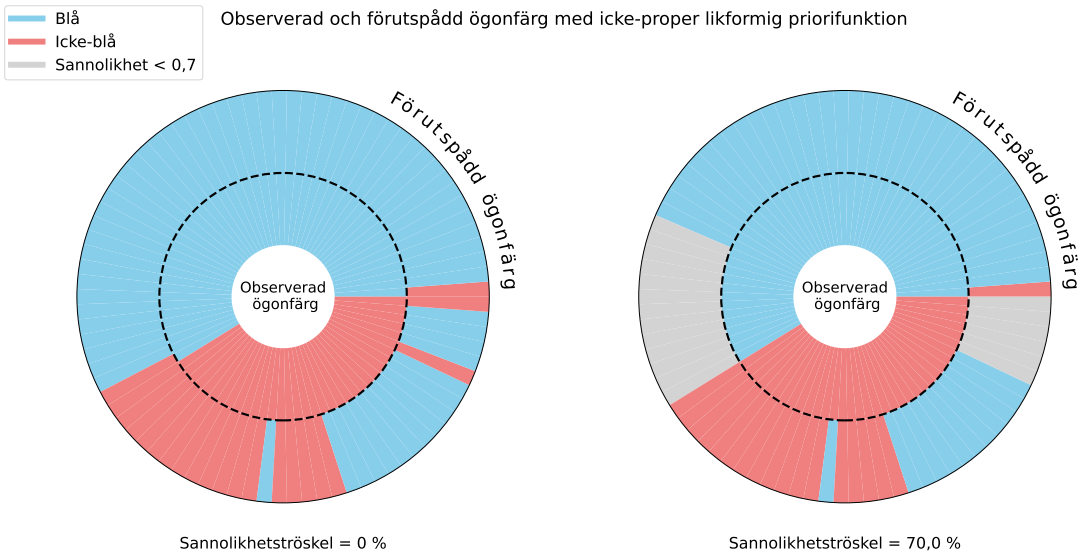
Figur 19: Konfusionsmatris för multinomial prediktion av hårfärg. Prediktionen har gjorts med två olika priorifunktioner.

B.3 Figurer för modellernas prediktioner

I avsnittet nedan illustreras resultaten för de olika prediktionsmodellerna i figurerna 20, 21, 22 och 23. Samtliga figurer visar de observerade fenotyperna i den inre cirkelskivan samt de färger som modellerna har predicerat i den yttre cirkelskivan.

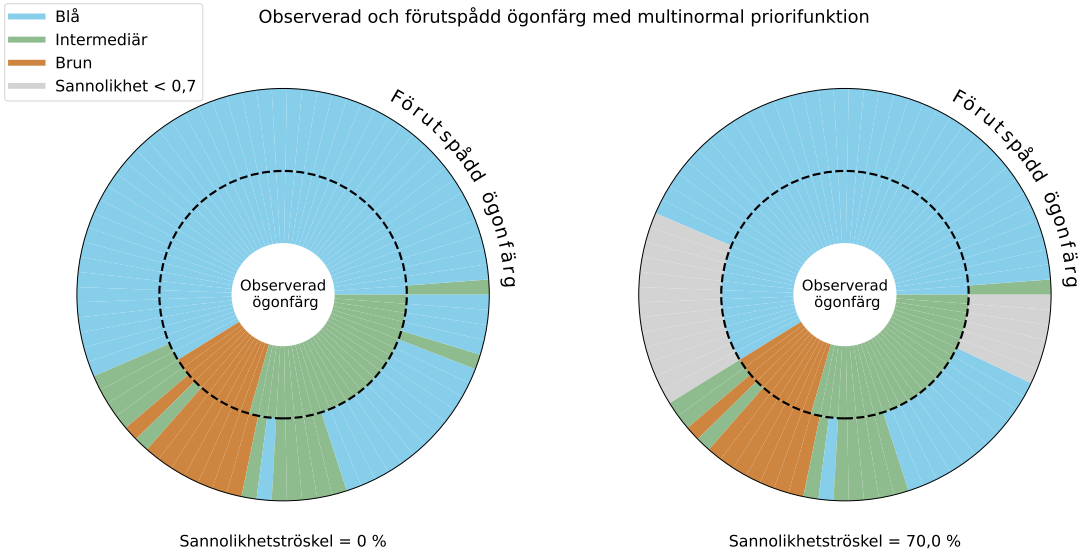


(a)

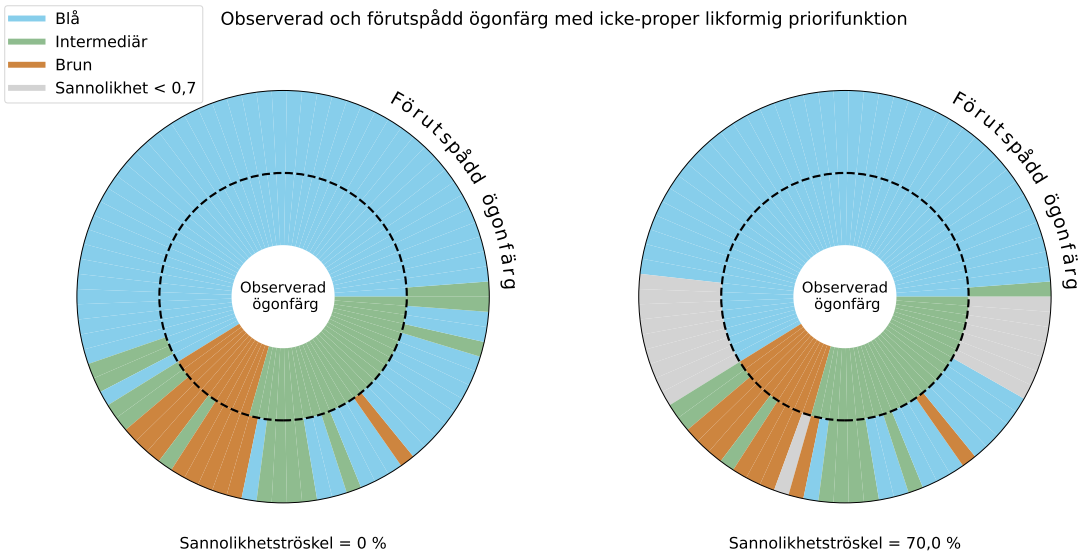


(b)

Figur 20: Resultat av den förutspådda ögonfärgen jämfört med den observerade ögonfärgen för en binär prediktionsmodell. För prediktionen användes i (a) en multinormal priorifunktion medan ingen användes i (b). Det högra cirkeldiagrammet i både (a) och (b) har en sannolikhetströskel på 70%.

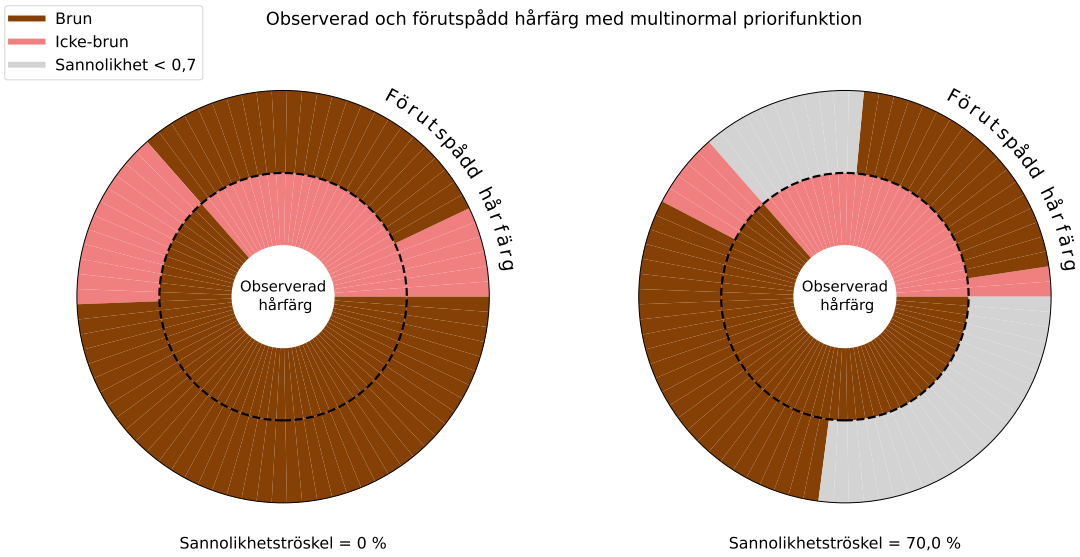


(a)

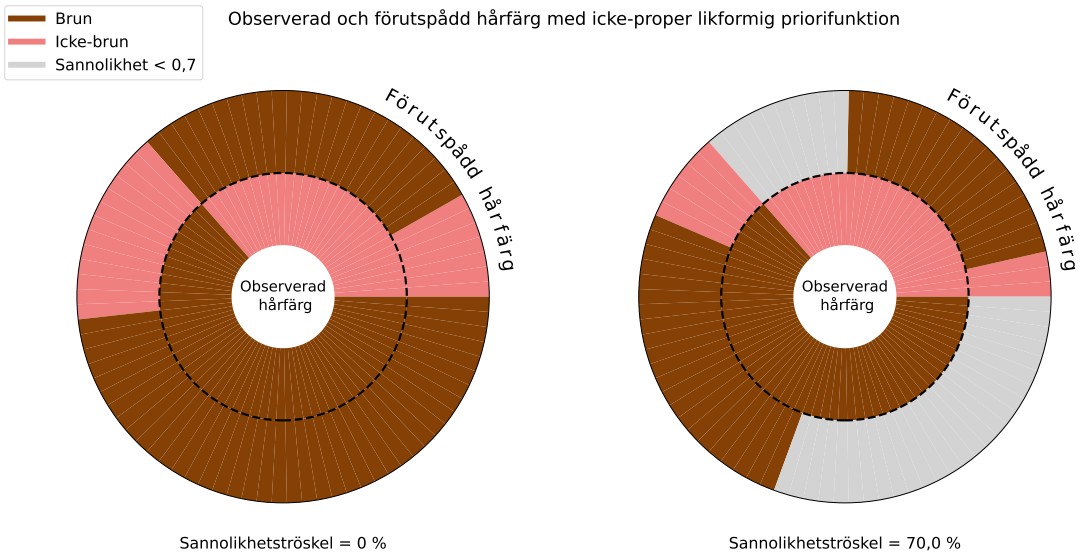


(b)

Figur 21: Resultat av den förutspådda ögonfärgen jämfört med den observerade ögonfärgen. För prediktionen användes i (a) en multinormal priorifunktion medan ingen användes i (b). Det högra cirkeldiagrammet i både (a) och (b) har en sannolikhetströskel på 70%.

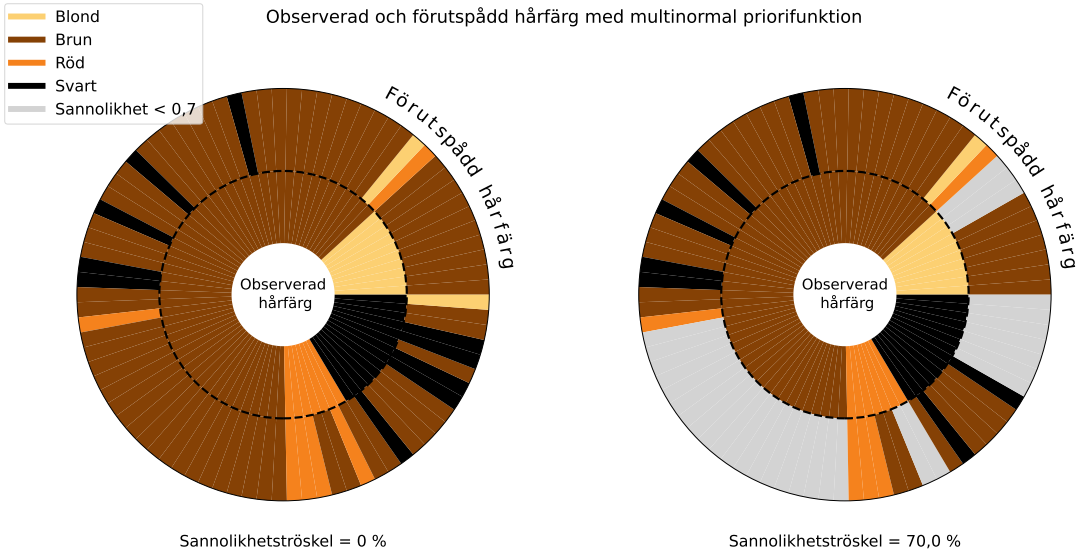


(a)

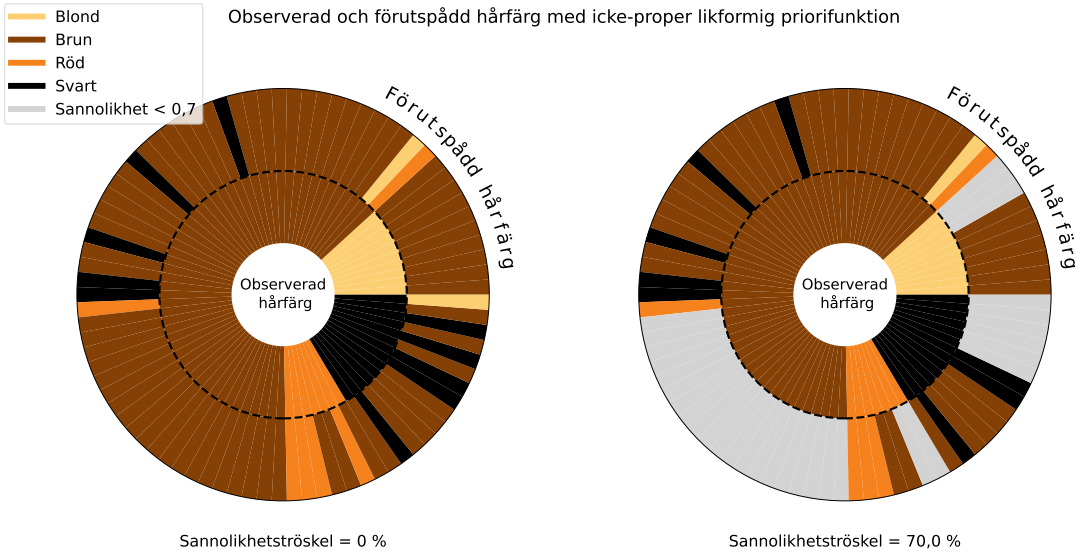


(b)

Figur 22: Resultat av den förutspådda hårfärgen jämfört med den observerade hårfärgen för en binär prediktionsmodell. För prediktionen användes i (a) en multinormal priorifunktion medan ingen användes i (b). Det högra cirkeldiagrammet i både (a) och (b) har en sannolikhetströskel på 70%.



(a)

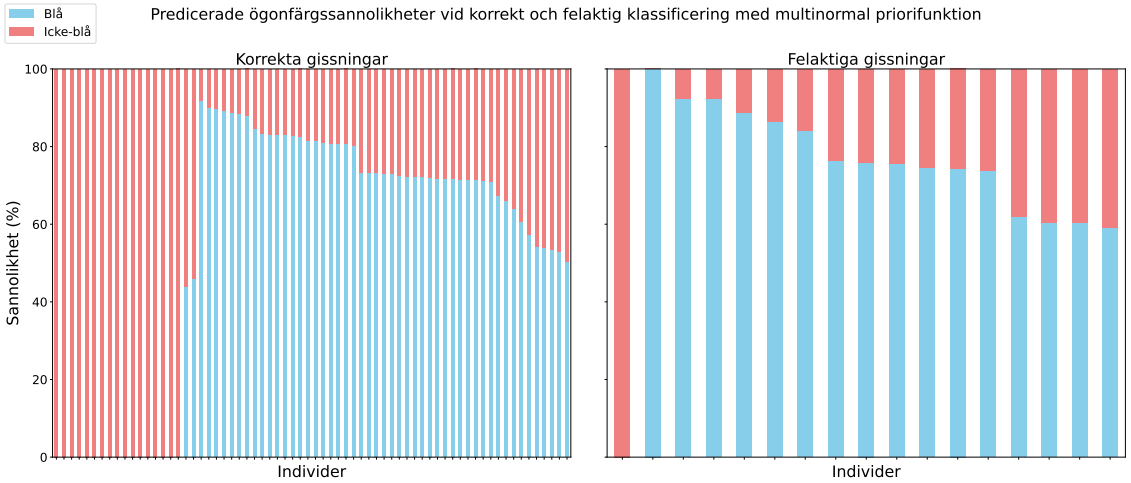


(b)

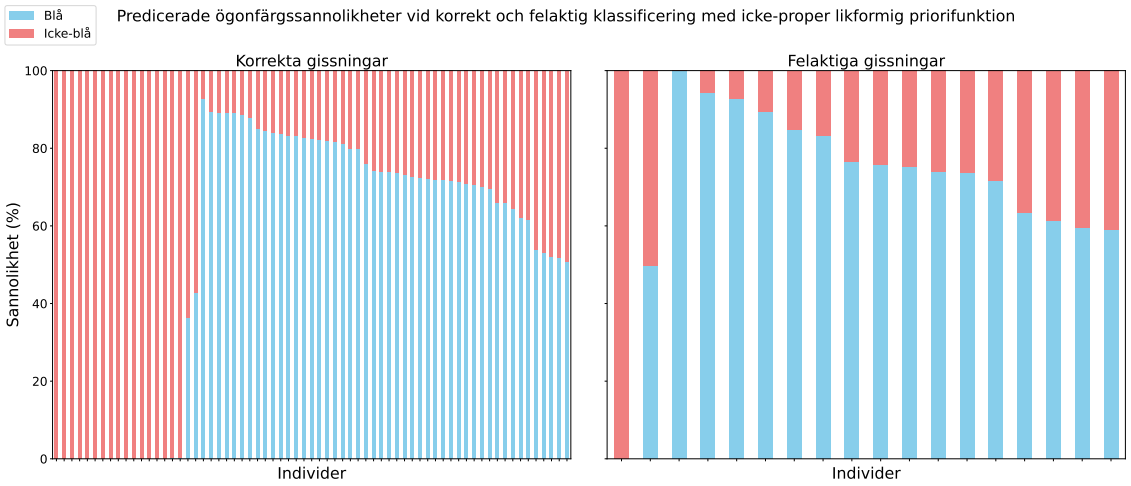
Figur 23: Resultat av den förutspådda hårfärgen jämfört med den observerade hårfärgen. För prediktionen användes i (a) en multinormal priorifunktion medan ingen användes i (b). Det högra cirkeldiagrammet i både (a) och (b) har en sannolikhetströskel på 70%.

B.4 Figurer över prediktionsmodellernas sannolikhetsfördelning

Figurerna 24, 25, 26 och 27 i nedanstående avsnitt visar resultatet över de predicerade sannolikheterna. Procentsatserna för respektive gissning är staplade på varandra i figurerna. Figurerna är uppdelade i korrekta gissningar och felaktiga gissningar.

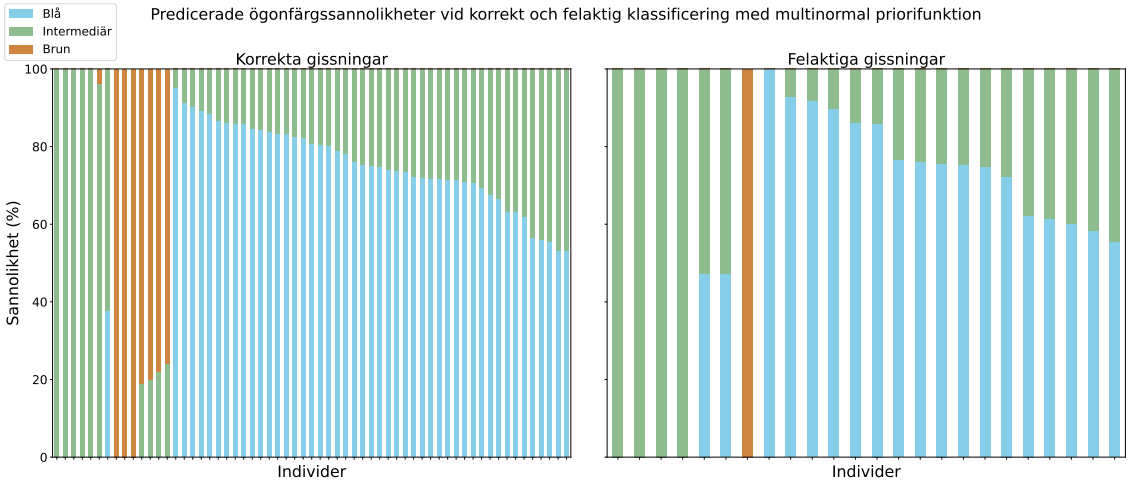


(a)

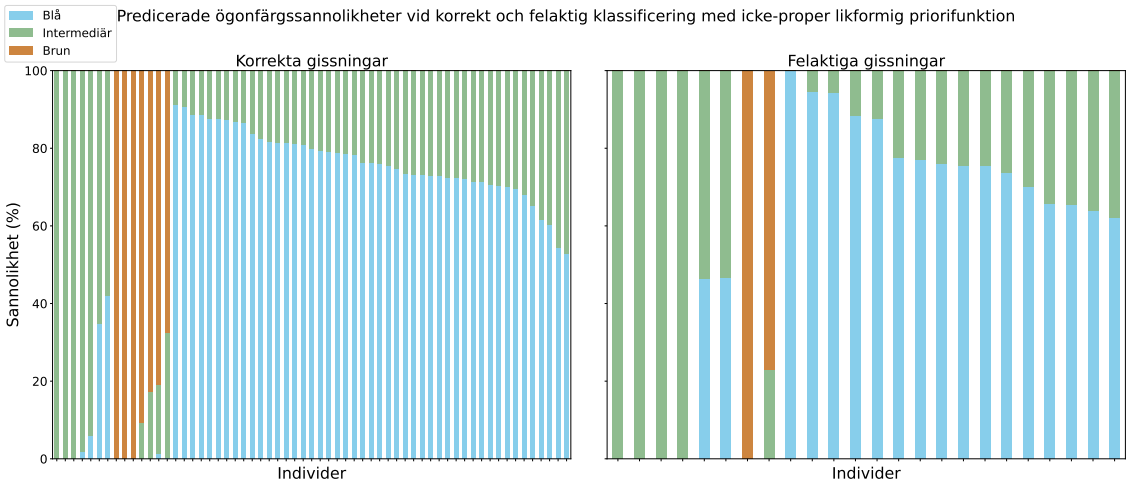


(b)

Figur 24: Resultat över sannolikheterna för respektive ögonfärg efter prediktion. Figur (a) visar sannolikheterna för prediktionsmodellen med en multinormal priorifunktion och (b) med en icke-proper likformig priorifunktion.

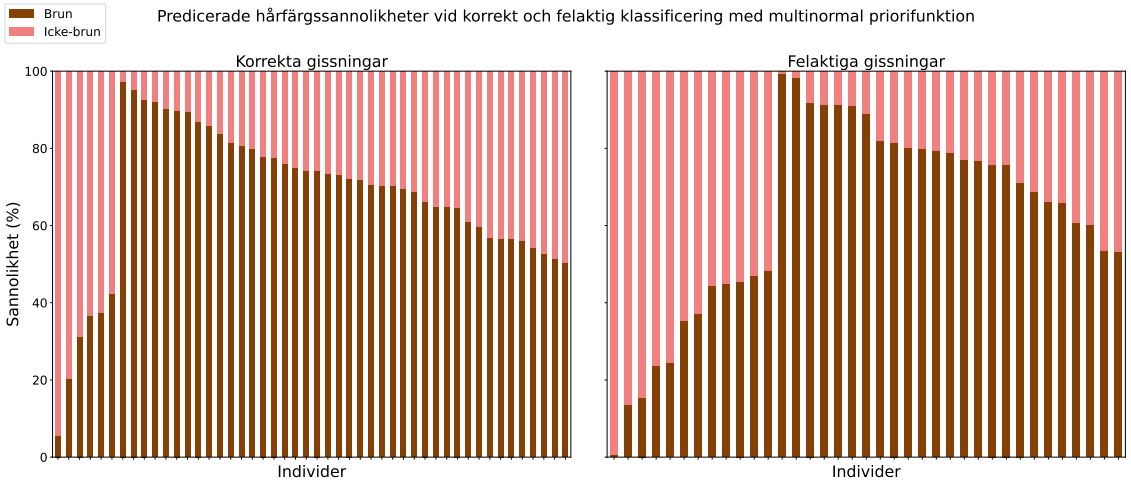


(a)

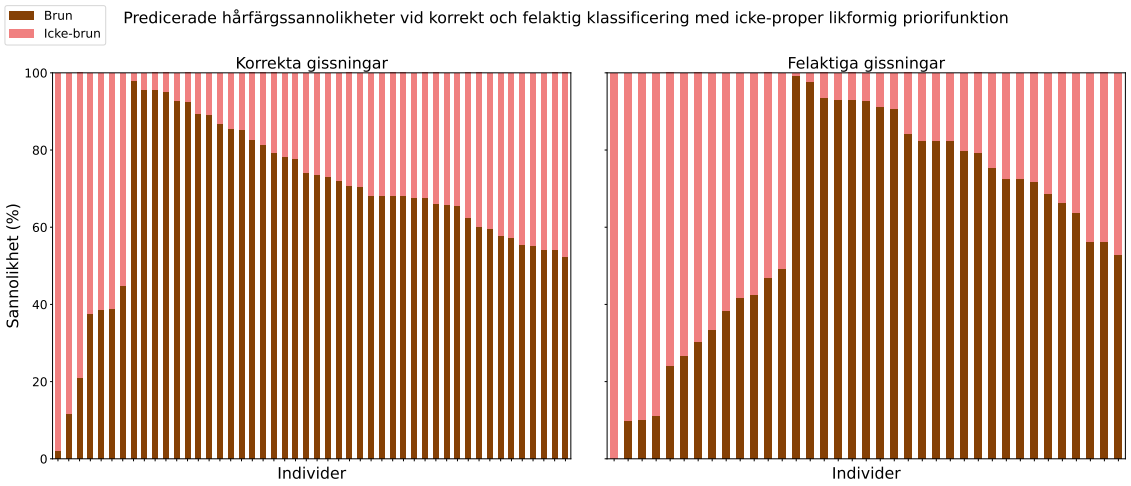


(b)

Figur 25: Resultat över sannolikheterna för respektive ögonfärg efter prediktion. Figur (a) visar sannolikheterna för prediktionsmodellen med en multinormal priorifunktion och (b) med en icke-proper likformig priorifunktion.

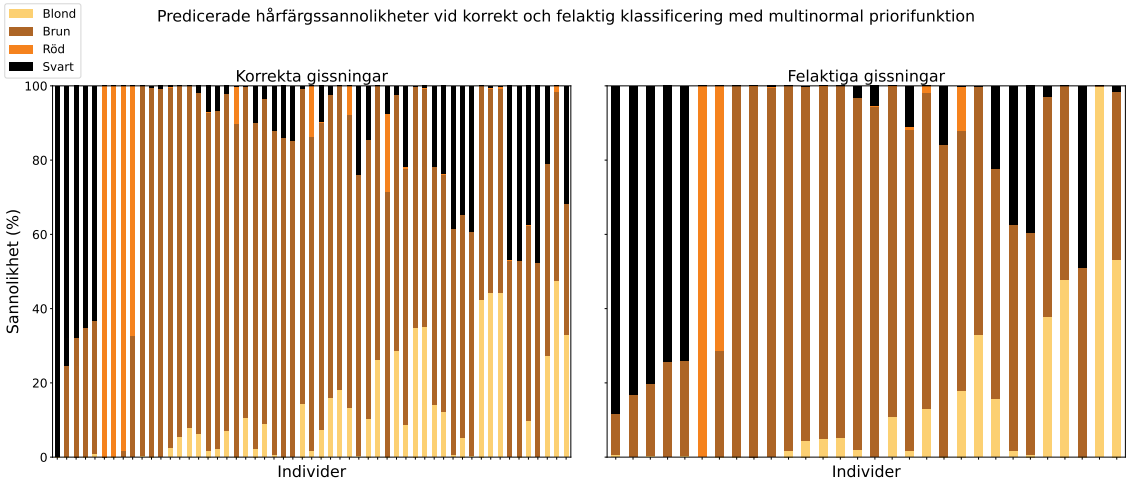


(a)

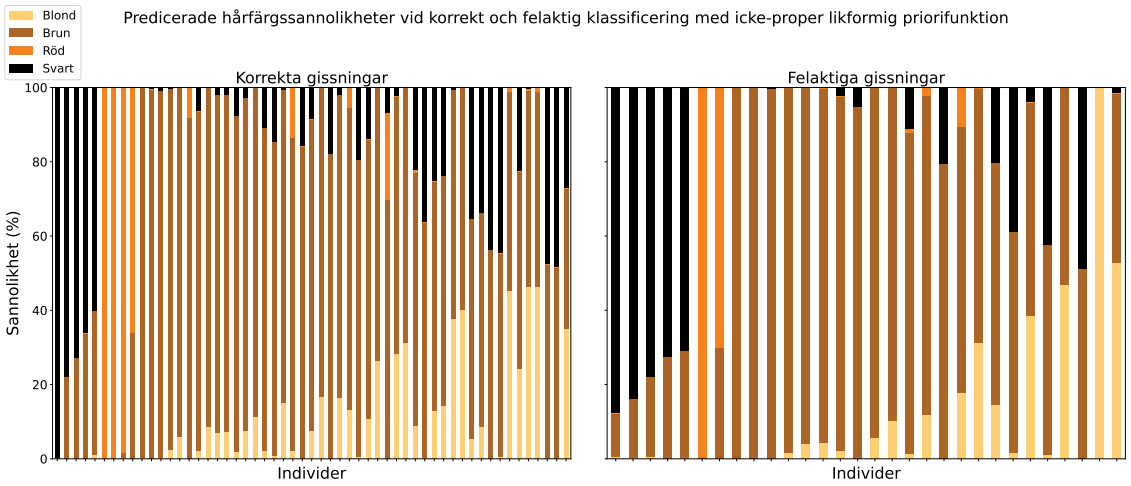


(b)

Figur 26: Resultat över sannolikheterna för respektive hårfärg efter prediktion. Figur (a) visar sannolikheterna för prediktionsmodellen med en multinormal priorifunktion och (b) med en icke-proper likformig priorifunktion.



(a)

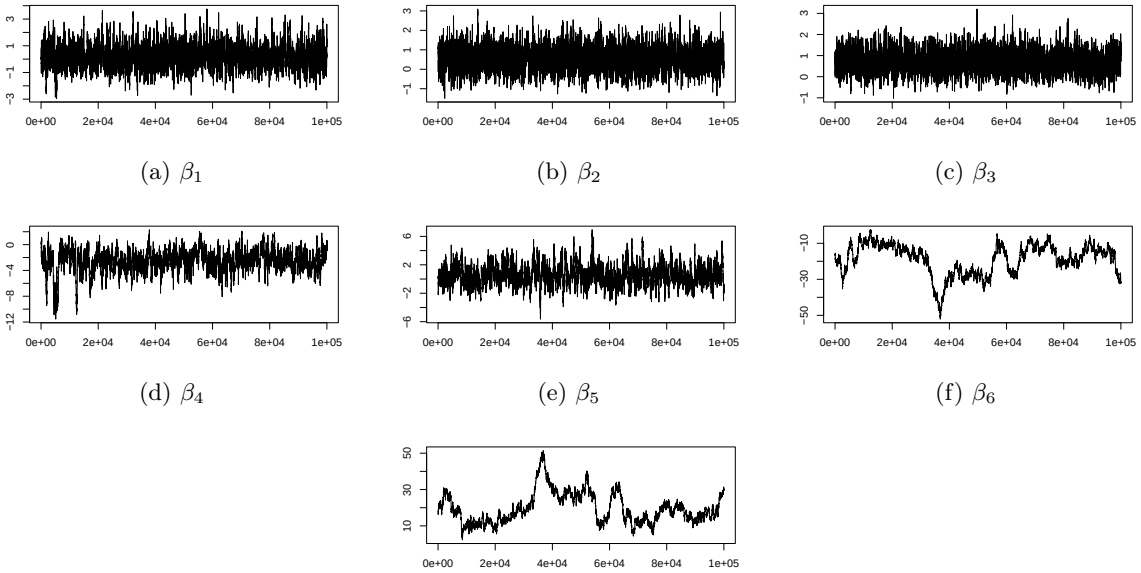


(b)

Figur 27: Resultat över sannolikheterna för respektive hårfärg efter prediktion. Figur (a) visar sannolikheterna för prediktionsmodellen med en multinormal priorifunktion och (b) med en icke-proper likformig priorifunktion.

B.5 Figurer för parametrarnas konvergens

Figurerna 28, 29, 30, 31, 32, 33 och 34 visar markovkedjorna för parametrarna som producerades av de olika modellerna.

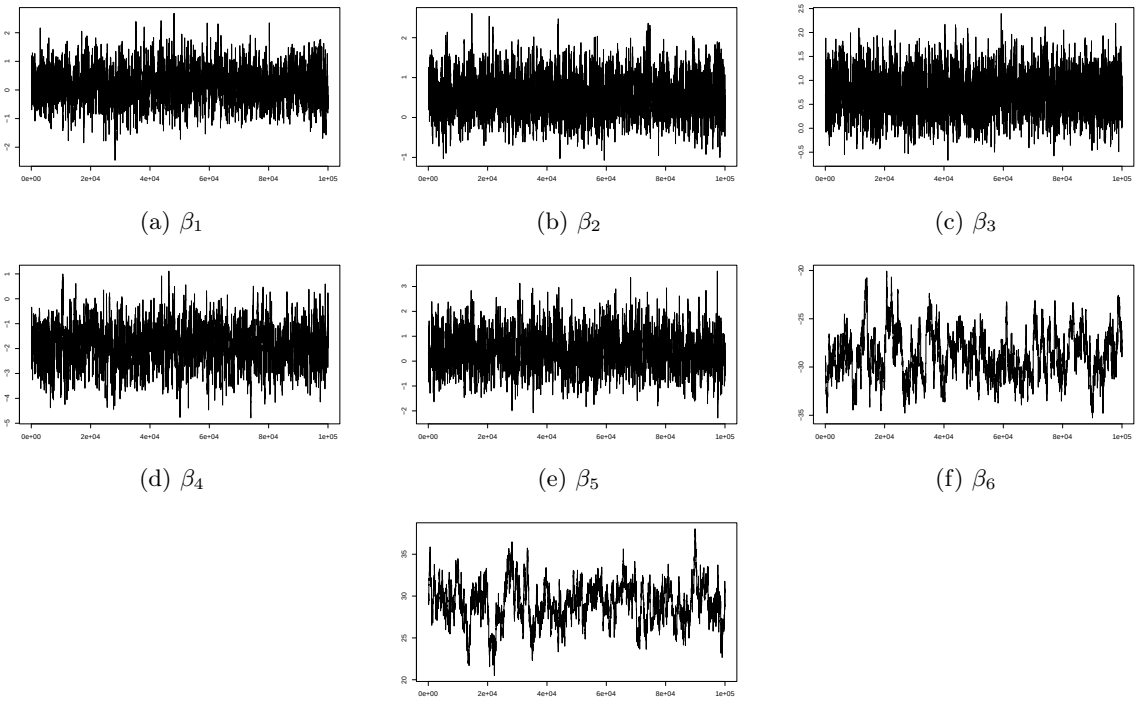


(a) β_1 (b) β_2 (c) β_3

(d) β_4 (e) β_5 (f) β_6

(g) α

Figur 28: Markovkedjor producerade av den binära modellen för ögonfärg vid användande av en prior som är proportionell mot den likformiga sannolikhetsfördelningen.

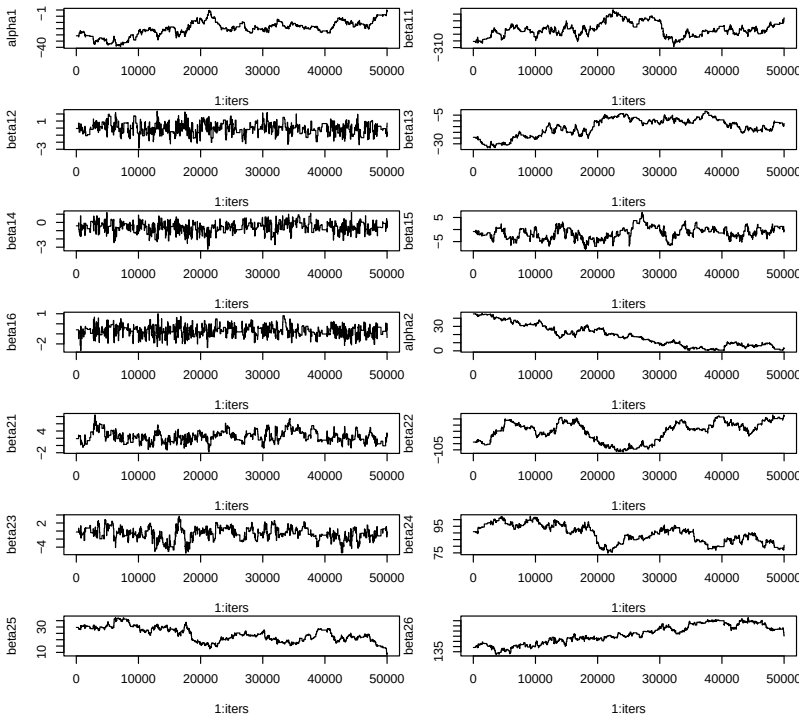


(a) β_1 (b) β_2 (c) β_3

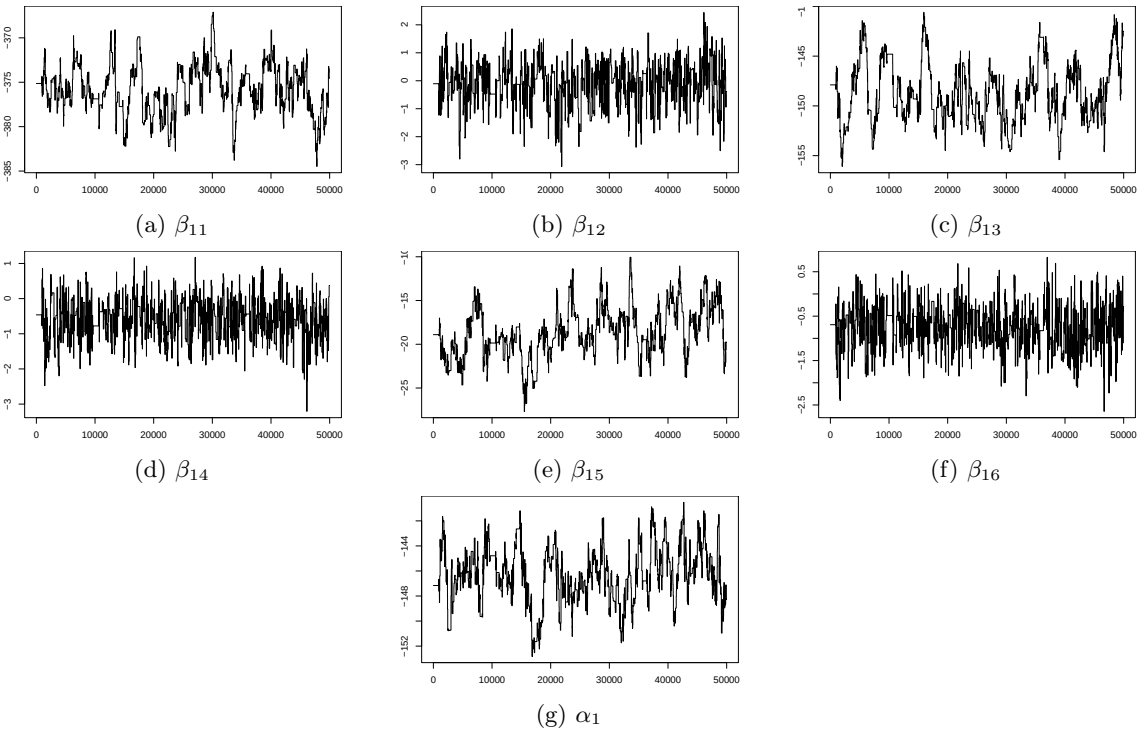
(d) β_4 (e) β_5 (f) β_6

(g) α

Figur 29: Markovkedjor producerade av den binära modellen för ögonfärg vid användande av en multinormal prior.



Figur 30: Markovkedjor för modellens parametrar producerade av den multinomiala modellen för ögonfärg vid användande av en prior proportionell mot den likformiga fördelningen

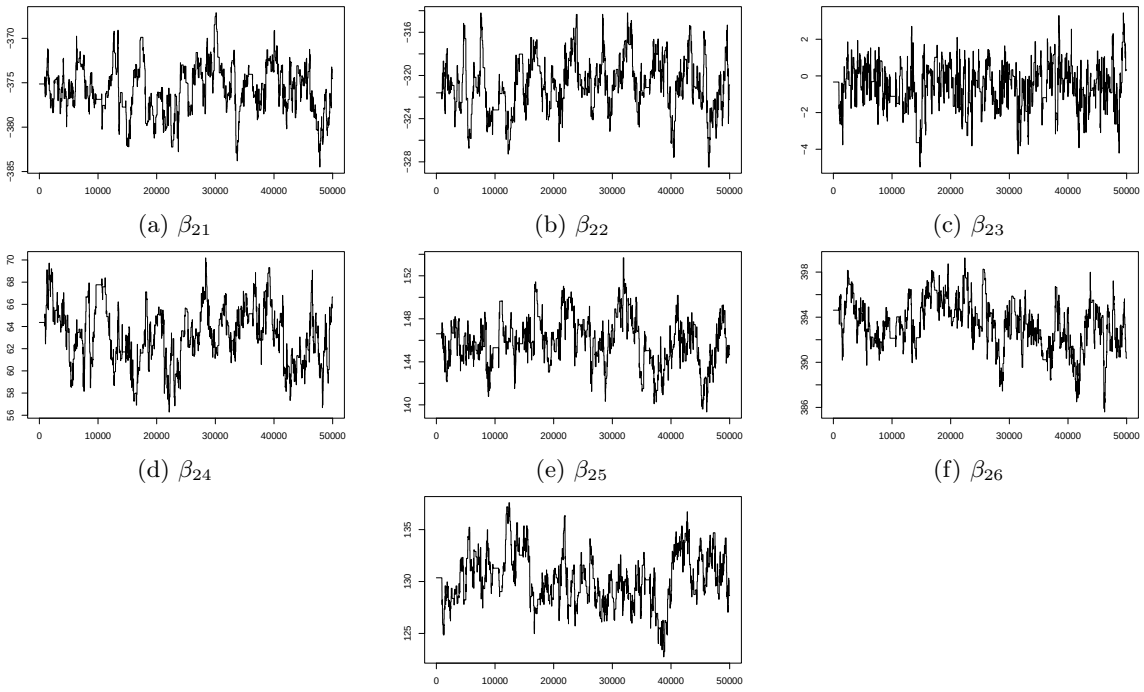


(a) β_{11} (b) β_{12} (c) β_{13}

(d) β_{14} (e) β_{15} (f) β_{16}

(g) α_1

Figur 31: Markovkedjor producerade av den multinomiala modellen vid användande av en multinormal prior, parametrarna är associerade med intermediär ögonfärg.

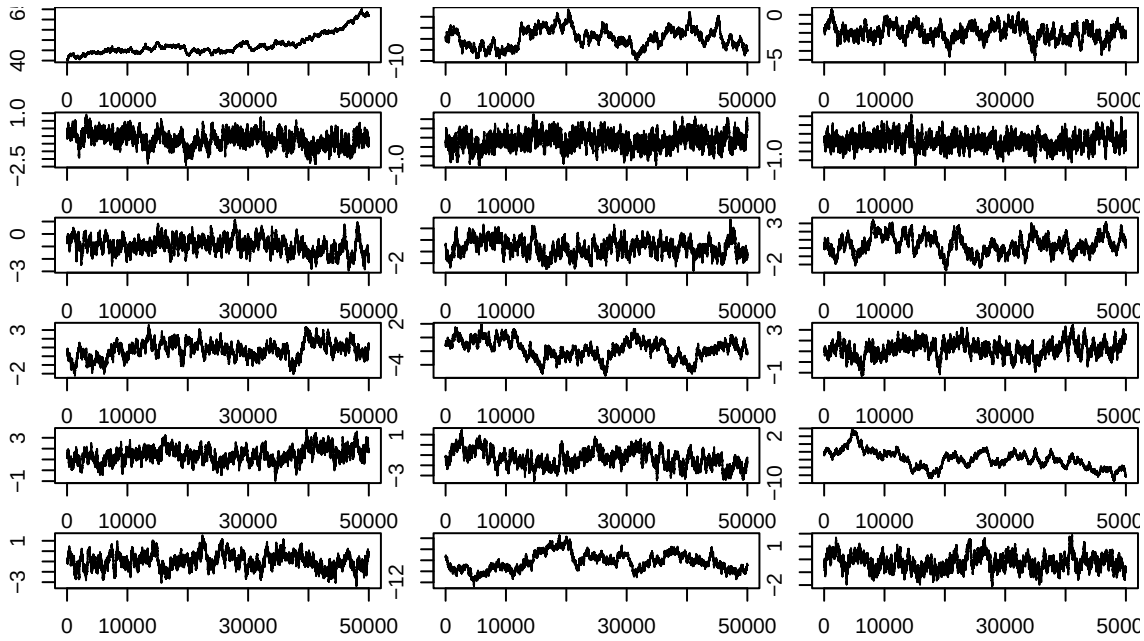


(a) β_{21} (b) β_{22} (c) β_{23}

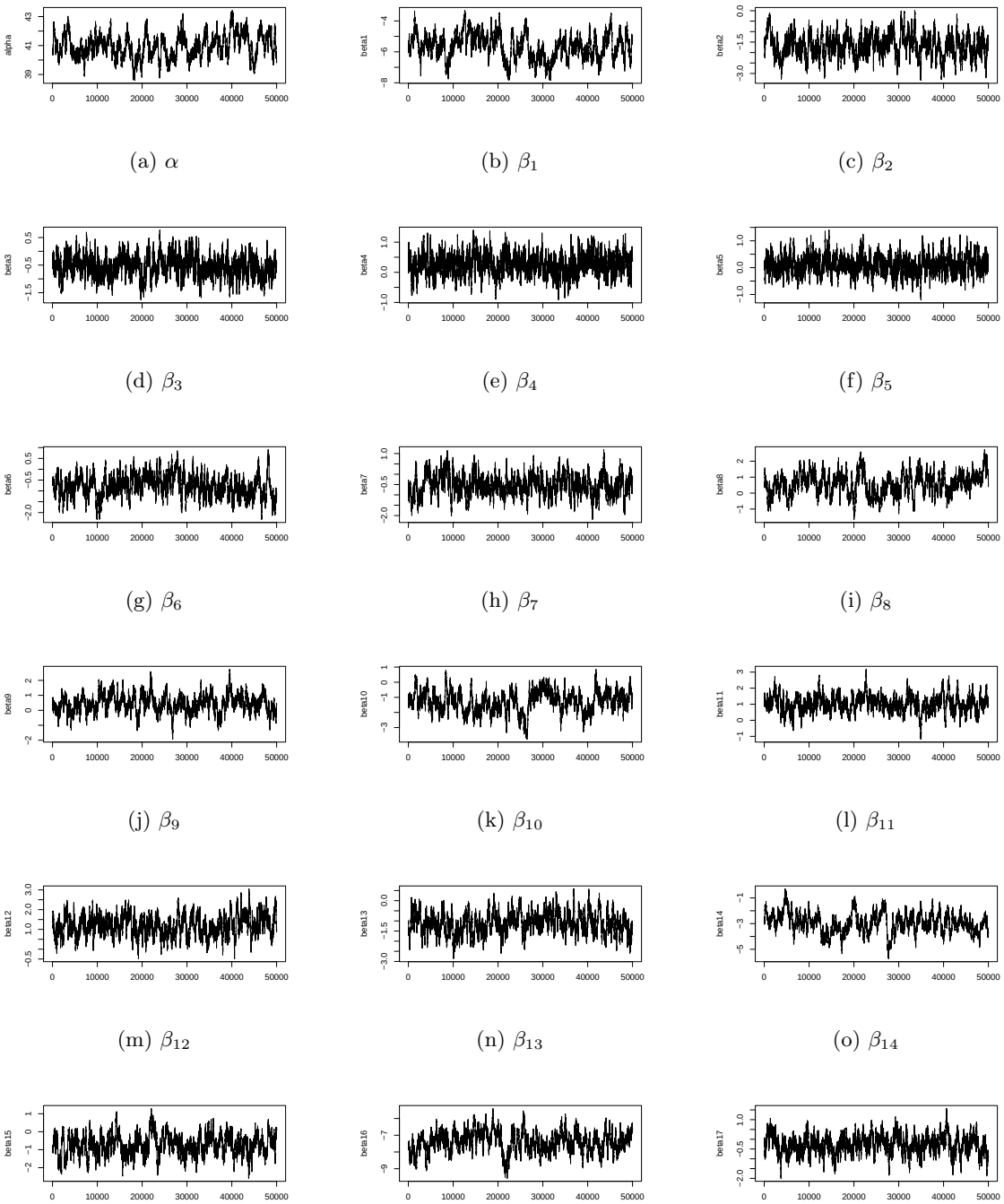
(d) β_{24} (e) β_{25} (f) β_{26}

(g) α_2

Figur 32: Markovkedjor producerade av den multinomiala modellen vid användning av en multinormal prior, parametrarna är associerade med intermediär ögonfärg.



Figur 33: Markovkedjor producerade av modellen för binär hårfärg vid användande av en icke-proper likfördelad prior.



(a) α (b) β_1 (c) β_2

(d) β_3 (e) β_4 (f) β_5

(g) β_6 (h) β_7 (i) β_8

(j) β_9 (k) β_{10} (l) β_{11}

(m) β_{12} (n) β_{13} (o) β_{14}

(p) β_{15} (q) β_{16} (r) β_{17}

Figur 34: Markovkedjor producerade av modellen som predicerar binär hårfärg vid användning av en multinormal prior.

C Källkod

Koden för de binära modellerna visas nedan. Koden för de multinomiala modellerna är nästan identisk. Det enda som skiljer dem åt är hur klassificeringarna av datamängden hanteras och att trolighetsfunktionen är en vektor. För att få de olika modellerna behöver värdena på parametrarna

justeras, vilket beskrivs i metoden.

```
# Choose the most numerous classification as the target variable,
    and convert the target variables in the prediction column to 1
    and assign the value 0 to the rest of the classifications.
df = data.frame(matrix(nrow = nrow(table), ncol = 0))
df$color = table[,2]
target_variable = unique(df$color)[which.max(tabulate(match(df$
    color, unique(df$color))))]
df$color = ifelse(df$color == target_variable, 1, 0)

# Transform the unique alleles in each SNP column to consecutive
    integers. Also remove redundant columns with exactly one unique
    value.
snp_cols = colnames(table)[-c(1, 2)]
for (col in snp_cols) {
  unique_vals = unique(table[[col]])
  if (length(unique_vals) == 1){
    next
  }
  val_map = setNames(seq_along(unique_vals), unique_vals)
  df[[col]] = val_map[as.character(table[[col]])]
}

# Log probability functions for numerical stability used in the
    likelihood.
log_prob_1 = function(h) {
  if (h >= 0) {
    return(-log1p(exp(-h)))
  } else {
    return(h - log1p(exp(h)))
  }
}
log_prob_2 = function(h) {
  if (h >= 0) {
    return(-h - log1p(exp(-h)))
  } else {
    return(-log1p(exp(h)))
  }
}

# Compute the log likelihood (used in the MCMC).
log_like = function(data, theta){
  # The data is a dataframe where the first column is the
      prediction column and the rest of the columns are observed
      data that will be used to make predictions.

  # theta is a vector of values for the parameters where the first
      element is the intercept, and the rest of the parameters are
      the coefficients that corresponds to exactly one unique column
       in the observed data.

  # Spits the data into two matrices, one for the prediction column
      and the other for the observed data. Also ensures that all
      values are numeric.
  data_matrix = as.matrix(data[,-1])
```

```
  data_labels = as.matrix(data[,1])
  class(data_matrix) = "numeric"
  class(data_labels) = "numeric"

  # Spits the theta vector into two matrices, one for the intercept
      and the other for the coefficients.
  m = matrix(theta, nrow = 1, ncol = ncol(data))
  beta_matrix = m[,-1]
  alpha_vec = m[,1]

  # Computes the cumulative value for the log likelihood for all
     rows of observed data.
  log_likelihood = 0
  n = length(data_labels)
  for(i in 1:n){
    h_vec = alpha_vec + beta_matrix %*% as.matrix(data_matrix[i,])
    log_prob1 = log_prob_1(h_vec[1])
    log_prob2 = log_prob_2(h_vec[1])
    log_likelihood = log_likelihood + data_labels[i]*log_prob1 +
       (1-data_labels[i])*log_prob2
  }
  log_likelihood
}

# Define the negative log likelihood.
negloglike = function(data, theta) - log_like(data,theta)

# MCMC with Metropolis Hastings algorithm.
mcmc = function(data, n, start){
  # The data is a dataframe where the first column is the
     prediction column and the rest of the columns are observed
     data that will be used to make predictions.

  # n is the amount iterations in the MCMC algorithm.

  # start is the initial of the vector of parameters for the MCMC
     algorithm.
  theta_vec = as.vector(start)

  # Simulate the chain of values for the parameters, and use
     Metropolis Hastings algorithm to determine whether to update
     the chain with the simulated values or to keep the same values
      for the next simulation.
  out = matrix(rep(theta_vec, n), nrow=n, byrow = TRUE)
  for(i in 2:n){
    # Simulate new values for the parameter that are normal
       distributed with mean being the value in the previous
       iteration and standard deviation being choosen.
    newtheta = theta_vec + rnorm(length(theta_vec), mean = 0, sd =
       0.7)

    # Compute the log likelihood for the current and previous
       iteration.
    like_new = log_like(data, newtheta)
    like_old = log_like(data, theta_vec)
    loga = like_new - like_old
```

```
85
86      # Compute the log prior if it is normal.
87      covariance_matrix = diag(1, nrow=length(theta_vec))
88      covariance_matrix[1,1] = 10
89      covariance_matrix[7,7] = 10
90
91      log_prior_new = log(dmnorm(newtheta, mean = as.vector(start),
            covariance_matrix))
92      log_prior_old = log(dmnorm(theta_vec, mean = as.vector(start),
            covariance_matrix))
93
94      # Add the prior to the log likelihood.
95      loga = loga + log_prior_new - log_prior_old
96
97      # Compute acceptance by using Metropolis Hastings.
98      if(log(runif(1)) < loga){
99        theta_vec = newtheta
100     }
101     # Store the current values for the parameters.
102     out[i,] = theta_vec
103   }
104   out
105 }
106
107 # Probability calculation of the prediction.
108 calc_probs = function(ynew_data, theta, iters){
109   # ynew_data is the test data that will be predicted.
110
111   # theta is the Markov chain for the parameters.
112
113   # iters is the length of the chain.
114
115   # Create a loop for each step in the chain where each iteration
        compute the probabilities for each classification. This
        essentially computes the MCMC sum that approximates the
        probabilities prediction.
116   result_vec = as.vector(c(0,0))
117   y_data = ynew_data[-1]
118   class(y_data) = "numeric"
119   for(i in ((iters+2)/2):iters){
120     theta_m = matrix(theta[i,], nrow = 1, ncol = ncol(ynew_data))
121     beta_matrix = theta_m[,-1]
122     alpha_vec = theta_m[,1]
123
124     h_vec = alpha_vec + beta_matrix %*% as.matrix(y_data)
125     denominator = 1 + sum(sapply(h_vec, exp))
126     numerator = sapply(h_vec,exp)
127
128     result_vec[1] = result_vec[1] + numerator/denominator
129     result_vec[2] = result_vec[2] + (1 - numerator/denominator)
130   }
131
132   # Normalize the probabilities.
133   result_vec = result_vec/(iters/2)
134   result_vec
135 }
```

```
136
137 set.seed(1)
138
139 # Number of iterations in the MCMC.
140 iters = 10000
141
142 # Prepare the confusion matrix.
143 conf_matr = data.frame(
144   V1 = c(0, 0),
145   V2 = c(0, 0)
146 )
147 colnames(conf_matr) = c(
148   paste('Predicted', target_variable),
149   paste('Predicted not', target_variable)
150 )
151 rownames(conf_matr) = c(
152   paste('Actual', target_variable),
153   paste('Actual not', target_variable)
154 )
155
156 # Initialize an empty data frame to store the probabilities, the
    prediction and the correct answer.
157 probability_infos = data.frame(matrix(ncol = 4, nrow = 0))
158 colnames(probability_infos) = c(paste(target_variable, 'probability
    '),
159           paste('not', target_variable, 'probability'), "correct
              answer", "Predicted")
160
161 # Loop over the dataset to apply LOOCV to measure the models
    predictive ability.
162 for(n in 1:dim(df)[1]){
163   cat("Testing on ", n,"\n")
164   # Choose the new test and train data
165   test_table = df[n,]
166   train_table = df[-n,]
167
168   # Compute the maximum likelihood by choosing the initial vector
      as ones.
169   theta_start = matrix(1, 1, ncol(df))
170   maxlikelihood = nlm(negloglike, p=as.vector(theta_start), data =
      train_table)$estimate
171   ml_param = matrix(maxlikelihood, nrow = 1, ncol = ncol(df))
172
173   # Apply MCMC to obtain a chain for the parameters.
174   mcmc10000 = mcmc(train_table, iters, ml_param)
175
176   # Find the prediction by finding the maximum of the predictive
      probabilities.
177   label = test_table[1]
178   probabilities = calc_probs(test_table, mcmc10000, iters)
179   prediction = which.max(probabilities)
180   max_probability = probabilities[prediction]
181
182   # Tabulate the result.
183   if(label == 1){
184     if(prediction == 1){
```

```
185       conf_matr[1,1] = conf_matr[1,1] + 1
186     } else{
187       conf_matr[1,2] = conf_matr[1,2] + 1
188     }
189   }else{
190     if(prediction == 1){
191       conf_matr[2,1] = conf_matr[2,1] + 1
192     } else{
193       conf_matr[2,2] = conf_matr[2,2] + 1
194     }
195   }
196   # Store the data for each iteration.
197   probability_infos[nrow(probability_infos) + 1,] = c(probabilities
        , test_table[1], prediction*-1 + 2)
198 }
199
200 # Compute the ROC-curves and AUC-values.
201 rocobj = roc(probability_infos[, 3], probability_infos[, 1], levels
      = c(0,1))
202 auc = round(auc(probability_infos[, 3], probability_infos[, 1],
     levels = c(0,1)),4)
```