KLASSIFIKATION AF GENOTYPER MED POSTERIOR SANDSYNLIGHEDER

I RETSGENETISKE PROBLEMSTILLINGER

Micklas Visby Christiansen, Sarah Elise Steen
& Sofie Lovise Uhrbrand

Matematik, Gruppe 4.206a, 22. December 2022

Speciale

AALBORG UNIVERSITET
STUDENTERRAPPORT

Det Teknisk-Naturvidenskabelige Fakultet
Institut for Matematiske Fag
Skjernvej 4a
9220 Aalborg Ø
https://www.math.aau.dk/

Titel

Klassifikation af genotyper med posterior sandsynligheder i retsgenetiske problemstillinger

Emne

Statistik, Speciale,
Matematik, Vintersemester 2022

Projekt periode

05/09/2022 - 22/12/2022

Gruppe

4.206a

Gruppemedlemmer

Micklas Visby Christiansen
Sarah Elise Steen
Sofie Lovise Uhrbrand

Vejleder

Mikkel Meyer Andersen

Opslagstal 5
Sideantal 61
Afsluttet d. 22/12/2022

Synopsis

I denne rapport bearbejdes et udleveret datasæt fra Retsgenetisk Afdeling, Retsmedicinsk Institut på Københavns Universitet, som består af 1100 SNPer fra fire anonymiserede personer. I rapporten konstrueres en klassifikationsmodel fra et supervised learning-perspektiv med henblik på at kunne forudsige genotyperne i tre forskellige varianter ud fra observationers signalintensiteter. Hertil anvendes multinomial logistisk regression, hvor K-fold krydsvalidering anvendes til validering af den multinomiale logistiske regressionsmodel. I rapporten vil det yderligere blive undersøgt, om der genotyperne er symmetrisk omkring identitetslinjen med henblik på at konstruere en simplere klassifikationsmodel. Den underliggende statistiske analyse udarbejdes i R-Studio og R (version 4.2.1), hvor tilhørende scripts vil kunne findes i Github, [1].

Forord

Dette speciale er skrevet i efteråret 2022 af Micklas Visby Christiansen på 10. semester af hans kandidat i Matematik og Fysik, Sofie Lovise Uhrbrand på 10. semester af hendes kandidat i Matematik og Kemi, og Sarah Elise Steen på 11. semester af hendes kandidat i Matematik og Dansk, på Det Teknisk-Naturvidenskabelige Fakultet på Aalborg Universitet under kyndig vejledning af Mikkel Meyer Andersen. Det overordnede tema for projektet er *Statistik*.

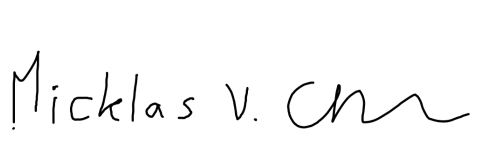
Ved kildehenvisning anvendes Vancouver Style. Ved citering anvendes $[n]$, hvor n angiver placeringen af kilden i litteraturlisten. Bliver der derudover, henvist til et specifik kapitel eller en specifik side, anvendes citering $[n, \text{kap. } m]$ eller $[n, \text{s. } m]$, hvor m angiver kapitelnummer eller sidetal. Anvendte kilder introduceres i starten af hvert kapitel.

Sætninger, definitioner, propositioner, korollarer, eksempler, figurer og tabeller nummereres efter kapitel. Nogle udregninger er angivet ved kapitelnummer. Henvisning til tidligere udregninger foretages med $(n.m)$, hvor n angiver kapitel og m angiver udregningens nummer.

I projektet behandles data udleveret af Retsgenetisk Afdeling, Retsmedicinsk Institut på Københavns Universitet ved hjælp af software environmentet R-Studio og R (version 4.2.1) til statistiske beregninger og modellering af data.

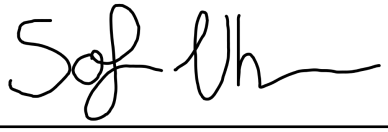
Aalborg Universitet, 22. december 2022

Underskrifter



Micklas Visby Christiansen
»mvch17@student.aau.dk«

Sarah Elise Steen
»sest17@student.aau.dk«



Sofie Lovise Uhrbrand
»suhrbr16@student.aau.dk«

Abstract

The purpose of this report was to construct a model, that could predict genetic variants, called genotypes, of specific DNA observations with the lowest possible amount of misclassifications. A dataset containing 4400 observations of DNA sequences, equivalent to 1100 Single Nucleotide Polymorphisms (SNPs) from four anonymous persons, was used to construct the model. The data was provided by the Department of Forensic Genetics at Copenhagen University and contained two signals, A and B, and the variable WGS, which contained the *true* genotype of the observation. The signals was obtained from the method *Next Generation Sequencing* and the variable WGS was obtained from the method *Whole Genome Sequencing*. The classification model was constructed from a supervised learning point of view such that the signals was used to train the model to classify as WGS. The trained model classified future observations into genotypes based on the signal intensities from A and B. In this report, the trained model was based on multinomial logistic regression (MLR) and the percentage of correctly classified observations is defined as the accuracy of the model. To increase the precision of the model, a group called no-call was introduced for data points with maximum posterior probability beneath a given threshold.

When visualizing the clusters corresponding to the three genotypes, it was observed that a symmetrical relationship could exist between the clusters around the identity line. If the genotypes were found to be symmetrical around the identity line, a simplified classification model could be constructed. The simplified model would then be tested to investigate if a similar classification to the MLR model could be obtained. It was tested whether the covariance matrices and means of the homozygous genotypes were symmetrical using Box's M-test and Hotelling's T^2-test. These tests showed that the covariance matrices were not symmetrical, but the means were symmetrical. When testing whether the heterozygous genotype was symmetrical around the identity line a correlation test and a t-test were conducted. The tests showed that the heterozygous genotype was not symmetrical around the identity line. As the genotypes were shown not to be symmetrical around the identity line, a guide to construct a simplified classification model was made, rather than constructing an actual model.

It can be concluded that a model constructed with multinomial logistic regression classified observations with an accuracy of 96.3% from the signals A and B compared to WGS, and after introduction of no-call, an accuracy of 97.7% was achieved. The MLR model obtained an acceptable accuracy, however the MLR model still misclassified some observations after the introduction of no-call.

As further development, it could be investigated whether the genotypes would become symmetrical around the identity line by manipulating the sizes of the groups.

Symbolregister

C	Antal grupper	S_i, Σ, Q_i	Kovariansmatrix
c	c'te gruppe	S	Pooled kovariansmatrix
$CV_{(K)}$	Fejlfrekvens	T_{krit}	Kritisk værdi
df	Frihedsgrader	$\mathcal{T}^2(df, df_2)$	Hotellings T^2-fordeling
$\mathbb{E}$	Forventet værdi	T^2	Hotellings teststørrelse
Err_i	Testfejl	t	teststørrelse t-test
$\mathcal{F}$	Fishers transformation	$W_p(\Sigma, df)$	Wishart-fordeling
$F(df, df_2)$	F-fordeling	Y	Responsvariabel
$\mathcal{H}_0$	Nulhypotese	$\mathcal{Z}$	Fishers teststørrelse
$\mathcal{H}_1$	Alternativ hypotese	X_A	Signal $\log(A+1)$
$h(t)$	Model	X_B	Signal $\log(B+1)$
K	Antal folder	β	Koefficientvektor
k	k'te fold	λ	Likelihood ratio teststørrelse
$L(\cdot)$	Logit	μ	Middelværdi
$\mathcal{L}$	Likelihood funktion	$\hat{\rho}$	Pearsons korrelationskoefficient
M	Box's M-teststørrelse	σ	Standard afvigelse
$N_p(\mu, \Sigma)$	Multivariat normalfordeling	σ^2	Varians
$O(\cdot)$	Odds	$\chi^2(df)$	Chi-i-anden fordeling
$P(\cdot)$	Sandsynlighed	ω	Fastlagt tærskel
$\lvert Q \rvert$	Determinant af Q	$\hat{}$	Estimat

Forkortelsesregister

GLM	Generaliseret lineære modeller
MLR	Multinomial logistisk regression
MLE	Maksimum likelihood estimat
NC	No-call
NGS	Next Generation Sequencing
SNP	Single nukleotid polymorfi
UTAG	Uden tab af generalitet
WGS	Helgenomsekventering (Whole Genome Sequencing)

Indhold

Forord **i**

Abstract **iii**

Symbolregister **v**

Forkortelsesregister **vi**

1 Introduktion **1**

2 Logistisk regression **4**
2.1 Modeller 4
2.2 Estimation af logistiske regressionskoefficienter 7
2.3 Udvidelse: Multipel logistisk regression 7
2.4 Multinomial logistisk regression (MLR) 9

3 K-fold krydsvalidering **12**

4 Præsentation af data **13**

5 Klassifikation af genotyper **16**
5.1 Klassifikation med MLR 16
5.2 Bestemmelse af beslutningsgrænser 18
5.2.1 Beslutninggrænse for no-call bånd 19
5.2.2 Konstruktion af no-call bånd 21
5.3 Test af model 22
5.3.1 Test på 400 observationer 23
5.3.2 Krydsvalidering af klassifikation med MLR 24

6 Hypotesetest af symmetri **27**
6.1 Korrelationstest 28
6.2 t-test 30
6.3 Box's M-test 31
6.4 Hotellings T^2-test 33

7 Test af symmetri **35**
7.1 Symmetri i homozygote genotyper omkring identitetslinjen 35
7.2 Symmetri i heterozygot genotype omkring identitetslinjen 37
7.3 Konstruktion af symmetrimodel 40

8 Diskussion **42**

9 Konklusion **44**

10 Perspektivering **45**

11 Litteratur **48**

Appendiks **50**

A Appendiks A **51**
A.1 Generaliserede lineære modeller (GLM) 51
A.2 Maksimum likelihood estimat (MLE) 52
A.3 Fordelinger 53
A.4 Multivariat normalfordeling 54
A.4.1 MLE for multivariat normalfordeling 56
A.4.2 Likelihood ratio test for multivariat normalfordeling 56

B Appendiks B **59**
B.1 Kritiske værdier 59

C Appendiks C **61**
C.1 Beregninger til Box's M-test i Maple (version 2021.2) 61

1 Introduktion

Indenfor retsgenetik spiller DNA en væsentlig rolle i identifikationen af personer i eksempelvis kriminalitet, faderskab, familiesammenføring, identifikation af ofre m.m. DNA eksisterer i alle levende organismer og er opbygget af baserne A (*adenin*), C (*cytosin*), G (*guanin*) og T (*thymin*). DNA består af to strenge, som snor sig som en dobbelt helix, der holdes sammen af svage hydrogenbindinger mellem de fire baser, hvorfor baserne A og T er forbundet, samt C og G er forbundet. Kombinationen af disse baser udgør en unik sekvens, som gør enhver levende organismes DNA unikt. [2, kap. 1]

I denne rapport analyseres et datasæt udleveret af Retsgenetisk Afdeling, Retsmedicinsk Institut på Københavns Universitet. Heri fokuseres der på den del af DNAet, hvor der er genetisk variation mellem personer. Variationer i DNA finder sted på specifikke placeringer kaldet *single nucleotid polymorfi* (SNP) med længde af én base. Basesekventeringen af DNAet kan foretages med helgenomssekventering (*Whole Genome Sequencing*, WGS), hvor sammensætningen af baser i DNA kortlægges for hele genomet, altså hele DNAet fra en person, eller med en billigere metode *Next Generation Sequencing* (NGS). I NGS bliver hele genomet ikke sekventeret, men målrettes SNPer, som er jævnt fordelt i hele genomet. I NGS anvendes en såkaldt snipchip, hvor der bruges to laserlys til at bestemme de baser, som er tilstede i en SNP. I denne rapport betegnes laserlysene som A- og B-signaler. De tilstedeværende baser i SNPen afgør, hvilken genetisk variant (genotype) SNPen klassificeres som. Da DNA nedarves i to kopier, består SNPen af to versioner, der enten består af samme basepar eller forskellige basepar. Hvis laserlysene emitterer to ens baser, vil SNPen klassificeres som en homozygot genotype, hvorimod hvis laserlysene emitterer to forskellige baser, vil SNPen klassificeres som en heterozygot genotype. [3] [2, kap. 4]

Det udleverede datasæt indeholder en kortlægning af 1100 SNPer fra fire anonymiserede personer med snipchip og WGS. Datasættet består af signaler A og B fra snipchippen samt en variabel med klassifikationen af genotypen i den enkelte SNP bestemt med WGS. Denne variabel betragtes som en responsvariabel, som indeholder den *sande* genotype. I denne rapport ønskes det at konstruere en klassifikationsmodel, som kan oversætte signalerne A og B til tre klynger, som korresponderer til tre forskellige genotyper. Modellen kan baseres på enten *supervised-* eller *unsupervised-learning*. Hvis de uafhængige variabler (prædiktorer) er kendte, men den afhængige variabel (responsvariabel) er ukendt, vil der være tale om unsupervised learning. Hvis prædiktorerne og responsvariablen er kendte, vil der være tale om supervised learning. I supervised learning ønskes det at finde en sammenhæng mellem prædiktorerne og responsvariablen med henblik på at kunne lave en nøjagtig forudsigelse af responsvariablen på fremtidige observationer. [4, kap. 2]

I denne rapport tages der udgangspunkt i supervised learning, hvor WGS betragtes som *sandheden*, og den sande genotype for hver SNP er derfor kendt. Da responsvariablen WGS er kvalitativ og indeholder mere end to grupper, anvendes multinomial logistisk regres-

sion (MLR) som klassifikationsmodel, hvilket er en udvidelse af logistisk regression. MLR-modellen er baseret på sandsynligheder for observationernes inddeling, kaldt posterior sandsynligheder, og en observation tildeles den klynge, hvor posterior sandsynligheden er størst. I supervised learning trænes MLR-modellen med WGS, hvorfor modellen evalueres ved at betragte konfusionsmatricer.

Det er vigtigt at minimere antallet af fejlklassifikationer, da de kan have fatale konsekvenser i retgenetiske undersøgelser. Derfor fastsættes en beslutsningsgrænse for, hvor sikker MLR-modellen skal være, før en genotype bliver kaldt for en observation. Dette leder os frem til følgende problemformulering:

Kan en model trænes ved hjælp af signalerne fra sekventeringsmetoden NGS, så den trænede model opnår klassifikationer tilsvarende sekventeringsmetoden WGS?

Kapiteloversigt

I **Kapitel 2** introduceres en logistisk regressionsmodel, hvor en responsvariabel indeholder to grupper. Denne udvides til multipel logistisk regression, hvor modellen indeholder flere prædiktorer. Herefter udvides multipel logistisk regression til MLR, hvor klassifikationen består af mere end to grupper ved anvendelse af flere prædiktorer. MLR danner udgangspunktet for modellerne anvendt i denne rapport. Det introduceres løbende, hvordan koefficienterne i modellerne estimeres vha. *maximum likelihood estimate* (MLE).

I **Kapitel 3** introduceres *resampling*-metoden K-fold krydsvalidering, som i denne rapport anvendes til at evaluere klassifikationsmodellen, som konstrueres i Kapitel 5. Denne evalueringsmetode anvendes, da denne metode kan give et indblik i testfejl og fejlfrekvens.

I **Kapitel 4** gives en nærmere introduktion af det udleverede datasæt og metoden NGS. Ydermere gives en forklaring på, hvordan signalerne A og B bliver anonymiseret, og afslutningsvis visualiseres klyngedannelserne opnået med helgenomssekventering.

I **Kapitel 5** konstrueres klassifikationsmodellen MLR. Da forskellige valg af referenceniveau påvirker koefficienterne anvendt i MLR-modellen, men valget af referenceniveau ikke har indflydelse på posterior sandsynlighederne, tages der i denne rapport kun udgangspunkt i genotype AB som referenceniveau. MLR-modellen konstrueres både for signalerne A og B, og log-transformerede signaler A og B. Det vurderes i hvilket tilfælde, der opnås den største modelpræcision målt på *accuracy*. Hvordan MLR-modellen klassificerer observationer visualiseres i konfusionsmatricer. I den forbindelse introduceres en beslutningsgrænse, som fastsætter, hvor sikker MLR-modellen skal være, før en genotype bliver kaldt. Dette evalueres med test af MLR-modellen og 11-fold krydsvalidering.

I **Kapitel 6** opstilles hypoteser til at undersøge, om genotyperne er symmetriske omkring identitetslinjen, hvorefter testene korrelationstest, t-test, Box's M-test og Hotellings T^2-test bliver introduceret. Introduktionen af testene tager udgangspunkt i deres anvendelse i denne rapport, hvorfor der ikke bevises resultater om teststørrelsernes egenskaber.

I **Kapitel 7** anvendes testene fra Kapitel 6 til at teste hypoteserne opstillet i Kapitel 6. Afslutningsvis gives en metode til konstruktion af en symmetrimodel.

2 Logistisk regression

Dette kapitel er baseret på [4, kap. 4] og [5, kap. 4].

En tilgang til at prædiktere en responsvariabel y er kendt som en proces kaldt klassifikation. At prædiktere y for en observation x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, kan betegnes som at klassificere en observation til en bestemt gruppe, hvor n er antallet af observationer, og p er antallet af prædiktorer. I et klassifikationsproblem trænes en model på et træningsdata $X = [x_{ij}]$ med henblik på, at modellen kan prædiktere en responsvariabel y. Denne model skal fungere godt på træningsdata såvel som testdata, som ikke er anvendt til at træne modellen.

En logistisk regressionsmodel betragtes også som en generaliseret lineær model (GLM). Såfremt læseren ikke er bekendt med GLM henvises læseren til at læse Appendiks A.1, som er et supplement til dette kapitel.

2.1 Modeller

Logistisk regression er en model for en responsvariabel y for to grupper, $C = 2$, og et antal kovariater. En logistisk regressionsmodel er intuitivt opbygget omkring en lineær regression,

$$g(x) = \beta_0 + \beta_1 x, \tag{2.1}$$

der fokuserer på at modellere en kvantitativ responsvariabel y direkte, men fokuset for en logistisk regressionsmodel er i stedet at modellere sandsynligheden for, at y tilhører en bestemt gruppe, dvs. $h(x_{ij}) = \mathrm{P}(y = 1|X = x_{ij})$, $h(x_{ij}) \in [0, 1]$. For responsvariablen y anvendes den generiske 0/1-kodning, hvor 1 indikerer, at en observation tilhører den positive gruppe, og 0 indikerer, at en observation tilhører den komplementære gruppe.

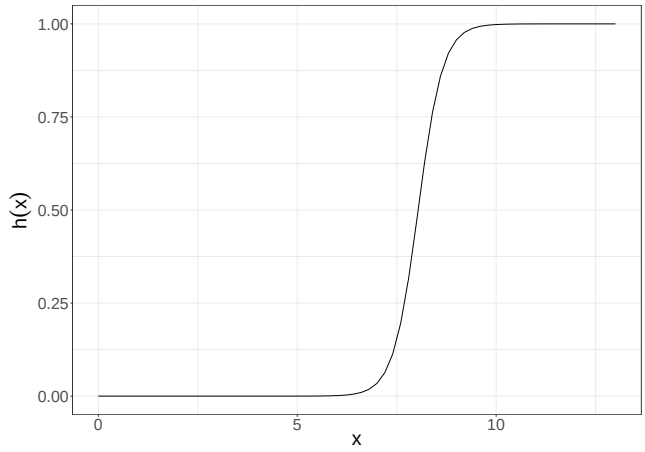
Præsentationen af en lineær regressionsmodel giver en problematik, når en lineær linje passer til et binært svar. Hvis x ikke er begrænset, vil der prædikteres en sandsynlighed for $g(x) < 0$ for nogle værdier af x og $g(x) > 1$ for andre værdier af x. Prædikationer som disse er ikke fornuftige, da en sandsynlighed skal ligge i intervallet $[0, 1]$. For at undgå denne problematik introduceres en funktion $h : \mathbb{R} \to [0, 1]$. Denne kan eksempelvis være en logistisk regressions kurve

$$h(t) = P(t = 1) = \frac{\exp(t)}{1 + \exp(t)}, \tag{2.2}$$

hvor t kan være enhver linear kombination af prædiktorer. Et eksempel på en linear kombination kan være den lineære regressionsmodel $g(x)$ introduceret i (2.1). Antag nu, at $t = \beta_0 + \beta_1 x$, hvorfra

$$h(t) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Funktionen h kaldes en logistisk funktion, hvis resultat kan fortolkes som posterior sandsynligheden for, at en observation tilhører en bestemt gruppe, altså $P(y = 1|X = x_{ij})$. Grafen for en sådan funktion h vil ligne en S-formet kurve uanset værdien af x. Dette er illustreret i Figur 2.1.



Figur 2.1: Illustration af S-kurven. I kurven betragtes tilfældende $P(y = 1|X = x_{ij})$ og $P(y = 0|X = x_{ij})$. Hvis den undersøgte observation har en posterior sandsynlighed større end 0.5, tilhører observationen den positive gruppe. Hvis observationen har en posterior sandsynlighed mindre end 0.5, tilhører observationen den komplementære gruppe.

Dermed kan den logistiske regressionsmodel defineres som følger.

Definition 2.1. Logistisk regressionsmodel
Lad $h(x) \in [0, 1]$. En logistisk regressionsmodel er defineret ved

$$h(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{2.3}$$

Udtrykket for modellen $h(x)$ kan udtrykkes simplere, hvor tilfældet $h(x) = P(y = 1|X = x_{ij})$ anvendes og divideres med sandsynligheden for det komplementære tilfælde, altså sandsynligheden for $1 - h(x) = P(y = 0|X = x_{ij})$. Dvs.,

$$O(t) = \frac{h(t)}{1 - h(t)} = \frac{\dfrac{\exp(t)}{1 + \exp(t)}}{1 - \dfrac{\exp(t)}{1 + \exp(t)}} = \exp(t)$$

for $t = \beta_0 + \beta_1 x$. Dette udtryk kaldes for *odds*. Denne funktion opfylder ikke det tidligere krav om en funktion, som afbilder over i $[0, 1]$, men den afbilder fra $\mathbb{R}$ til $(0, \infty)$.

Odds er dog ækvivalent med modellen i (2.3) i den forstand, at odds tæt på 0 eller ∞ indikerer henholdsvis lave eller høje posterior sandsynligheder.

Proposition 2.2. Odds, logistisk regression
Modellen (2.3) kan omskrives til

$$O(x) = \frac{h(x)}{1-h(x)} = \exp(\beta_0 + \beta_1 x), \tag{2.4}$$

hvor $O(x)$ kaldes odds, og kan antage enhver værdi i intervallet $(0, \infty)$.

Bevis:
Antag, at $O(t) \in (0, \infty)$ og $t = \beta_0 + \beta_1 x$. Anvend modellen fra Definition 2.1, hvorfra det følger, at

$$\begin{aligned} O(t) = \frac{h(t)}{1-h(t)} &= \frac{\exp(t)}{1+\exp(t)} \; \frac{1}{1 - \frac{\exp(t)}{1+\exp(t)}} \\ &= \frac{\exp(t)}{1+\exp(t)} \; \frac{1}{\frac{1+\exp(t)-\exp(t)}{1+\exp(t)}} \\ &= \frac{\exp(t)}{1+\exp(t)} \; (1+\exp(t)) = \exp(t). \end{aligned}$$

Dermed er (2.4) opfyldt. ■

Da odds kun angiver inputvektoren af en positiv gruppe defineres nu en invers funktion $L : (0, \infty) \to \mathbb{R}$, som opfylder, at

$$L\left(O(t)\right) = t, \tag{2.5}$$

hvor $t = \beta_0 + \beta_1 x$. Udtrykket i (2.5) opfyldes ved at anvende logaritmen på den logistiske regression. Dette udtryk kaldes *logit* eller *log odds* og fungerer intuitiv som en manipulation af Proposition 2.2.

Proposition 2.3. Logit, logistisk regression
Logit for en logistisk regressionsmodel er givet ved

$$L(x) = \log\left(\frac{h(x)}{1-h(x)}\right) = \beta_0 + \beta_1 x, \tag{2.6}$$

hvor $L : (0, \infty) \to \mathbb{R}$.

Ved at lave en algebraisk manipulation af odds vil det gøre det nemmere at fortolke koefficientvektoren β, imens den lineære form bevares. For eksempelvis Proposition 2.3, kan det nærmere undersøges, hvad en enhedsforøgelse i x får logit til at stige med. Dermed kan

output fra logit også indikerer, hvor posterior sandsynligheden for en given gruppe er. En negativ logitværdi vil indikere en posterior sandsynlighed mindre end 0.5, hvor en positiv logitværdi vil indikere en posterior sandsynlighed større end 0.5. For logit er dette forhold symmetrisk.

2.2 Estimation af logistiske regressionskoefficienter

Parametrene β_0 og β_1 er ukendte, og må estimeres ud fra træningsdataet vha. MLE. Intuitionen ved at anvende MLE til at tilpasse en logistisk regressionsmodel er, at der findes estimater for β_0 og β_1, noteret $\hat{\beta}_0$ og $\hat{\beta}_1$. Ved at indsætte disse estimater i modellen (2.3) fås en posterior sandsynlighed så tæt på 1 som muligt for alle observationer i en bestemt gruppe, og en posterior sandsynlighed så tæt på 0 som muligt for de observationer, der ikke tilhører den bestemte gruppe. Ved at anvende en likelihood funktion defineret i Appendix A.2, kan dette formaliseres ved,

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1) &= \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \\ &= \prod_{i:y_i=1} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \prod_{i':y_{i'}=0} \left(1 - \frac{\exp(\beta_0 + \beta_1 x_i')}{1 + \exp(\beta_0 + \beta_1 x_i')}\right),\end{aligned} \tag{2.7}$$

hvor $\hat{\beta}_0$ og $\hat{\beta}_1$ vælges til at maksimere (2.7). Når koefficienter til den logistiske model er estimeret, kan de statistiske prædikationer $\hat{p}(x_i)$ for en given klassifikation udregnes. Disse skal opfylde, at $0 \leq P(x_i) \leq 1$, og $\sum_{j=1}^{C} P(x_i)_j = 1$.

2.3 Udvidelse: Multipel logistisk regression

Hvis et datasæt indeholder mere end en kvalitativ prædiktor, kan logistisk regression udvides til multipel logitisk regression, dvs. at klassifikationsproblemet udvides til at forudsige en binær responsvariabel ved at anvende flere prædiktorer. Udvidelsen laves ved at give hver prædiktor en separat parameter β_j, $j = 1, \ldots, p$ i en enkelt model med p forskellige prædiktorer. Dermed udvides X fra en vektor til en $n \times p$ matrix med $(x_1, \ldots, x_n)^T$ rækker og $(x_1, \ldots, x_p)$ søjler. Altså,

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}.$$

For den logistiske kurve præsenteret i (2.2) antages

$$t = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \sum_{j=0}^{p} \beta_j x_j,$$

hvor $x_0 = 1$. Dermed kan multipel logistisk regression defineres som følger.

Definition 2.4. Multipel logistisk regression
Lad $h(X) \in [0,1]$. En multipel logistisk regressionsmodel er defineret ved

$$h(X) = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_j\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_j\right)}, \quad x_0 = 1. \tag{2.8}$$

Ligesom for logistisk regression kan udtrykket for $h(x)$ i (2.8) udtrykkes simplere, hvor tilfældet $h(X) = P(y = 1|X = x_{ij})$ anvendes og divideres med sandsynligheden for det komplementære tilfælde, altså sandsynligheden for $1 - h(x) = P(y = 0|X = x_{ij})$. Dvs.,

$$O(t) = \frac{h(t)}{1 - h(t)} = \frac{\frac{\exp(t)}{1 + \exp(t)}}{1 - \frac{\exp(t)}{1 + \exp(t)}} = \exp(t).$$

Dette udtryk kaldes for *odds*. Denne funktion afbilder fra $\mathbb{R}$ til $(0, \infty)$, og er ækvivalent med modellen i (2.8) i den forstand, at odds tæt på 0 eller ∞ indikerer henholdsvis lave eller høje sandsynligheder.

Proposition 2.5. Odds, multipel logistisk regression
Modellen (2.8) kan omskrives til

$$O(X) = \frac{h(X)}{1 - h(X)} = \exp\left(\sum_{j=0}^{p} \beta_j x_j\right), \tag{2.9}$$

hvor $x_0 = 1$ og $O(X)$ kaldes odds, og kan antage enhver værdi i intervallet $(0, \infty)$.

Bevis:
Resultatet følger af Definition 2.4 og Proposition 2.2. ■

Ligesom for logistisk regression, vil en algebraisk manipulation af odds gøre det nemmere at fortolke koefficientvektoren β for multipel logistisk regression.

Proposition 2.6. Logit, multipel logistisk regression
Logit for en multipel logistisk regressionsmodel er givet ved

$$L(X) = \log\left(\frac{h(X)}{1 - h(X)}\right) = \sum_{j=0}^{p} \beta_j x_j, \tag{2.10}$$

hvor $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ er en koefficientvektor og $x_0 = 1$.

Til estimation af koefficientvektoren $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ anvendes en udvidelse af likelihood funktionen i (2.7):

$$\mathcal{L}(\beta) = \prod_{i:y_i=1} \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ji}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ji}\right)} \prod_{i':y_i'=0} \left(1 - \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ji'}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ji'}\right)}\right), \tag{2.11}$$

hvor $x_0 = 1$ og $\hat{\beta}$ vælges til at maksimere (2.11). Når koefficienter til den multiple logistiske model er estimeret, kan posterior sandsynligheden for en given klassifikation udregnes.

2.4 Multinomial logistisk regression (MLR)

I tilfælde, hvor en responsvariabel indeholder mere end to grupper, $C > 2$, kan multinomial logistisk regression anvendes. Her vælges en af grupperne som referenceniveau. UTAG vælges den C'te gruppe som referenceniveau. Antag, at for hver gruppe c, så er

$$t_c = \beta_{0,c} + \beta_{1,c} x_1 + \cdots + \beta_{p,c} x_p = \sum_{j=0}^{p} \beta_{j,c} x_j, \tag{2.12}$$

hvor $x_0 = 1$ og t_c angiver en linear kombination for en gruppe c, $c = 1, \ldots, C-1$. Med udgangspunkt i logit betyder dette, at der nu undersøges logaritmen af posterior sandsynligheden for gruppe c, som den positive gruppe, divideret med posterior sandsynligheden for referenceniveauet, altså hvad der bliver tilsvarende den komplementære gruppe. Dermed er logit for multinomial logistisk regression givet som følger.

Proposition 2.7. Logit, multinomial logitisk regression
Logit for en multinomial logistisk regressionsmodel er givet ved

$$L_c(X) = \log\left(\frac{\mathrm{P}(y = c|X = x_{ij})}{\mathrm{P}(y = C|X = x_{ij})}\right) = \sum_{j=0}^{p} \beta_{j,c} x_j. \tag{2.13}$$

hvor $\beta_c = (\beta_{0,c}, \beta_{1,c}, \ldots, \beta_{p,c})$ er en koefficientvektor for gruppe c og $x_0 = 1$.

Logit gør det nemmere at fortolke koefficientvektoren β_c. Koefficienterne i modellen kan variere efter valg af referenceniveau, men prædikationerne vil forblive de samme uanset valg af referenceniveau. Fra Proposition 2.3 ses det, at logit funktionen er udledt af at anvende logaritmen af odds, hvorfor odds for multinomial logistisk regression kan udledes fra (2.13) ved at anvende eksponential-funktionen, altså

$$O_c(X) = \frac{\mathrm{P}(y = c|X = x_{ij})}{\mathrm{P}(y = C|X = x_{ij})} = \exp\left(\sum_{j=0}^{p} \beta_{j,c} x_j\right),$$

hvor $x_0 = 1$. Dermed er odds for multinomial logistisk regression givet som følger.

Proposition 2.8. Odds, multinomial logistisk regression
Odds for en multinomial logistisk regression er givet ved

$$O_c(X) = \frac{\mathrm{P}(y = c|X = x_{ij})}{\mathrm{P}(y = \mathrm{C}|X = x_{ij})} = \exp\left(\sum_{j=0}^{p} \beta_{j,c} x_j\right), \tag{2.14}$$

hvor $x_0 = 1$, og $O_c(X)$ kan antage enhver værdi i intervallet $(0, \infty)$.

Bevis:
Antag, at $O_c(t_c) \in (0, \infty)$ og anvend t_c defineret i (2.12). Så følger resultatet af Proposition 2.2. ■

Imidlertid minder denne notation ikke nødvendigvis om den, der blev præsenteret ved odds for logistisk regression og multipel logistisk regression. Dog kan en lignende notation udledes. Definer en funktion $h : \mathbb{R} \to [0, 1]$, sådan at

$$\frac{h_c(X)}{h_C(X)} = \frac{\mathrm{P}(y = c|X = x_{ij})}{\mathrm{P}(y = C|X = x_{ij})} = O_c(X), \tag{2.15}$$

hvorom det gælder, at $\sum_{c=1}^{C} h_c(X) = 1$. Så er

$$1 = \sum_{c=1}^{C} h_c(X) = h_C(X) + \sum_{c=1}^{C-1} h_c(X) = h_C(X) + \sum_{c=1}^{C-1} h_C(X) O_c(X) = h_C(X)\left(1 + \sum_{c=1}^{C-1} O_c(X)\right)$$

Ved at isolere for h_C fås

$$h_C(X) = \frac{1}{1 + \sum_{c=1}^{C-1} O_c(X)},$$

og fra (2.15) ses det, at

$$h_c(X) = O_c(X) h_C(X) = \frac{O_c(X)}{1 + \sum_{c=1}^{C-1} O_c(X)}.$$

Dermed kan den multinomiale logistiske regressionsmodel defineres som følger.

Definition 2.9. Multinomial logitisk regression
Lad $h_c(X) \in [0, 1]$. En multinomial logistisk regressionsmodel er defineret ved

$$h_c(X) = \mathrm{P}(y = c|X = x_{ij}) = \frac{\exp\left(\sum_{j=0}^{p} \beta_{j,c} x_j\right)}{1 + \sum_{c=1}^{C-1} \exp\left(\sum_{j=0}^{p} \beta_{j,c} x_j\right)} \tag{2.16}$$

for $c = 1, \ldots, C-1$ og $\beta_c = (\beta_{0,c}, \beta_{1,c}, \ldots, \beta_{p,c})$.

Til estimation af koefficientvektoren β_c anvendes en udvidelse af likelihood funktionen i (2.7), som er givet i [6, s. 10-12]:

$$\mathcal{L}(\beta_c) = \prod_{i=1}^{n} \prod_{c=1}^{C-1} \left(\frac{h_c(X_i)}{h_C(X_i)} \right)^{y_{i,c}} \cdot h_C(X_i)^{n_i}, \tag{2.17}$$

hvor $y_{i,c}$ angiver antallet af observationer i den c'te gruppe ifølge responsvariablen. $\hat{\beta}_c$ vælges til at maksimere (2.17). Når $\hat{\beta}_c$ er estimeret, kan posterior sandsynligheden for en given klassifikation udregnes.

I tilfælde af $p = 2$ og $C = 3$, vil (2.17) være givet ved

$$\mathcal{L}(\beta_c) = \prod_{i=1}^{n} \prod_{c=1}^{2} \frac{\exp\left(\beta_{0,c} + \beta_{1,c} X_{1,i} + \beta_{2,c} X_{2,i}\right)^{y_{i,c}}}{\left(1 + \sum_{j=0}^{2} \exp\left(\beta_{0,j} + \beta_{1,j} X_{j,i} + \beta_{2,j} X_{j,i}\right)\right)^{-n_i}}.$$

Bemærkning. *I dette kapitel er funktionen h_c konstrueret med henblik på at vise, hvordan MLR udledes fra logistisk regression. For at lette notationen i rapporten, vil vi fremover benytte notationen*

$$P_c = \frac{P(y = c | X = x_{ij})}{P(y = C | X = x_{ij})}.$$

3 K-fold krydsvalidering

Dette kapitel er baseret på [4, ap. 5.1] og [7, kap. 7.10].

I denne rapport vil der fokuseres på evalueringsmetoden krydsvalidering, som er en *resampling*-metode. Denne tilgang kan give mulighed for at få oplysninger, som ikke kunne fås ved at træne modellen én gang ved hjælp af den oprindelig opdeling af datasættet. Dette kan være oplysninger om stabiliteten af modellen, eller om testudfaldet har tilfældige eller generelle tendenser.

I K-fold krydsvalidering opdeles datasættet i K-folder, og en model trænes på $K-1$ folder, og testdataet vælges som den k'te fold. Målet med at udføre K-fold krydsvalidering er at bestemme, hvor godt en model klassificerer nye observationer. Hvis responsvariablen y er kvalitativ, kan testfejlen eksempelvis angive antallet af fejlklassifikationer. Fejlfrekvensen (3.1) i klassifikationsproblemer udregnes ved gennemsnittet af testfejl fra hver fold k,

$$CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K}\text{Err}_i \tag{3.1}$$

hvor $\text{Err}_i = \sum_{i\in k}\mathbb{1}(y_i \neq \hat{y}_i)$ er testfejlen. Fejlfrekvensen $CV_{(K)}$ udregnes analogt for træningsdata og testdata. K-folds krydsvalidering implementeres i Algoritme 1.

Algoritme 1 K-fold krydsvalidering

1. Inddel observationerne vilkårligt i K folder.
2. For hver fold $k = 1, 2, \ldots, K$
   (a) Vælg fold k som testdata og anvend de resterende $K-1$ folder som træningsdata.
   (b) Tilpas og træn modellen på træningsdata.
   (c) Test den tilpassede model på testdata, fold k.
   (d) Bevar testfejlen og kassér modellen.
3. Opsummer testfejl og udregn fejlfrekvensen $CV_{(K)}$.

4 Præsentation af data

Dette kapitel baseret på [3], [2, kap. 4] og med inspiration fra [8] og [9].

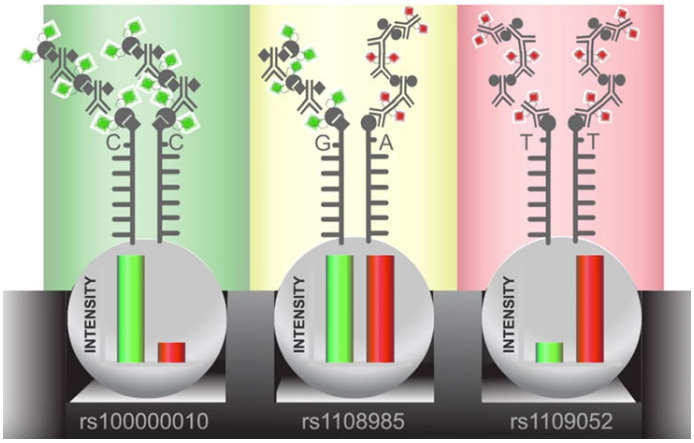
I denne rapport behandles et datasæt udleveret af Retsgenetisk Afdeling, Retsmedicinsk Institut på Københavns Universitet. Datasættet består af et udpluk af målinger fra snipchipper og *sande* genotyper. Snipchipperne indeholder 4,3 millioner positioner, som ligger spredt i et genom. Indledningvist i rapporten blev det introduceret, at et genom indeholder den DNA-kode, som udgør det genetiske arvemateriale i alle levende organismer.

Et genom består af 46 kromosoner, altså 23 kromosonpar, som opdeles i 22 autosomale kromosonpar og et kønskromosonpar. Kromosoner findes i par, idet individet har et kromoson fra hver af sine forældre. I et kromoson findes alleller i bestemte positioner, hvor en allel vil fortælle om indholdet i dette kromoson. Dermed kan det siges, at alleler er en tilstand i et genom og markerer de genetiske markører i DNAet. Alleler kan have størrelse fra én base til et helt gen, hvor en variation i én base i en bestemt position kaldes SNP. SNPer udgør over 90% af de genetiske variationer i arvematerialet i et menneske, hvorfor disse har stor interesse, når variationer i DNAet mellem to eller flere mennesker undersøges. En genotype dannes af to alleler. Hvis samme basepar er tilgængeligt i begge alleler, kaldes allelen homozygot, og hvis baseparerne er forskellige kaldes allelen heterozygot. Til at kortlægge, hvilke basepar, der er tilgængelige i en bestemt SNP, anvendes processen DNA-sekventering. I denne proces bliver basesekvensen kortlagt med udgangspunkt i sekventeringmetoden af hele genomet, kaldt helgenomsekventering (WGS), eller den billigere metode Next Generation Sequencing (NGS).

I NGS kortlægges udvalgte SNPer i genomet, hvor der i det udleverede datasæt anvendes Illuminas BeadArray-teknologi, hvor den anvendte snipchip er beklædt med perler. Perlerne målretter sig specifikke positioner i genomet. Disse positioner kaldes loci. Sekventeringen på snipchippen består af at aflæse en base af gangen på markerede nukleotider. I sekventeringscyklussen påsættes én nukleotid til DNA-fragmenter i klyngerne på perlerne. Den valgte flowcelle behandles derefter med laserlys, hvorfra de markerede nukleotider vil emittere lys. Dette er illustreret i Figur 4.1.

De markerede nukleotider emitterer et rødt lys for baserne A og T samt et grønt lys for baserne C og G. På grund af den kemiske opbygning af DNA, vil en base bestemme dens komplementære base på den anden streng, hvorfor A og T bliver registreret som et signal, og C og G registreres som et andet. Genotyper bestemmes ud fra intensiteten af de to signaler. En overvægt af enten det røde eller grønne lyssignal vil betyde, at personen i alle aflæsninger af en specifik SNP har samme base i begge alleler og klassificeres som en homozygot genotype. Hvis lysintensiteten fra begge signaler er tilnærmelsesvis lige store, fås en heterozygot genotype.

Bemærkning. *I denne rapport vil lysintensitenterne af signalerne blive kaldt signalintensiteterne.*



Figur 4.1: Illustration fra [3] af det emitterede lys fra de markerede nukleotider. Illustrationen viser signalet i en enkelt syntesecyklus af tre forskellige perler. Her ses eksempler på, hvordan hver af de tre genotyper i bestemte SNP'er kan prædikteres.

Det udleverede datasæt består af 4400 observationer tilsvarende 1100 SNPer fra fire anonymiserede personer. Intensiteten af de emmiterede lyssignaler behandles som et A- og B-signal, og vil fungere som datasættets primære variabler. Da hver SNP er binær, er der to mulige alleler. Den ene kaldes reference allel, og den anden kaldes alternativ allel. Informationen om, hvilke baser det drejer sig om for hver SNP, kan findes i den såkaldte manifest-fil (se [10]). Signalerne A og B er navngivet konventionelt, og dermed er det ukendt, hvorvidt det er baserne A/T eller C/G, der er tilstede i eksempelvis A-signalet eller B-signalet.

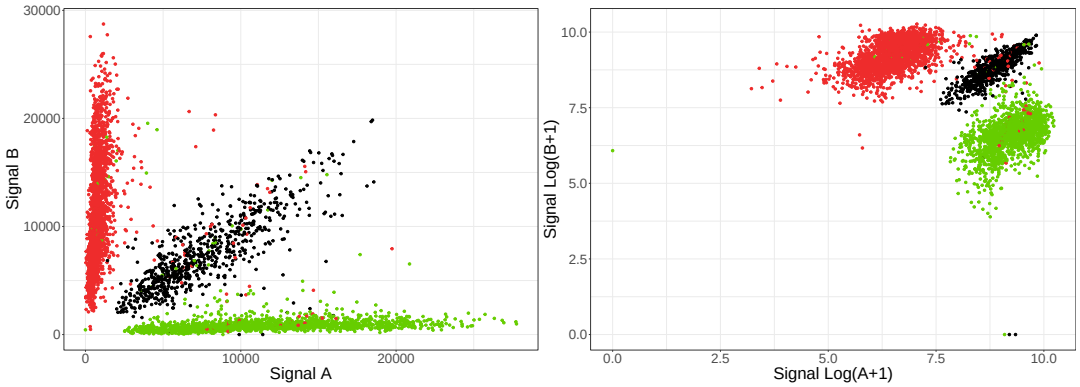
Datasættet anonymiseres, sådan at A- og B-signalerne ikke korresponderer til henholdsvis det røde og grønne signal, og SNP-navnene nummeres $1, 2, \ldots, 1100$. På denne måde kan datasættet kun anvendes til at afgøre, om en af testpersonerne har en homozygot genotype eller en heterozygot genotype i en given SNP, men ikke hvilke baser som er til stede i genotyperne. På den måde navngives homozygote genotyper AA og BB og den heterozygote genotype navngives AB. Nedenfor ses de første fire observationer i datasættet.

	SNP	Person	A_sig_Mean	A_sig_SD	A_sig_N	B_sig_Mean	B_sig_SD	B_sig_N	WGS
1	1	1	10792	1316	18	558	387	12	AA
2	1	2	13404	1553	14	491	367	9	AA
3	1	3	7837	1151	18	8356	871	14	AB
4	1	4	13097	1481	16	581	405	8	AA

For hver observation er der registreret ni oplysninger, som i denne rapport betragtes som ni variabler. De første variabler i datasættet noterer, hvilken SNP der bearbejdes, samt hvilken person denne SNP tilhører. De tre variabler, som primært bruges fra datasættet, er *A_sig_Mean*, *B_sig_Mean* og *WGS*. A_sig_Mean og B_sig_Mean angiver gennemsnittet for hvert signal for hver observation, hvor antallet af lyssignaler findes under henholdvis *A_sig_N* og *B_sig_N*, og *A_sig_SD* og *B_sig_SD* angiver standardafvigelsen. Variablen WGS angiver den *sande* genotype på hver observation opnået ved hjælp af metoden WGS, der betragtes som den gyldne standard.

Rapportens formål er at bestemme genotypen af hver observation ud fra de givne signaler ved hjælp af supervised learning. Her trænes en MLR-model på baggrund af A_sig_Mean, B_sig_Mean og responsvariablen WGS til at klassificere genotyper på fremtidige observationer. I denne rapport udtages der 4000 observationer til træningsdata, og 400 observationer til test af modellen fra de i alt 4400 observationer.

I Figur 4.2a illustreres datasættet ud fra signalerne A_sig_Mean og B_sig_Mean, hvor observationerne fordeler sig i tre klynger tilsvarende de tre genotyper bestemt med WGS. De homozygote genotyper AA og BB er henholdsvis grøn og rød, og den heterozygote genotype AB er sort. Her ses det, at klyngerne ligger tæt omkring origo, men ved anvendelse af en log-transformation af signalerne, med den naturlige logaritme, adskilles klyngerne mere fra hinanden. Dette er illustreret i Figur 4.2b, hvor det ses, at klyngerne adskiller sig mere samt at klynger er mere kompakte. Log-transformationen bidrager til at stabilisere variansen, samt at gøre datasættet mere normalfordelt. Når signalerne log-transformeres benyttes $\log(X+1)$, da nogle observationer ikke har et registreret A- eller B-signal. For at undgå problematikken om $\log(0)$ lægges 1 til alle signaler, og dermed ligger alle log-transformerede intensiteter i intervallet $[0, \infty)$.



(a) Signal på rå data **(b)** Signal på log-transformeret data

Figur 4.2: Visualisering af de tre genotyper: AA (grøn), BB (rød) og AB (sort).

Bemærkning. *I denne rapport anvender vi i matematiske udtryk notationen* X_A *for* $\log(A+1)$ *og* X_B *for* $\log(B+1)$.

5 Klassifikation af genotyper

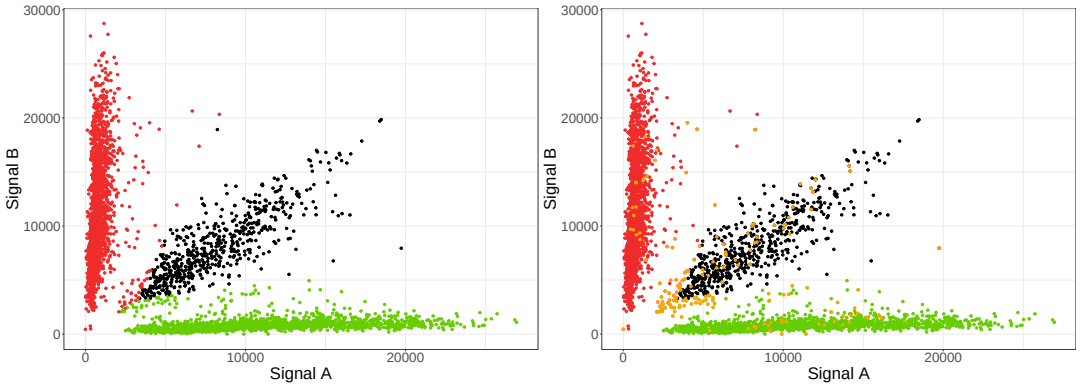
I dette kapitel fremlægges modeller til klassifikation af genotyper med udgangspunkt i MLR, som præsenteret i Kapitel 2. Scripts findes i Github (se [1]).

5.1 Klassifikation med MLR

MLR-modellen kan benytte en af de tre genotyper AA, BB eller AB som referenceniveau. Modellens koefficienter afhænger af valget af referenceniveau, men posterior sandsynligheden for, at en observation bliver klassificeret som en bestemt genotype, forbliver det samme uanset valg af referenceniveau. I denne rapport udarbejdes modellerne med genotypen AB som referenceniveau.

Først trænes MLR-modellen på datasættets første 4000 observationer for at undersøge, om der opnåes en bedre klassifikationsmodel på rå-data eller på log-transformeret data. Modellens præcision, målt på *accuracy*, vurderes ud fra den procentvise andel af korrekt klassificeret genotyper sammenlignet med responsvariablen WGS.

Klyngedannelsen baseret på rå-data er illustreret i Figur 5.1a, og denne model giver en modelpræcision på 95.8%. Figur 5.1b illustrerer, hvor modellen laver fejlklassifikationer, og hvor disse er placeret.



(a) MLR på rå-data.

(b) Fejlklassifikationer med MLR på rå-data.

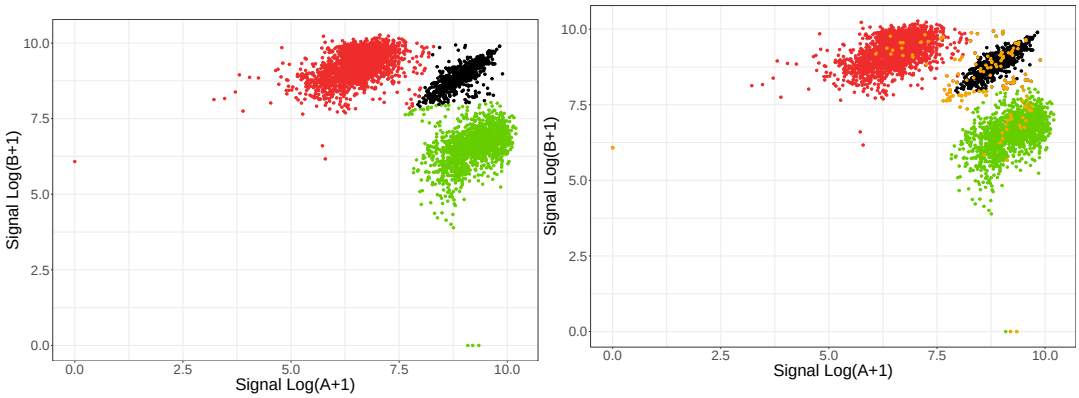
Figur 5.1: Illustration af klassifikation med MLR på rå-data. Genotype AA er grøn, BB er rød, AB er sort og fejlklassifikationer er orange.

		MLR			
		AA	**AB**	**BB**	**Total**
WGS	**AA**	1473	16	21	1510
	AB	46	534	37	617
	BB	26	24	1823	1873
	Total	1545	574	1881	4000

Tabel 5.1: Konfusionsmatrix for rå-data: Antallet af observationer i hver genotype bestemt af MLR sammenlignet med WGS. Modelpræcision 95.8%.

I Tabel 5.1 ses en konfusionsmatrix for MLR-modellen på rå-data, som visualiserer, hvilke observationer som klassificeres korrekt, og hvilke observationer som fejlklassificeres sammenlignet med responsvariablen WGS. MLR-modellen kalder en genotype for en observation der, hvor posterior sandsynligheden er størst. Søjlerne AA, BB og AB viser modellens prædikationer, hvor rækkerne AA, BB og AB viser observationernes klassifikation med responsvariablen WGS. I tabellen fremgår antallet af fejlklassifikationer udenfor diagonalen. Her ses det blandt andet, at for det antal observationer responsvariablen WGS tildeler genotypen AB, så fejlklassificerer MLR-modellen hhv. 46 og 37 observationer som AA og BB.

I MLR med log-transformeret data opnås en bedre procentdel af korrekt klassificerede observationer, 96.3%. Disse klassifikationer er illustreret i Figur 5.2a, og Figur 5.2b illustrerer, hvor modellen laver fejlklassifikationer, og hvor disse er placeret. I Tabel 5.1 fås 170 fejlklassifikationer, hvor der fås 149 fejlklassifikationer i Tabel 5.2. Derfor anses log-transformationen som et væsentligt værktøj i forhold til klyngeinddeling, da det giver en bedre modelpræcision samt log-tranformationen formindsker afstanden mellem alle observationerne og andelen af fejlklassifikationer.



(a) MLR på log-transformeret data.

(b) Fejlklassifikationer med MLR på log-transformeret data.

Figur 5.2: Illustration af klassifikation med MLR på log-transformeret data. Genotype AA er grøn, BB er rød, AB er sort og fejlklassifikationer er orange.

		MLR			
		AA	**AB**	**BB**	**Total**
WGS	**AA**	1459	32	19	1510
	AB	29	574	14	617
	BB	22	33	1818	1873
	Total	1510	639	1851	4000

Tabel 5.2: Konfusionsmatrix for log-transformeret data: Antallet af observationer i hver genotype bestemt af MLR sammenlignet med WGS. Modelpræcision 96.3%.

I denne rapport tager vi udgangspunkt i MLR-modellen med log-transformeret data, og modellen kan opskrives som i Definition 2.9:

$$P_c = \frac{\exp\left(\hat{\beta}_{0,c} + \hat{\beta}_{1,c}X_A + \hat{\beta}_{2,c}X_B\right)}{1 + \sum_{l \in \{AA,BB\}} \exp\left(\hat{\beta}_{0,l} + \hat{\beta}_{1,l}X_A + \hat{\beta}_{2,l}X_B\right)} \tag{5.1}$$

hvor $c = \{\text{AA, BB}\}$, da referenceniveauet er valgt til AB. Da fås følgende koefficienter

$$\hat{\beta}_{AA} = \begin{bmatrix} \hat{\beta}_{0,AA} \\ \hat{\beta}_{1,AA} \\ \hat{\beta}_{2,AA} \end{bmatrix} = \begin{bmatrix} 18.47 \\ 0.31 \\ -2.66 \end{bmatrix} \quad \text{og} \quad \hat{\beta}_{BB} = \begin{bmatrix} \hat{\beta}_{0,BB} \\ \hat{\beta}_{1,BB} \\ \hat{\beta}_{2,BB} \end{bmatrix} = \begin{bmatrix} 16.32 \\ -2.66 \\ 0.59 \end{bmatrix}. \tag{5.2}$$

Posterior sandsynligheden for AB udregnes ved

$$P_{AB} = \frac{1}{1 + \sum_{l \in \{AA,BB\}} \exp\left(\hat{\beta}_{0,l} + \hat{\beta}_{1,l}X_A + \hat{\beta}_{2,l}X_B\right)}.$$

5.2 Bestemmelse af beslutningsgrænser

Selvom MLR-modellen for log-transformeret data viste sig at være bedre til klassifikation af genotyper, lykkedes det ikke for denne model at klassificere observationerne perfekt med hensyn til responsvariablen WGS. Modellen opnåede en præcision på 96.3%, men havde forsat 149 fejlklassifikationer. MLR-modellen er trænet ud fra WGS, hvor en observations genotype bestemmes ud fra den største posterior sandsynlighed. Dog sker der forsat fejlklassifikationer, hvor det til enhver tid vil være bedre ikke at klassificere en genotype end at fejlklassificere en genotype. En del af rapportens formål er at udarbejde en model, der kan prædiktere genotyper for nye observationer ud fra signalintensiteter, og der introduceres derfor i dette afsnit en ny gruppe, *no-call*, samt der fremlægges en metode til at tildele observationer til no-call.

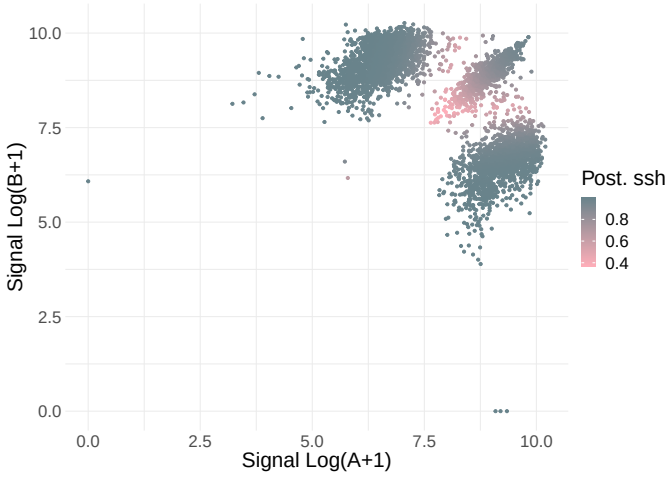
Denne metode baseres på posterior sandsynlighederne fra træningsdatasættet bestemt af MLR-modellen, hvor der herudfra udarbejdes et grid, som afspejler områderne for de tre forskellige genotyper og den nye gruppe no-call. På den måde skabes en visualisering af, hvordan områderne for hver genotype defineres ud fra posterior sandsynligheder, og hvilke

områder, hvor modellen er sikker nok til at kalde en genotype. I den forbindelse bliver det også tydeliggjort, hvilke områder der defineres som no-call, altså hvor det er for usikkert at kalde en genotype, når vi fastsætter en bestemt tærskel ω.

5.2.1 Beslutninggrænse for no-call bånd

For at bestemme en beslutningsgrænse for no-call bånd tages der udgangspunkt i posterior sandsynlighederne bestemt med MLR-modellen. Det ønskes at definere et område, hvor der er for usikkert at kalde en genotype på en observation. Grænsen til dette område bestemmes ved en tærskel, som angiver, hvor sikker en modellen skal være før en genotype bliver kaldt. Hvis den maksimale posterior sandsynlighed ligger under denne tærskel tildeles observationen gruppen no-call.

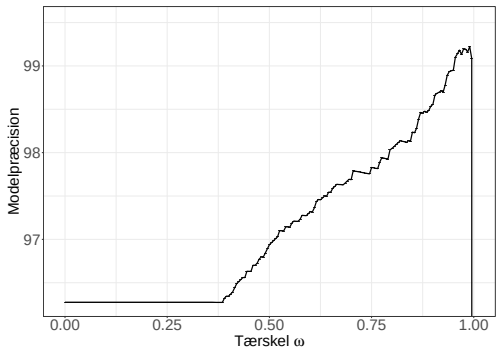
Der findes flere forskellige måder at bestemme en tærskel for posterior sandsynlighed. En sådan tærskel er oftest bestemt ud fra formålet med analysen og krav fra udbyder. I denne rapport er der intet krav fra udbyder, hvorfor vi vælger at fokusere på at træne en model til at være så sikker som mulig i sine klassifikationer. Til at bestemme en tærskel for, hvor sikker en model skal være for at kalde en genotype betragtes Figur 5.3, som illustrerer den højeste posterior sandsynlighed for hver observation med en farveskala.



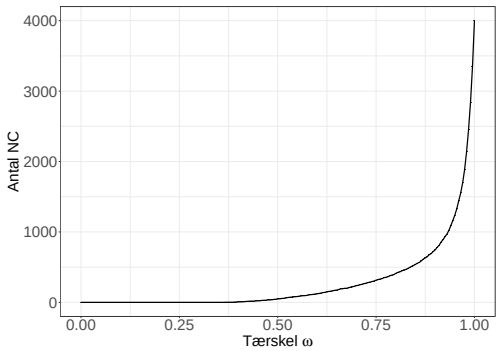
Figur 5.3: Illustration med farveskala af den maksimale posterior sandsynlighed for hver observation.

Her observeres det, at de fleste posterior sandsynligheder ligger over 80%. Vælges denne tærskel vil en stor del af klyngen for genotype AB tildeles no-call, og dette medfører, at en stor andel af korrekt klassificerede observationer tildeles no-call. Til trods for, at nogle observationer er korrekt klassificerede i henhold til responsvariablen WGS, er posterior sandsynligheden ikke tilstrækkelig høj til, at modellen er sikker i sin klassifikation, og derfor bliver nogle korrekt klassificerede observationer tildelt no-call. Vælges tærsklen til 70% vil en mindre andel af observationer tildeles no-call. Heriblandt vil andelen af korrekt klassificerede observationer, som tildeles no-call, være mindre end ved 80%, men det vil være på bekostning af modellens præcision.

I Figur 5.4 visualiseres det, hvordan modelpræcisionen ændrer sig, når der fastsættes en tærskel for, hvor sikker MLR-modellen skal være før, der kaldes en genotype. Her ses det, at modelpræcisionen stiger i takt med, at tærsklen nærmer sig 1, men modelpræcision falder til tilnærmelsvist 0, hvis tærsklen vælges til 1.



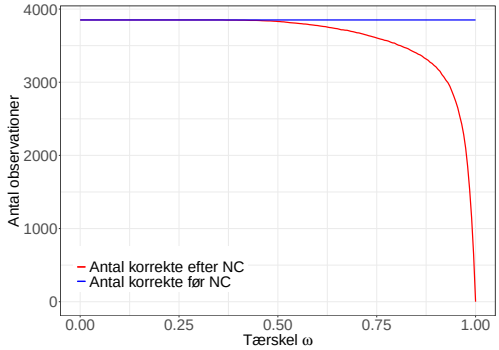
Figur 5.4: Modelpræcision ved tærskel ω efter no-call.



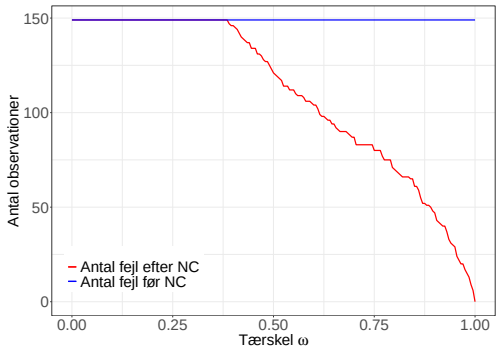
Figur 5.5: Forholdet mellem observationer, som tildeles no-call og tærsklen ω.

Når modelpræcisionen forsat stiger frem til en tærskel på 1, kan dette indikere at valget om en høj tærskel, vil være det bedste valg for MLR-modellen. Et valg om en høj tærskel vil være på bekostning af både antallet af korrekt klassificerede observationer, fejlklassikationer og andelen af no-call, da andelen af observationer, som tildeles no-call er større desto højere tærskel, der vælges. De observationer, som tildeles no-call kan både være korrekt klassificerede observationer og fejlklassifikationer. Hvilke observationer, som tildeles no-call afhænger kun af observationens posterior sandsynlighed. Dette fremgår af Figur 5.5.

På Figur 5.6a illustreres forholdet mellem antallet af korrekt klassificerede observationer og valg af tærskel, og på Figur 5.6b illustreres forholdet mellem antallet af fejlklassifikationer og valg af tærskel. Her fremgår det mere tydeligt, at en høj tærskel skaber et markant fald i antallet af fejlklassifikationer sammenlignet med faldet i antallet af korrekt klassifikationer.



(a) Forhold: korrekte klassifikationer og tærsklen ω.



(b) Forhold: fejlklassifikationer og tærsklen ω.

Figur 5.6: Funktioner af forholdet mellem korrekte klassifikationer, fejlklassifikationer og tærsklen ω.

Det betyder, at der er en betydelig sammenhæng mellem valg af tærskel, mængden af no-call, korrekt klassificerede observationer og fejlklassifikationer. Med udgangspunkt i disse overvejelser og Figur 5.3, arbejdes der i denne rapport videre med en tærskel på 70%.

5.2.2 Konstruktion af no-call bånd

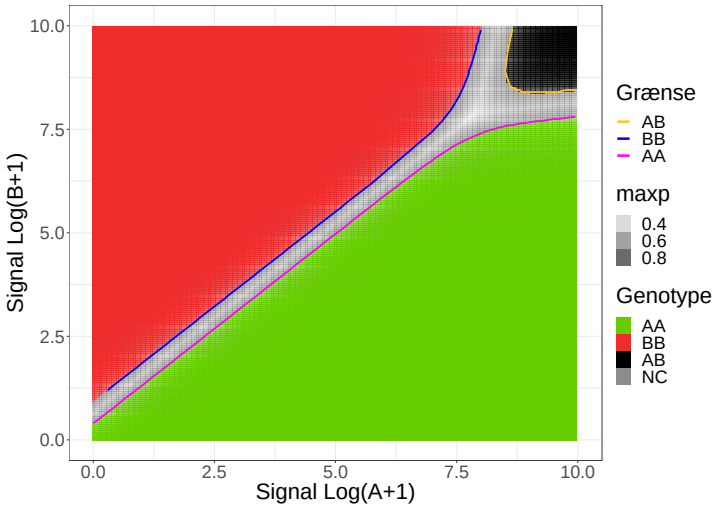
Visualiseringen af områderne for de tre genotyper og no-call baseret på den maksimale posterior sandsynlighed fra MLR-modellen er illustreret i Figur 5.7. Det røde område afspejler, at hvis en ny observation falder indenfor dette område, vil observationen klassificeres som genotype BB. Hvis den falder indenfor det sorte eller grønne område, klassificeres en ny observation som hhv. genotype AB eller genotype AA. Falder en observation derimod indenfor det grå no-call bånd, tildeles observationen no-call. No-call båndet bestemmes af tærsklen for posterior sandsynligheden, hvor MLR-modellen defineret i (5.1):

$$P_c = \frac{\exp(\hat{\beta}_{0,c} + \hat{\beta}_{1,c}X_A + \hat{\beta}_{2,c}X_B)}{1 + \sum_{l \in \{AA,BB\}} \exp(\hat{\beta}_{0,l} + \hat{\beta}_{1,l}X_A + \hat{\beta}_{2,l}X_B)},$$

hvor $c \in \{AA, BB\}$ og

$$P_{AB} = \frac{1}{1 + \sum_{l \in \{AA,BB\}} \exp(\hat{\beta}_{0,l} + \hat{\beta}_{1,l}X_A + \hat{\beta}_{2,l}X_B)},$$

anvendes til at bestemme områderne, hvor modellen er sikker i at klassificere en genotype.

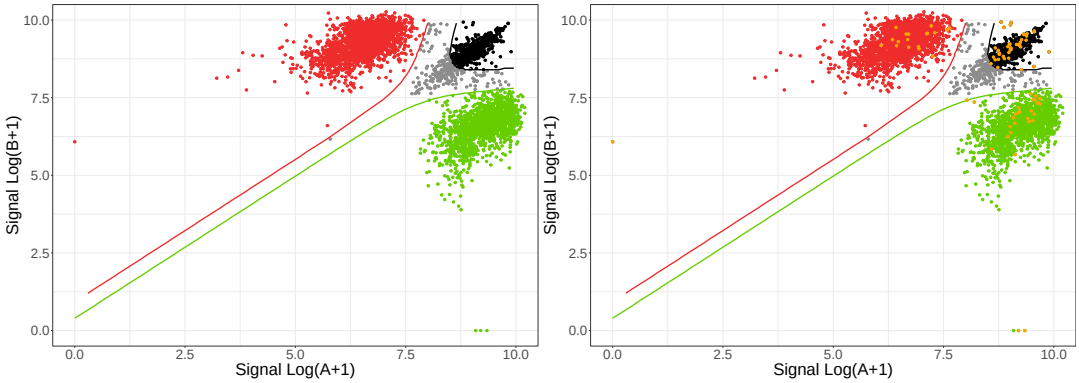


Figur 5.7: Visualisering af MLR-modellens posterior-områder.

Størrelsen af no-call båndet afhænger af, hvilken tærskel for posterior sandsynligheden, der fastsættes. Grænsen for hvert område bestemmes ved at undersøge, hvornår $P_c < \omega$ og $P_C < \omega$, hvor ω angiver tærsklen. Ved en fastsat tærskel $\omega = 0.7$, betragtes det, hvornår en observation med en posterior sandsynlighed på minimum 70% kan blive klassificeret som

en genotype AA, BB eller AB. Hvis posterior sandsynligheden er mindre end 70% tildeles observationen no-call. Grænserne definerer seperationslinjerne mellem de tre genotyper og no-call båndet.

I Figur 5.8a illustreres klyngetildelingen fra Figur 5.2 med no-call bånd. De observationer, som tildeles no-call, er markeret med grå. I Figur 5.8b illustreres fejlklassifikationerne med MLR-modellen og deres placering efter no-call.



(a) MLR med no-call bånd. **(b)** Fejlklassifikationer med no-call bånd.

Figur 5.8: Illustration af klassifikation af træningsdata med MLR og no-call bånd. Genotype AA er grøn, BB er rød, AB er sort no-call er grå og fejlklassifikationer er orange.

I Tabel 5.3 ses konfusionsmatricen for MLR-modellen med en tærskel for no-call på 70%. No-call er markeret med rød. Modellen klassificerer 97.7% af træningsdataet korrekt sammenlignet med responsvariablen WGS, men modellen tildeler 237 observationer til no-call, tilsvarende 5.9% af træningsdataet. Det er bemærkelsesværdig fra Tabel 5.3, at modellen fortsat fejlklassificerer 87 observationer efter no-call. Ved test af modellen, er det derfor forventeligt, at fejlklassifikationer kan opstå. Dette undersøges i Afsnit 5.3.

		MLR				
		AA	**AB**	**BB**	**NC**	**Total**
WGS	**AA**	1448	13	18	31	1510
	AB	7	427	1	182	617
	BB	21	27	1801	24	1873
	Total	1476	467	1820	237	4000

Tabel 5.3: Konfusionsmatrix: Klassifikation med MLR-modellen på træningsdata og andelen af observationer som tildeles no-call sammenlignet med responsvariablen WGS. No-call er markeret med rød.

5.3 Test af model

Da det ønskes at kunne forudsige genotypen af nye observationer ud fra signalintensiteter, foretages der i dette afsnit to forskellige tests af MLR-modellen. Ved at evaluere modellen

ses det, hvordan modellen klassificerer nye observationer, som den ikke er trænet på. Første test baseres på de 400 observationer fra 4000/400-opdelingen, som blev præsenteret i Kapitel 4. Den anden test baseres på 11-fold krydsvalidering, som vil give et mere realistisk billede af modellens præcision ved at træne og teste den på tilfældige opdelinger af datasættet, som i gennemsnit opfylder en 4000/400-opdeling. Det ønskes at undersøge, om den procentvise andel af no-call ændrer sig.

5.3.1 Test på 400 observationer

I første test benyttes MLR-modellen til at forudsige genotyper på 400 nye observationer. Dette gøres ved først at betragte posterior sandsynligheden for en ny observation x_{n+m}, $m = 1, \ldots, 400$, ud fra (5.1), dvs.

$$P_c = \frac{\exp\left(\hat{\beta}_{0,c} + \hat{\beta}_{1,c} X_A + \hat{\beta}_{2,c} X_B\right)}{1 + \sum_{l \in AA,BB} \exp\left(\hat{\beta}_{0,l} + \hat{\beta}_{1,l} X_A + \hat{\beta}_{2,l} X_B\right)}, \tag{5.3}$$

hvor $c = \{\text{AA, BB}\}$ og

$$P_{AB} = \frac{1}{1 + \sum_{l \in AA,BB} \exp\left(\hat{\beta}_{0,l} + \hat{\beta}_{1,l} X_A + \hat{\beta}_{2,l} X_B\right)}, \tag{5.4}$$

da referenceniveauet af valgt til AB. Koefficienterne $\hat{\beta}_{AA}$ og $\hat{\beta}_{BB}$ er bestemt i (5.2). Herefter forudsiges en genotype først, når det vides, hvorvidt en ny observations posterior sandsynlighed overstiger den fastlagte tærskel. Hvis tærsklen ikke overstiges tildeles observationen no-call. Fremgangsmåden hertil er beskrevet i Algoritme 2.

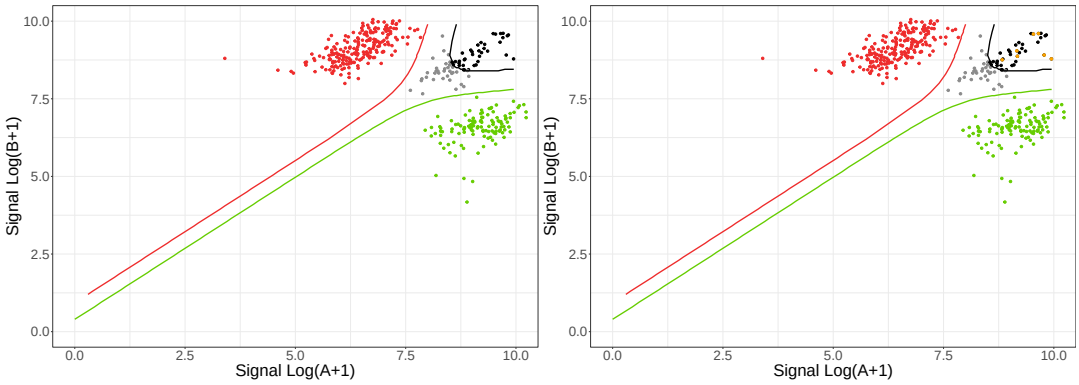
Algoritme 2 Klassifikation af genotype for nye observationer

Input: En ny observation x_{n+m}, koefficienterne fra (5.2) og fastlagt tærskel ω.

1. Beregn posterior sandsynlighed vha. MLR-modellen i (5.3) og (5.4).
2. Tjek hvorvidt $\max_c (P_c) > \omega$, $c = \{AA, BB, AB\}$, hvis
   Sand: fortsæt til trin 3.
   Falsk: tildel observation til no-call.
3. Find maksimal posterior sandsynlighed og kald genotype herefter.

Output: AA, BB, AB eller NC.

Trin 3 indikerer, at når en genotype kaldes, tildeles observation x_{n+m} til klyngen, hvor den har størst posterior sandsynlighed. I Figur 5.9a illustreres de 400 observationer fra testdataet klassificeret ud fra Algoritme 2, hvor Figur 5.9b illusterer, hvor fejlklassifikationerne placerer sig efter no-call. Her ses det, at de observationer som bliver tildelt no-call placerer sig i samme område som det illustreres på Figur 5.8.



(a) Testdata med MLR og no-call bånd.

(b) Fejlklassifikationer med no-call bånd.

Figur 5.9: Illustration af klassifikation af testdata med MLR og no-call bånd. Genotype AA er grøn, BB er rød, AB er sort, no-call er grå og fejlklassifikationer er orange.

I Tabel 5.4 ses det, hvordan observationerne placeres ved klassificering af genotype sammenlignet med WGS for testdataet samt andelen og placering af no-call, som er markeret med rød. MLR-modellen klassificerer 98.1% af alle de nye observationer korrekt sammenlignet med responsvariablen WGS efter no-call, men modellen tildeler 36 observationer til no-call, tilsvarende 9% af testdataet. Det er bemærkelsesværdig, at modellen forsat fejlklassificerer 7 observationer efter no-call. Dette blev ligeledes observeret for træningsdataet i Tabel 5.3, hvorfor det er forventelig, at det også ville ske for testdata.

		MLR				
		AA	**AB**	**BB**	**NC**	**Total**
WGS	**AA**	114	4	0	1	119
	AB	0	30	0	34	64
	BB	0	3	213	1	217
	Total	114	37	213	36	400

Tabel 5.4: Konfusionsmatrix: Klassifikation med MLR-modellen på testdata og placeringen af observationer som tildeles no-call sammenlignet med responsvariablen WGS. No-call er markeret med rød.

Denne test er afhængig af opdelingen af datasættet, og vil derfor ikke nødvendigvis give et retvisende billede af, hvor god MLR-modellen er til at klassificere nye observationer. Testfejlen bliver i denne test kun estimeret på en enkelt opdeling. Da det kan være tilfældigt, at netop disse 400 observationer vil give et godt testresultat, evalueres MLR-modellen med 11-fold krydsvalidering i Afsnit 5.3.2.

5.3.2 Krydsvalidering af klassifikation med MLR

I anden test benyttes MLR-modellen til at forudsige genotyper på nye observationer med 11-fold krydsvalidering. Da det udleverede data består af 4400 observationer er 11-fold valgt, sådan at inddelingen af data i gennemsnit opfylder en 4000/400-inddeling. Denne

inddeling ønskes for at kunne sammenligne resultaterne fra krydsvalideringen med resultaterne i Afsnit 5.3.1.

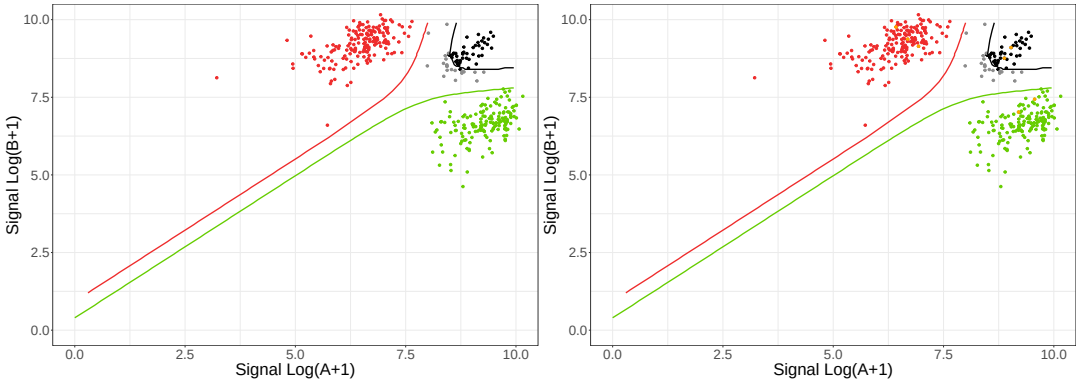
Algoritme 3 K-fold krydsvalidering

Input: data, model, antal folder K og fastlagt tærskel ω.

1. Inddel data i K tilfældige folder.
2. For hver fold $k = 1, 2, \ldots, K$:
   (a) Brug $\{1, \ldots, K\} \backslash \{k\}$ folder som træningsdata og vælg fold k som testdata.
   (b) Træn MLR-model på træningsdata.
   (c) Test MLR-model på fold k vha. Algoritme 2.
   (d) Bevar konfusionsmatrix og kassér modellen.
3. Opsummer konfusionsmatrix.

Output: konfusionsmatrix.

I Kapitel 3 introduceres Algoritme 1 for K-fold krydvalidering, som danner udgangspunktet for Algoritme 3. Algoritmen opdeler dataet i K unikke folder, som hver især anvendes som enten trænings- eller testdata. Herefter tilpasses MLR-modellen på træningsdata og testes på testdataet vha. Algoritme 2. I hver iteration af algoritmen bevares konfusionsmatricen. I Figur 5.10a illustreres klassifikation af testdata fra iteration 6 fra Algoritme 3, og i Figur 5.10b illustreres placeringen af fejlklassifikationer. Observationer, som tildeles no-call for hver fold, placerer sig forsat i samme område, som på Figur 5.8.



(a) Fold 6 med no-call bånd.

(b) Fejlklassifikationer i fold 6 med no-call bånd.

Figur 5.10: Illustration af klassifikation af testdata med MLR og no-call bånd i iteration 6 af 11-fold krydsvalidering. Genotype AA er grøn, BB er rød, AB er sort, no-call er grå og fejlklassificering er orange.

I Tabel 5.5 ses antal observationer og en opsummering af konfusionsmatricer for hver fold i Algoritme 3, hvor GNMS angiver gennemsnittet af antal observationer, fejlfrekvensen $CV_{(K)}$, antal af observationer tildelt no-call og den gennemsnitslige model præcision af alle folder.

K	**Antal i test-fold**	**Testfejl**	**No-call total**	**Præcision (%)**
1	379	11	24	96.9
2	407	7	21	98.2
3	387	4	17	98.9
4	412	11	23	97.2
5	436	8	31	98.0
6	396	7	21	98.1
7	388	11	20	97.0
8	406	6	20	98.5
9	392	11	19	97.1
10	392	10	24	97.3
11	405	10	19	97.4
GNMS	400	$8.72 \approx 9$	$21.73 \approx 22$	97.7

Tabel 5.5: Opsummering af konfusionsmatricer i 11-fold krydsvalidering. GNMS angiver gennemsnittet.

I 11-fold krydsvalidering klassificerer MLR-modellen i gennemsnit 97.7% af alle nye observationer korrekt sammenlignet med responsvariablen WGS. Dog ses det, at MLR-modellen i gennemsnit tildeler 22 observationer til no-call, tilsvarende 5.5% af testdataet, og forsat fejlklassificerer i gennemsnit 9 observationer efter no-call. Dette blev ligeledes observeret i Tabel 5.3 og Tabel 5.4, hvorfor det er forventelig, at det også ville ske for testdata i 11-fold krydsvalidering.

Sammenlignes resultatet fra Tabel 5.4 og Tabel 5.5 ses det, at der opnås en bedre procentvis præcision i første test (Tabel 5.4, 98.1%) end ved 11-fold krydsvalidering (Tabel 5.5, 97.7%), og den gennemsnitslige fejlfrekvens $CV_{(K)}$ forstørres for Tabel 5.5, da den stiger fra 7 til 9. Modsat ses det, at antallet af observationer, som tildeles no-call i Tabel 5.4 (36) falder i Tabel 5.5 (gennemsnitlig 22). Selvom den procentvise modelpræcision af 11-fold krydsvalidering er lavere end første test, er andelen af no-call mindre. Dermed er der flere observationer, hvor modellen er mere sikker på at klassificere genotyper end første test. Det øjebliks-billede som første test viser, bekræftes derudover af, at den procentvise andel som tildeles no-call er højere, 9%, end både for træningsdataet (Tabel 5.3), 5.9%, og 11-folds krydsvalidering (Tabel 5.5), 5.5%. Det er derfor mere realistisk at forvente, at den procentvise andel af no-call er omkring $5.5 - 6\%$ af dataet.

6 Hypotesetest af symmetri

Visualiseringen af datasættet i Figur 4.2 kan antyde, at datasættet er symmetrisk omkring identitetslinjen. Hvis det er tilfældet, vil der kunne opstilles en simplere model. Hvis en symmetrimodel klassificerer tilsvarende MLR-modellen, vil symmetrimodellen være at foretrække, da den kræver færre estimater og er dermed mindre beregningsmæssig tung. Det undersøges derfor, om genotyperne er symmetriske omkring identitetslinjen. Det antages, at signalerne følger bivariate normalfordelinger $N_2(\hat{\mu}_i, \hat{\Sigma}_i)$, hvor $i \in \{AA, BB, AB\}$, og

$$\begin{aligned} \hat{\mu}_{AA} &= \begin{bmatrix} 9.19 \\ 6.62 \end{bmatrix} \quad \text{og} \quad \hat{\Sigma}_{AA} = \begin{bmatrix} 0.34 & 0.06 \\ 0.06 & 0.53 \end{bmatrix}, \\ \hat{\mu}_{BB} &= \begin{bmatrix} 6.58 \\ 9.20 \end{bmatrix} \quad \text{og} \quad \hat{\Sigma}_{BB} = \begin{bmatrix} 0.52 & 0.07 \\ 0.07 & 0.28 \end{bmatrix}, \quad \text{og} \\ \hat{\mu}_{AB} &= \begin{bmatrix} 8.84 \\ 8.78 \end{bmatrix} \quad \text{og} \quad \hat{\Sigma}_{AB} = \begin{bmatrix} 0.21 & 0.16 \\ 0.16 & 0.46 \end{bmatrix}, \end{aligned} \tag{6.1}$$

og definer

$$\hat{\mu}'_{BB} = \begin{bmatrix} 9.20 \\ 6.58 \end{bmatrix} \quad \text{og} \quad \hat{\Sigma}'_{BB} = \begin{bmatrix} 0.28 & 0.07 \\ 0.07 & 0.52 \end{bmatrix}. \tag{6.2}$$

Til at undersøge, om genotype AB er symmetrisk omkring identitetslinjen, opstilles nulhypotesen $\mathcal{H}_0$:

$$\mathcal{H}_0 : \mu_{AB} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} \quad \text{og} \quad \Sigma_{AB} = \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix}, \tag{6.3}$$

Denne nulhypotese kan imidlertid være vanskelig at teste, hvorfor nulhypotesen opdeles i to tests. I første del anvendes en korrelationstest, som er introduceret i Afsnit 6.1, og i anden del anvendes en t-test, som er introduceret i Afsnit 6.2.

Til at undersøge om klyngerne for genotype AA og BB er symmetriske omkring identitetslinjen, skal det både gælde, at $\Sigma_{AA} = \Sigma'_{BB}$ og $\mu_{AA} = \mu'_{BB}$. Hermed kan følgende hypotese opstilles:

$$\begin{aligned} &\mathcal{H}_0 : \Sigma_{AA} = \Sigma'_{BB} \quad \text{og} \quad \mu_{AA} = \mu'_{BB}, \quad \text{og} \\ &\mathcal{H}_1 : \text{Mindst ét af kravene i } \mathcal{H}_0 \text{ er ikke opfyldt.} \end{aligned} \tag{6.4}$$

Denne hypotese opdeles i to delhypoteser. For at teste om $\Sigma_{AA} = \Sigma'_{BB}$ anvendes Box's M-test, som er introduceret i Afsnit 6.3, og for at teste om $\mu_{AA} = \mu'_{BB}$ anvendes Hotellings T^2-test, som er introduceret i Afsnit 6.4.

6.1 Korrelationstest

Dette afsnit er baseret på [11, kap. 3.5.7], [12] og [13].

I første del af at teste nulhypotensen $\mathcal{H}_0$ i (6.3) udarbejdes en test af korrelation med udgangspunkt i Fishers Z-test.

Definition 6.1. Pearsons korrelationskoefficient
Lad X og Y være to stokastiske variabler. Da er Pearsons korrelationskoefficient defineret ved

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

hvor σ_X og σ_Y angiver standardafvigelserne for hhv. X og Y.

At teste nulhypotesen $\mathcal{H}_0$ i (6.3) er imidlertid ikke lige til, hvorfor der anvendes et trick, hvor der konstrueres to normalfordelte stokastiske variabler

$$U = X_A - X_B = (u_1, \ldots, u_n)^T \quad \text{og} \quad V = X_A + X_B = (v_1, \ldots, v_n)^T.$$

Hvis U og V er uafhængige, vil tricket være ækvivalent med nulhypotesen $\mathcal{H}_0$ i (6.3) jf. Sætning 6.2.

Sætning 6.2.
Lad U og V være normalfordelte stokastiske variabler. Da gælder det, at

$$U, V \text{ uafhængige} \quad \Longleftrightarrow \quad \rho(U,V) = 0.$$

Beviset for denne sætning er udeladt fra denne rapport.

Da U og V er konstrueret fra det udleverede data, benyttes Pearsons stikprøve korrelationskoefficient til at teste, om der eksisterer korrelation mellem U og V.

Definition 6.3. Pearsons stikprøve korrelationskoefficient
Lad (u_i, v_i), $i = 1, \ldots, n$, være observationspar fra en stikprøve. Da er Pearsons stikprøve korrelationskoefficient defineret ved

$$\hat{\rho}(U,V) = \frac{\sum_{i=1}^{n} u_i v_i - n\bar{U}\bar{V}}{(n-1)\sqrt{S_U S_V}} = \frac{\sum_{i=1}^{n} u_i v_i - n\bar{U}\bar{V}}{\sqrt{\left(\sum_{i=1}^{n} u_i^2 - n\bar{U}^2\right)\left(\sum_{i=1}^{n} v_i^2 - n\bar{V}^2\right)}}, \tag{6.5}$$

hvor $\hat{\rho} \in [-1, 1]$ og

$$S_U = \frac{1}{n-1}Q_U = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{U})^2 \quad \text{og} \quad S_V = \frac{1}{n-1}Q_V = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{V})^2.$$

Hvis $\hat{\rho}$ antager en værdi tæt på 0, er der ikke korrelation mellem U og V, og U og V er dermed uafhængige. Antager $\hat{\rho}$ derimod en værdi tæt på 1 eller -1, er U og V afhængige. Den numeriske størrelse af $\hat{\rho}$ kan ikke fortolkes som evidens for et årsag- og virkningsforhold mellem de to variabler, men giver derimod evidens for, om variablerne U og V er uafhængige eller afhængige.

Til korrelationstesten opstilles en delhypotese af (6.3), sådan at

$$\begin{aligned} &\mathcal{H}_0 : \rho = 0, \quad \text{og} \\ &\mathcal{H}_1 : \rho \neq 0. \end{aligned} \tag{6.6}$$

Nulhypotesen $\mathcal{H}_0$ i (6.6) testes vha. Fishers transformation. Denne er defineret som følger.

Definition 6.4. Fishers transformation
Lad U og V være normalfordelte stokastiske variabler, og lad $\hat{\rho}$ betegne Pearsons stikprøve korrelationskoefficient. Så er Fishers transformation defineret ved

$$\mathcal{F}(\hat{\rho}(U,V)) = \frac{1}{2}\log\left(\frac{1+\hat{\rho}(U,V)}{1-\hat{\rho}(U,V)}\right) = \tanh^{-1}(\hat{\rho}(U,V)).$$

Ifølge [11, kap. 3.5.7] følger $\mathcal{F}$ approksimativt en univariat normalfordeling $N\left(\mathcal{F}(\rho), \frac{1}{n-3}\right)$. Hertil anvendes en to-halet Z-test for at undersøge, om ρ er signifikant forskellig fra 0 ved at definere en teststørrelse vha. en *score*-funktion, sådan at

$$\mathcal{Z} = \frac{\mathcal{F}(\hat{\rho}(U,V)) - \mathcal{F}(0)}{\frac{1}{\sqrt{n-3}}} = (\mathcal{F}(\hat{\rho}(U,V)) - \mathcal{F}(0))\sqrt{n-3}.$$

Denne teststørrelse vil under $\mathcal{H}_0$ i (6.6) følge en standard normalfordeling. Hvis

$$|\mathcal{Z}| > |Z_{krit}|$$

forkastes hypotesen $\mathcal{H}_0$ i (6.6), og dermed er U og V ikke uafhængige. Værdierne for $|Z_{krit}|$ findes i Tabel B.4 for et signifikansniveau på 10%, 5% og 1%.

6.2 t-test

Dette afsnit er baseret på [14].

Generelt anvendes en t-test til at teste en stikprøve mod en population, altså at undersøge inferens for korrelation vha. en stikprøve. På den måde kan der udledes egenskaber for en population ud fra en parameter, og dette gøres ved at betragte middelværdien. I denne rapport ønskes det at undersøge, om genotype AB er symmetrisk omkring identitetslinjen, hvorfor det undersøges, om middelværdien for AB kan opskrives som i (6.3). Her følger en teststørrelse t en t-fordeling med $n-1$ frihedsgrader, som er defineret i Definition A.8. Denne t teststørrelse er defineret som følger.

Definition 6.5. Teststørrelsen t for one-sample t-test
Lad n betegne observationerne i datasættet $Z = (z_1, \ldots, z_n)^T$, $\bar{Z}$ betegne stikprøvemiddelværdien af Z,

$$S_i = \frac{1}{n-1} Q_i = \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{Z})^2$$

være estimatet for stikprøvekovariansen, og lad μ være en selvvagt middelværdi. Så er teststørrelsen t for en one-sample t-test defineret ved

$$t = (\bar{Z} - \mu) \sqrt{\frac{n}{S_i}}.$$

Bemærkning. *Stikprøvekovariansen Q_i er et skalar af $\hat{\Sigma}_i$, hvor $Q_i = n_i \hat{\Sigma}_i$, og n_i angiver antallet af observationerne i gruppen i. Dermed gælder det, at*

$$S_{AB} = \frac{1}{n_{AB} - 1} Q_{AB} \quad og \quad Q_{AB} = \frac{1}{n_{AB}} \hat{\Sigma}_{AB},$$

hvor n_{AB} angiver antallet af observationer i AB. Her er $\bar{Z} = (\bar{X}_A, \bar{X}_B)^T$, og den selvvalgte middelværdi vælges til 0.

Opstillingen af den alternative hypotese $\mathcal{H}_1$ og det endelige udtryk for teststørrelsen t afhænger af, hvilken t-test, der foretages. Der findes tre forskellige typer t-test: *one-sample*, *two-sample* og *paired*. I denne rapport fokuseres der på en one-sample to-halet t-test, da det ønskes at undersøge middelværdien af én stikprøve overfor en selvvalgt middelværdi. Dermed kan følgende delhypotese af (6.3) opstilles

$$\begin{aligned} &\mathcal{H}_0 : \bar{Z} = \mu, \quad \text{og} \\ &\mathcal{H}_1 : \bar{Z} \neq \mu. \end{aligned}$$

Hvis $|t| > |T_{krit}|$ forkastes nulhypotesen $\mathcal{H}_0$ i (6.3). Værdier for T_{krit} findes i Tabel B.5 for et signifikansniveau på 5%.

6.3 Box's M-test

Dette afsnit er baseret på [11, kap. 9], [15] og [16].

For at undersøge om, genotype AA og BB er symmetriske omkring identitetslinjen, foretages en analyse af klyngernes kovariansmatricer. Dette kan opstilles som en delhypotese af (6.4)

$$\begin{aligned} &\mathcal{H}_0 : \Sigma_{AA} = \Sigma'_{BB}, \quad \text{og} \\ &\mathcal{H}_1 : \Sigma_{AA} \neq \Sigma'_{BB}. \end{aligned} \tag{6.7}$$

Såfremt symmetri mellem kovariansmatricerne kan påvises, vil antagelsen i nulhypotesen $\mathcal{H}_0$ i (6.7) ikke forkastes. Mere generelt kan hypotesen for Box's M-test defineres som følger.

Definition 6.6. Hypotese for Box's M-test
Lad $Z = (Z_1, \ldots, Z_C)^T$ være uafhængige og identisk fordelte stokastiske variabler, hvor $Z_i \sim N_p(\mu_i, \Sigma_i)$, $i = 1, \ldots C$. Hypotesetesten defineres som

$$\begin{aligned} &\mathcal{H}_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_C, \quad \text{og} \\ &\mathcal{H}_1 : \text{mindst ét par, hvor } \Sigma_i \neq \Sigma_j,\ i, j = 1, 2, \ldots, C. \end{aligned} \tag{6.8}$$

Box's M-test er en metode til at teste symmetri mellem kovariansmatricer blandt C grupper, og er baseret på en *likelihood ratio* teststørrelse under den multivariate normalfordeling (se Appendiks A.4.2).

Definition 6.7. Likelihood ratio teststørrelse for Box's M-test
Lad $Z = (Z_1, \ldots, Z_C)^T$ være uafhængige og identisk fordelte stokastiske variabler, hvor $Z_i \sim N_p(\mu_i, \Sigma_i)$, $i = 1, \ldots C$. Så er likelihood ratio teststørrelsen for nulhypotesen i (6.7) bestemt ved

$$\lambda = \frac{\prod_{i=1}^{C} |Q_i|^{n_i/2} n^{pn/2}}{\left| \sum_{i=1}^{C} Q_i \right|^{n/2} \prod_{i=1}^{C} n_i^{pn_i/2}}. \tag{6.9}$$

Antag, at der er givet n_i elementer fra den $i$te gruppe sådan at $Z_i = ((Z_i)_1, \ldots, (Z_i)_{n_i})$ for $i = 1, \ldots, C$ så er Q_i for den i'te gruppe defineret ved

$$n_i \hat{\Sigma}_i = Q_i = \sum_{j=1}^{n_i} ((Z_i)_j - \bar{Z}_i)((Z_i)_j - \bar{Z}_i)^T, \tag{6.10}$$

hvor $\hat{\Sigma}_i$ er stikprøvekovariansen, Q_i er den observationsskalerede stikprøvekovarians, $|Q_i|$ angiver determinanten af Q_i, og $n = \sum_i n_i$ er antallet af observationer.

Inden Box's M-teststørrelse defineres, introduceres først en *pooled* kovariansmatrix. Denne anvendes til at udregne Box's M-teststørrelse, da en pooled kovariansmatrix er et estimat på en fælles kovariansmatrix mellem grupperne. Det betyder, at under nulhypotesen $\mathcal{H}_0$ vil den pooled kovariansmatrix være den *sande* kovariansmatrix, da det antages $\Sigma_{AA} = \Sigma'_{BB}$.

Definition 6.8. Pooled kovariansmatrix
Lad $S_i = \frac{1}{n_i - 1} Q_i$ angive stikprøvekovarianser for $i = 1, 2, \ldots, C$, og lad $n = \sum_i n_i$. Så er pooled kovariansmatrix defineret ved

$$S = \frac{1}{n - C} \sum_{i=1}^{C} (n_i - 1) S_i, \tag{6.11}$$

Når nulhypotesen er sand og n er stor, så følger $-2\log(\lambda)$ approksimativt en χ^2-fordeling med frihedsgrad $\nu_1 = \frac{1}{2} p(p+1)(C-1)$. Dog kan en bedre χ^2-approksimation opnås ved en modifikation af λ. Her bliver n_i erstattet af antallet af frihedsgrader $df_i = n_i - 1$ associeret med Q_i, hvilket leder frem til Box's M-teststørrelse.

Definition 6.9. Box's M-teststørrelse
Lad antallet af frihedsgrader være defineret ved $df = \sum_i df_i = n - C$, og lad $S_i = Q_i / df_i$ være stikprøvekovarianser og $S = \sum_i Q_i / df$ er en pooled kovariansmatrix. Funktionen $M : \mathbb{R} \to [0, 1]$ definerer Box's M-teststørrelse ved

$$M = \frac{\prod_{i=1}^{C} |S_i|^{df_i/2}}{|S|^{df/2}} = \frac{\prod_{i=1}^{C} |Q_i|^{df_i/2} df^{df/2}}{\left| \sum_{i=1}^{C} Q_i \right|^{df/2} \prod_{i=1}^{C} df_i^{df_i/2}}. \tag{6.12}$$

Fordelingen af Box's M-teststørrelse kan approksimeres på to forskellige måder. I første metode kan Box's M-teststørrelse approksimeres vha. af en χ^2-fordeling, sådan at M opfylder, at $-2\log(M)(1 - \gamma_1) \sim \chi^2(\nu_1)$, hvor

$$\gamma_1 = \frac{2p^2 + 3p - 1}{6(p+1)(C-1)} \left(\sum_{i=1}^{C} \frac{1}{n_i - 1} - \frac{1}{n - C} \right). \tag{6.13}$$

Nulhypotesen $\mathcal{H}_0$ i (6.7) forkastes, hvis $-2\log(M)(1 - \gamma_1) > \chi^2_{krit}$. Værdier for χ^2_{krit} findes i Tabel B.1 for et signifikansniveau på 5%.

χ^2-approksimationen er god, når $n_i > 20, p \leq 5$ og $C \leq 5$ [17]. Hvis disse krav ikke er opfyldt, kan Box's M-teststørrelse approksimeres vha. en F-fordeling. Da datasættet i denne rapport opfylder kravene for en god χ^2-approksimation vil den anden metode ikke blive uddybet yderligere, men kan findes i [11, s. 448-451].

6.4 Hotellings T^2-test

Dette afsnit er baseret på [11, kap. 9] og [18, kap. 3.6].

I anden del af undersøgelsen af symmetri imellem klyngerne for genotype AA og BB, undersøges det, om der er klyngernes middelværdier er symmetriske vha. Hotellings T^2-test, som er en generalisering af t-testen. Givet to uafhængige stokastiske variabler Z og W, hvor $Z \sim N_1(\mu, \sigma^2)$, og $W \sim \sigma^2\chi^2(n)$, så gælder det, at teststørrelsen

$$T = \frac{(Z-\mu)/\sigma}{\sqrt{(W/n\sigma^2)}} = \frac{Z-\mu}{\sqrt{W/n}} \sim t(n).$$

Herfra ses det, at

$$T^2 = \frac{n(Z-\mu)^2}{W} = n(Z-\mu)W^{-1}(Z-\mu) \sim F(1,n). \tag{6.14}$$

Generaliseres dette til flere dimensioner fås den såkaldte Hotellings T^2-teststørrelse.

Definition 6.10. Hotellings T^2-teststørrelse
Lad $Z = (z_1, \ldots, z_p)^T \sim N_p(\mu, \Sigma)$, $W = (w_1, \ldots, w_p)^T \sim W_p(n, \Sigma)$, og lad Z og W være uafhængige. Så er Hotellings T^2-teststørelse defineret som

$$T^2 = n(Z-\mu)^T W^{-1}(Z-\mu). \tag{6.15}$$

Fra (6.14) ses det, at Hotellings T^2-teststørrelse i et én-dimensionalt tilfælde følger en F-fordeling. Dette kan udvides, sådan at teststørrelsen T^2 i Definition 6.10 også følger en F-fordeling for en multivariat normalfordeling. Dette er vist i [11, Kap. 2.4], altså

$$\frac{n-p+1}{p}\,\frac{T^2}{n} \sim F(p, n-p+1). \tag{6.16}$$

For at lette notationen i (6.16) indføres Hotellings $\mathcal{T}^2$-fordeling, sådan at teststørrelsen T^2 siges at følge en $\mathcal{T}^2$-fordeling, noteret $\mathcal{T}^2(p,n)$, hvis (6.16) gælder.

Sætning 6.11. $\mathcal{T}^2$-fordeling for grupper af forskellig størrelse
Lad X_1 og X_2 være designmatricer, hvor alle n_i rækker i X_i for $i = 1, 2$ er uafhængige og identiske fordelte som $N_p(\mu_i, \Sigma_i)$, sådan at alle n rækker i $X = [X_1, X_2]^T$ er uafhængige. Så når $\mu_1 = \mu_2$ og $\Sigma_1 = \Sigma_2$, så følger

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{X}_1 - \bar{X}_2)^T S^{-1}(\bar{X}_1 - \bar{X}_2) \sim \mathcal{T}^2(p, n_1+n_2-2),$$

hvor S angiver pooled kovariansmatrix, og $\bar{X}_i$ angiver stikprøvemiddelværdien.

Bevis:
Lad X_1 og X_2 være designmatricer, hvor alle n_i rækker i X_i for $i = 1,2$ er uafhængige og identiske fordelte som $N_p(\mu_i, \Sigma_i)$, sådan at alle n rækker i $X = [X_1, X_2]^T$ er uafhængige. Definer $D = \bar{X}_1 - \bar{X}_2$. Jf. Sætning 3.1 i [11, s. 63] så $\bar{X}_i \sim N_p\left(\mu_i, \frac{1}{n_i}\Sigma_i\right)$, $i = 1,2$. Dermed

$$D \sim N_p\left(\mu_1 - \mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right)$$

Når $\mu_1 = \mu_2$ og $\Sigma_1 = \Sigma_2 = \Sigma$, så

$$D \sim N_p\left(0, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right) = N_p\left(0, \left(\frac{n_2}{n_1 n_2} + \frac{n_1}{n_1 n_2}\right)\Sigma\right) = N_p\left(0, \frac{n_1 + n_2}{n_1 n_2}\Sigma\right).$$

Definer $a = \frac{n_1+n_2}{n_1 n_2}$. Jf. Definition 6.10 skal pooled kovariansmatricen $S \sim W_p(n, \Sigma)$ for at teststørrelsen T^2 kan skrives på formen af (6.15). Betragt nu $Q_i = n_i\hat{\Sigma}_i$, hvor $\hat{\Sigma}_i$ angiver stikprøvekovariansen for den i'te gruppe, og $S = (Q_1 + Q_2)/(n_1 + n_2 - 2)$ angiver pooled kovariansmatrix. Så følger $Q_i \sim W_p(n_i - 1, \Sigma_i)$ jf. Sætning A.11. Dermed, når $\Sigma_1 = \Sigma_2 = \Sigma$, så

$$Q = (n_1 + n_2 - 2)S = Q_1 + Q_2 \sim W_p(n_1 + n_2 - 2, \Sigma).$$

Fra Sætning A.11 følger det ligeledes, at $aQ \sim W_p(n_1 + n_2 - 2, a\Sigma)$, Q er uafhængig af D, da $\bar{X}_i$ er uafhængig af $\hat{\Sigma}_i$ og dermed også uafhængig af Q_i for $i = 1,2$, og X_1 og X_2 er uafhængige. Derfor gælder det, at

$$T^2 = (n_1 + n_2 - 2)D^T(aQ)^{-1}D = D^T(aS)^{-1}D = \frac{n_1 n_2}{n_1 + n_2}(\bar{X}_1 - \bar{X}_2)^T S^{-1}(\bar{X}_1 - \bar{X}_2)$$

for $S = \frac{1}{n_1+n_2-2}Q$. Dermed følger $T^2 \sim \mathcal{T}^2(p, n_1 + n_2 - 2)$ for to grupper af forskellige størrelser. ■

Dermed kan følgende delhypotese af (6.4) opstilles

$$\begin{aligned} &\mathcal{H}_0 : \mu_{AA} = \mu'_{BB}, \quad \text{og} \\ &\mathcal{H}_1 : \mu_{AA} \neq \mu'_{BB}. \end{aligned} \tag{6.17}$$

Under nulhypotesen $\mathcal{H}_0$ i (6.17) vil

$$T^2 = \frac{n_{AA} n_{BB}}{n_{AA} + n_{BB}}(\hat{\mu}_{AA} - \hat{\mu}'_{BB})^T S^{-1}(\hat{\mu}_{AA} - \hat{\mu}'_{BB}) \sim \mathcal{T}^2(p, n_{AA} + n_{BB} - 2).$$

Dette kan også skrives vha. en F-fordeling, hvor

$$F_0 = \frac{n_{AA} + n_{BB} - p - 1}{(n_{AA} + n_{BB} - 2)p}T^2 \sim F(p, n_{AA} + n_{BB} - 1 - p).$$

Nulhypotesen vil blive forkastet, hvis $F_0 > F_{krit}(p, n_{AA} + n_{BB} - 1 - p)$ med signifikansniveau α. Værdier for F_{krit} findes i Tabel B.2 for et signifikansniveau på 1% og Tabel B.3 for et signifikansniveau på 5%.

7 Test af symmetri

I dette kapitel anvendes testene beskrevet i Kapitel 6 til at undersøge om genotyperne er symmetriske omkring identitetslinjen. Hvis der er symmetri, afsluttes kapitlet med en konstruktion af en symmetrimodel med udgangspunkt i MLR, og hvis der ikke er symmetri, gives en metode til konstruktion af en symmetrimodel.

7.1 Symmetri i homozygote genotyper omkring identitetslinjen

Indledningvist i Kapitel 6 antages det, at de to homozygote genotyperne AA og BB følger en bivariat normalfordeling $N_2(\hat{\mu}_i, \hat{\Sigma}_i)$, hvor $i \in \{AA, BB\}$. Hvis der er symmetri mellem de to homozygote genotyper AA og BB vil vi se, at deres kovarianser og middelværdier er symmetriske. Her testes hypotesen opstillet i (6.4),

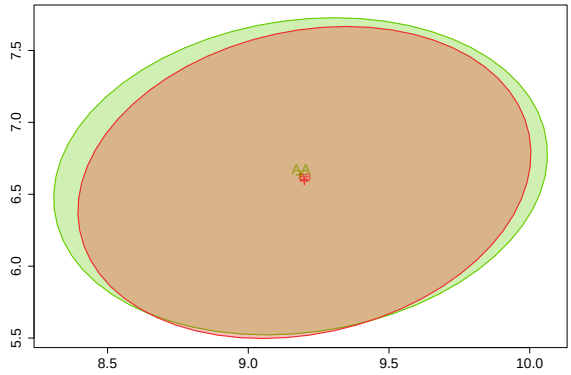
$$\begin{aligned}\mathcal{H}_0 &: \Sigma_{AA} = \Sigma'_{BB} \quad \text{og} \quad \mu_{AA} = \mu'_{BB}, \quad \text{og} \\ \mathcal{H}_1 &: \text{Mindst ét af kravene i } \mathcal{H}_0 \text{ er ikke opfyldt.}\end{aligned}$$

Først undersøges det, om deres kovariansmatricer er symmetriske vha. Box's M-test, og derefter undersøges, hvorvidt deres middelværdier er symmetriske vha. Hotellings T^2-test.

Til at teste, om der er symmetri mellem kovariansmatricerne for genotype AA og BB, testes delhypotesen (6.7) med Box's M-test,

$$\begin{aligned}\mathcal{H}_0 &: \Sigma_{AA} = \Sigma'_{BB}, \quad \text{og} \\ \mathcal{H}_1 &: \Sigma_{AA} \neq \Sigma'_{BB}.\end{aligned}$$

Indledningsvist visualiseres kovariansellipserne for Σ_{AA} og Σ'_{BB} i Figur 7.1 med centrum i henholdsvis μ_{AA} og μ'_{BB}. Her opnås en indikation af, at kovariansellipsen for Σ_{AA} (grøn) er større end kovariansellipsen for Σ'_{BB} (rød), hvilket kan indikere, at $\mathcal{H}_0$ i (6.7) kan forkastes.



Figur 7.1: Kovariansellipser for genotyperne AA (grøn) og BB (rød).

Ved anvendelse af `boxM()` funktionen fra R-pakken `heplots` forkastes nulhypotesen, da funktionen beregner $M = 274.01$, sådan at

$$-2\log(M)(1-\gamma_1) = -11.22,$$

hvor

$$\gamma_1 = \frac{1}{2}\left(\left(\frac{1}{1628} - \frac{1}{3717}\right) + \left(\frac{1}{2089} - \frac{1}{3717}\right)\right) = 0.0004$$

for $p = 2$ og $C = 2$ i (6.13). Dette kan ikke lade sig gøre, da χ^2 er en mængde, som består af kvadrede værdier, hvorfor χ^2 ikke kan antage negative værdier. Derfor er værdien for M ikke retvisende, hvorfor Box's M-test er implementeret manuelt i R. Her viser det sig, at i beregningen af teststørrelsen M evalueres tal, som R numerisk anser som 0, hvorfor det ikke er muligt manuelt at beregne teststørrelsen M i R. Anvendes derimod CAS-værktøjet Maple (version 2021.2) imødekommes det numeriske problem. Beregningerne hertil findes i Appendiks C, hvorfra det observeres, at

$$M = \frac{|S_{AA}|^{1628/2}|S_{BB}|^{2089/2}}{|S|^{3717/2}} = 0.00013$$

for

$$S = \begin{bmatrix} 0.31 & 0.07 \\ 0.07 & 0.52 \end{bmatrix}, \quad S_{AA} = \begin{bmatrix} 0.34 & 0.06 \\ 0.06 & 0.53 \end{bmatrix} \quad \text{og} \quad S'_{BB} = \begin{bmatrix} 0.28 & 0.07 \\ 0.07 & 0.52 \end{bmatrix}.$$

Her ses det, at værdierne for S_{AA} og S'_{BB} er tilsvarende med kovariansmatricerne i (6.1) på grund af afrunding af decimaler. I Tabel B.1 ses det, at $\chi^2_{krit} = 7.82$ ved et signifikansniveau på 5% med 3 frihedsgrader. Hermed observeres det, at

$$-2\log(M)(1-\gamma_1) = 17.88 \implies -2\log(M)(1-\gamma_1) > \chi^2_{krit}.$$

Dermed ses en mere retvisende evaluering af M, men nulhypotesen $\mathcal{H}_0$ i (6.7) forkastes. Altså, er der ikke symmetri imellem genotype AA og BBs kovariansmatricer.

Dermed ses det, at første krav for $\mathcal{H}_0$ i (6.4) ikke er opfyldt, og vi kan allerede forkaste hypotesen om, at genotyperne AA og BB ikke er symmetriske omkring identitetslinjen. På trods af dette, undersøges det forsat, om der eksisterer symmetri i middelværdierne for genotype AA og BB omkring identitetslinjen med henblik at konstruere en symmetrimodel for at undersøge om forskellen mellem $\hat{\Sigma}_{AA}$ og $\hat{\Sigma}'_{BB}$ er probabilistisk signifikant.

I anden del af testen, om genotype AA og BB er symmetriske omkring identitetslinjen, tester vi, om klyngernes middelværdier er symmetriske. Her testes delhypotesen (6.4), altså

$$\begin{aligned} \mathcal{H}_0 &: \mu_{AA} = \mu'_{BB}, \quad \text{og} \\ \mathcal{H}_1 &: \mu_{AA} \neq \mu'_{BB}. \end{aligned}$$

Hertil observeres det, at for F-fordelingen er

$$df_1 = p = 2 \quad \text{og} \quad df_2 = 1629 + 2090 - 1 - 2 = 3716,$$

og

$$T^2 = \frac{1629 \cdot 2090}{1629 + 2090} \left(\hat{\mu}_{AA} - \hat{\mu}'_{BB}\right)^T S^{-1} \left(\hat{\mu}_{AA} - \hat{\mu}'_{BB}\right) \approx 4.31,$$

sådan at

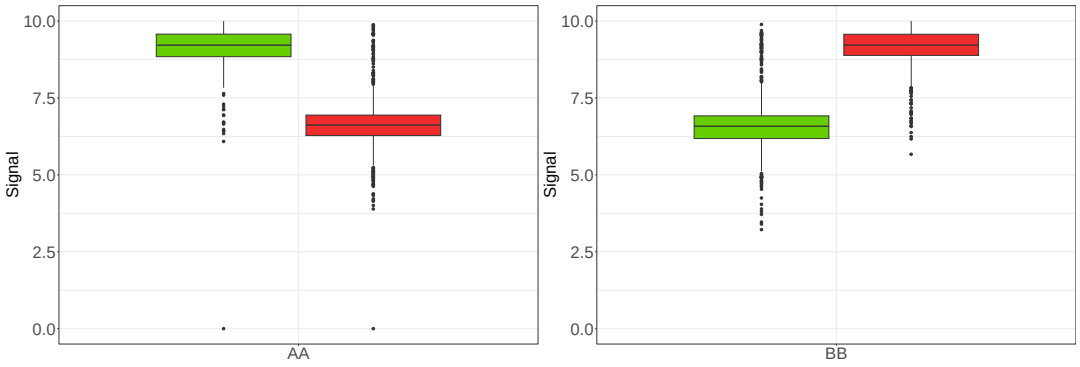
$$F(df_1, df_2) = \frac{1629 + 2090 - 2 - 1}{(1629 + 2090 - 2)2} 4.31 \approx 2.16.$$

Det observeres fra Tabel B.3, at $F_{krit} = 3.69$ for et signifikansniveau på 5%, hvor

$$F(df_1, df_2) < F_{krit}.$$

Det ses også, at dette er opfyldt for et signifikansniveau på 1%, hvor $F_{krit} = 4.61$ fra Tabel B.2. Dette betyder, nulhypotesen $\mathcal{H}_0$ i (6.17) ikke forkastes.

I Figur 7.2 visualiseres et boxplot af signalintensiteterne i $\hat{\mu}_{AA}$ og $\hat{\mu}'_{BB}$. Fra (6.1) ses det, at vi skal undersøge om middelværdien af signalet $\log(A+1)$ fra genotype AA er tilnærmelsesvis ens med middelværdien af signalet $\log(B+1)$ fra genotype BB. Tilsvarende med middelværdien af signalet $\log(B+1)$ fra genotype AA og middelværdien af signalet $\log(A+1)$ fra genotype BB. Her visualiseres det, at alle kvartiler tilnærmelsesvis ens, og det giver en god indikation af, at der er symmetri iblandt middelværdierne.



(a) Boksplot for $\hat{\mu}_{AA}$. **(b)** Boksplot for $\hat{\mu}'_{BB}$.

Figur 7.2: Visualisering af boksplot for $\hat{\mu}_{AA}$ og $\hat{\mu}'_{BB}$. Signal $\log(A+1)$ er grøn og signal $\log(B+1)$ er rød.

7.2 Symmetri i heterozygot genotype omkring identitetslinjen

Indledningvist i Kapitel 6 ses det også, at den heterozygote genotype AB følger en bivariat normalfordeling $N_2(\hat{\mu}_{AB}, \hat{\Sigma}_{AB})$. Til at teste, om genotypen AB er symmetriske omkring identitetslinjen, anvendes nulhypotensen opstillet i (6.3),

$$\mathcal{H}_0 : \mu_{AB} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} \quad \text{og} \quad \Sigma_{AB} = \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix},$$

At teste denne symmetri kan være vanskeligt, hvorfor vi til testen anvender tricket fremlagt i Afsnit 6.1. For dette trick gælder Sætning 7.1.

Sætning 7.1.
Lad Z_1 og Z_2 være stokastistiske variabler, og lad $Z_i \sim N(\mu, \sigma^2)$, og definer $U = Z_1 - Z_2$ og $V = Z_1 + Z_2$. Lad $\mathcal{H}_0$ være givet ved

$$\mu = \begin{bmatrix} \mathbb{E}(Z_1) \\ \mathbb{E}(Z_2) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} \quad \text{og} \quad \Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix}, \tag{7.1}$$

hvor det antages $\mathbb{E}(Z_1) = \mathbb{E}(Z_2)$. Nulhypotesen $\mathcal{H}_0$ er sand hvis og kun hvis

a) $U \sim N(0, 2\sigma^2(1-\rho)) = N(0, \sigma_U^2)$,

b) $V \sim N(2\mu, 2\sigma^2(1+\rho)) = N(\mu_V, \sigma_V^2)$, og

c) U og V er uafhængige.

Bevis:
Antag, at nulhypotesen $\mathcal{H}_0$ fra (7.1) er sand, så

$$\mathbb{E}(Z_1) = \mathbb{E}(Z_2) = \mu, \quad \text{Var}(Z_1) = \text{Var}(Z_2) = \sigma^2 \quad \text{og} \quad \text{Cov}(Z_1, Z_2) = \sigma^2\rho.$$

Lad $U = Z_1 - Z_2$ og $V = Z_1 + Z_2$ være normalfordelte stokastiske variabler. Så er middelværdi og varians for U givet ved

$$\mathbb{E}(U) = \mathbb{E}(Z_1 - Z_2) = \mathbb{E}(Z_1) - \mathbb{E}(Z_2) = \mu - \mu = 0$$

og

$$\begin{aligned} \text{Var}(U) &= \mathbb{E}((Z_1 - Z_2)^2) - \mathbb{E}(Z_1 - Z_2)^2 \\ &= \mathbb{E}(Z_1^2 + Z_2^2 - 2Z_1Z_2) - (\mathbb{E}(Z_1) - \mathbb{E}(Z_2))^2 \\ &= \mathbb{E}(Z_1^2) - \mathbb{E}(Z_1)^2 + \mathbb{E}(Z_2^2) - \mathbb{E}(Z_2)^2 - 2(\mathbb{E}(Z_1) - \mathbb{E}(Z_1)\mathbb{E}(Z_2)) \\ &= \text{Var}(Z_1) + \text{Var}(Z_2) - 2\text{Cov}(Z_1, Z_2) \\ &= \sigma^2 + \sigma^2 - 2\sigma^2\rho = 2\sigma^2(1-\rho). \end{aligned}$$

Tilsvarende er middelværdi og varians for V givet ved

$$\mathbb{E}(V) = \mathbb{E}(Z_1 + Z_2) = 2\mu \quad \text{og} \quad \text{Var}(V) = \text{Var}(Z_1 + Z_2) = 2\sigma^2(1+\rho).$$

Dermed følger $U \sim N_2(0, 2\sigma^2(1-\rho))$ og $V \sim N_2(2\mu, 2\sigma^2(1+\rho))$. For at vise, at U og V er uafhængige, skal $\text{Cov}(U, V) = 0$ jf. Sætning A.15 (2),

$$\begin{aligned} \text{Cov}(U, V) &= \text{Cov}(Z_1 - Z_2, Z_1 + Z_2) \\ &= \text{Cov}(Z_1, Z_1) + \text{Cov}(Z_1, Z_2) - \text{Cov}(Z_1, Z_2) - \text{Cov}(Z_2, Z_2) \\ &= \text{Cov}(Z_1, Z_1) - \text{Cov}(Z_2, Z_2) = \text{Var}(Z_1) - \text{Var}(Z_2) = 0, \end{aligned}$$

da $\text{Var}(Z_1) = \text{Var}(Z_2)$, og dermed er U og V uafhængige.

Det ønskes nu at vise, at a), b) og c) er ækvivalente med nulhypotesen $\mathcal{H}_0$ i (7.1). Antag derfor, at $U \sim N_2(0, 2\sigma^2(1-\rho))$, $V \sim N_2(2\mu, 2\sigma^2(1+\rho))$, og U og V er uafhængige, så er

$$\mathbb{E}(U) = \mathbb{E}(Z_1) - \mathbb{E}(Z_2) = 0 \quad \text{og} \quad \mathbb{E}(V) = \mathbb{E}(Z_1) + \mathbb{E}(Z_2) = 2\mu.$$

Fra $\mathbb{E}(U)$ ses det, at $\mathbb{E}(Z_1) = \mathbb{E}(Z_2)$, hvorfor

$$\mathbb{E}(V) = \mathbb{E}(Z_1) + \mathbb{E}(Z_1) = 2\mathbb{E}(Z_1) = 2\mu,$$

hvor det følger, at $\mathbb{E}(Z_1) = \mathbb{E}(Z_2) = \mu$. Dermed opfylder middelværdien under nulhypotesen $\mathcal{H}_0$, at

$$\mu = \begin{bmatrix} \mathbb{E}(Z_1) \\ \mathbb{E}(Z_2) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix},$$

hvor $\mathbb{E}(Z_1) = \mathbb{E}(Z_2)$. For variansen gælder det, at

$$\begin{aligned} \text{Var}(U) + \text{Var}(V) &= \text{Var}(Z_1 - Z_2) + \text{Var}(Z_1 + Z_2) \\ &= \text{Var}(Z_1) + \text{Var}(Z_2) - 2\text{Cov}(Z_1, Z_2) + \text{Var}(Z_1) + \text{Var}(Z_2) + 2\text{Cov}(Z_1, Z_2) \\ &= 2(\text{Var}(Z_1) + \text{Var}(Z_2)) = 2\sigma^2(1-\rho) + 2\sigma^2(1+\rho), \end{aligned}$$

hvor det følger, at

$$\text{Var}(Z_1) + \text{Var}(Z_2) = \sigma^2(1-\rho) + \sigma^2(1+\rho) = \sigma^2 - \sigma^2\rho + \sigma^2 + \sigma^2\rho = 2\sigma^2.$$

Da U og V er antaget uafhængige, gælder det, at

$$\begin{aligned} \text{Cov}(U, V) &= \text{Cov}(Z_1 - Z_2, Z_1 + Z_2) \\ &= \text{Cov}(Z_1, Z_1) + \text{Cov}(Z_1, Z_2) - \text{Cov}(Z_1, Z_2) - \text{Cov}(Z_2, Z_2) \\ &= \text{Var}(Z_1) - \text{Var}(Z_2) = 0. \end{aligned}$$

Dermed er

$$\text{Var}(Z_1) + \text{Var}(Z_2) = 2\sigma^2 \quad \text{og} \quad \text{Var}(Z_1) - \text{Var}(Z_2) = 0.$$

Da $\text{Var}(Z_1) = \text{Var}(Z_2)$, ses det, at

$$2\sigma^2 = \text{Var}(Z_1) + \text{Var}(Z_1) = 2\text{Var}(Z_1),$$

hvor det følger, at $\text{Var}(Z_1) = \text{Var}(Z_2) = \sigma$. Hermed opfylder variansen under nulhypotesen $\mathcal{H}_0$, at:

$$\hat{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix}.$$

■

Definer nu $Z_1 = X_A$ og $Z_2 = X_B$. Så kan nulhypotesen $\mathcal{H}_0$ i (6.3) testes ved:

$$\begin{aligned} &1: \text{ Korrelationstest for } U \text{ uafhængig af } V. \\ &2: \text{ Givet uafhængighed: one-sample } t\text{-test for } \mu_U = 0. \end{aligned} \tag{7.2}$$

I undersøgelse af første del i (7.2) observeres det, at estimatet for Pearsons korrelationskoefficienten $\hat{\rho} = -0.43$. Der udregnes derfor en $\mathcal{Z}$-teststørrelse

$$\mathcal{Z} = (\mathcal{F}(-0.43) - \mathcal{F}(0))\sqrt{681 - 3} = -11.94 \tag{7.3}$$

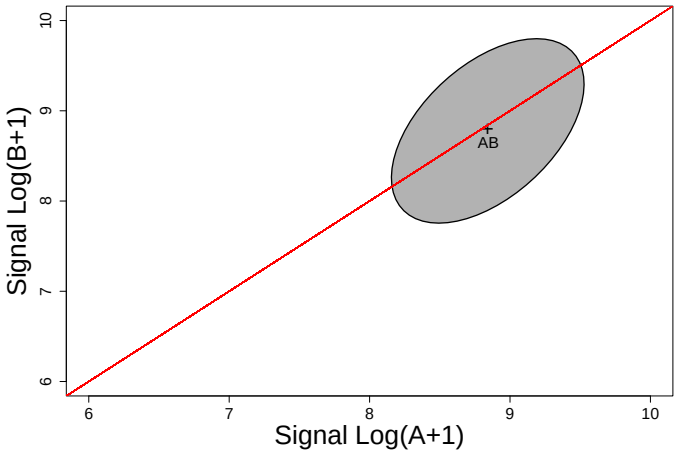
I Tabel B.4 findes værdierne for Z_{krit} og med signifikansniveau på 5% er $Z_{krit} = \pm 1.96$. Da $|\mathcal{Z}| > |Z_{krit}|$, så er der korrelation mellem U og V, og dermed er U og V afhængige. Dermed kan vi allerede forkaste nulhypotesen $\mathcal{H}_0$ i (6.3).

Dog udføres one-sample t-test forsat for at sammenligne middelværdier i anden del af (7.2). Til udførelsen af t-testen, antager vi, at U og V er uafhængige. For at teste om $\mu_U = 0$ anvendes en to-halet t-test, hvor vi undersøger, om $\bar{U} \neq \mu_U$. Her fås en teststørrelse

$$t = \frac{0.06 - 0}{0.59/\sqrt{681}} = 2.83$$

I Tabel B.5 er $|T_{krit}| = 1.96$ med signifikansniveau på 5% ved 680 frihedsgrader. Dermed er $|t| > |T_{krit}|$, hvorfor $\mathcal{H}_0$ i (6.3) forkastes.

I Figur 7.3 illustreres kovariansellipsen for $\hat{\Sigma}_{AB}$ med identitetslinjen. Her visualiseres det, at kovariansellipsen for $\hat{\Sigma}_{AB}$ ikke er symmetrisk omkring identitetslinjen, da kovariansellipsen ikke er ligeligt fordelt på begge sider af identitetslinjen.



Figur 7.3: Kovariansellipse for genotype AB og identitetslinjen.

7.3 Konstruktion af symmetrimodel

Som påvist forkastes nulhypotesen $\mathcal{H}_0$ i (6.3) og (6.4), hvorfor klyngerne ikke er symmetriske omkring identitetslinjen. Hvis nulhypoteserne $\mathcal{H}_0$ i (6.3) og (6.4) ikke blev forkastet,

ville en symmetrimodel med udgangspunkt i logit og referenceniveau AB kunne konstrueres ved

$$\begin{aligned} \log\left(\frac{P_{AA}}{P_{AB}}\right) &= \beta_{AA}X_A + \beta_{BB}X_B \quad \text{og} \\ \log\left(\frac{P_{BB}}{P_{AB}}\right) &= \beta_{BB}X_A + \beta_{AA}X_B. \end{aligned} \tag{7.4}$$

Modsat MLR-modellen opskrives symmetrimodellen i (7.4) uden en skæringskoefficient β_0, da denne koefficient påvirker, hvor beslutningsgrænsen mellem klyngerne for genotype AA og AB skærer med beslutningsgrænsen mellem klyngerne for genotype BB og AB. Ved at udelade skæringskoefficienten β_0 tvinges skæringspunktet til at ligge i origo. Dette kan have til fordel, at der kun bliver behov for at estimere to koefficienter til symmetrimodellen fremfor fire. Posterior sandsynlighederne for hver observation kan udregnes ved

$$\begin{aligned} P_{AA} &= \frac{\exp\left(\beta_{AA}X_A + \beta_{BB}X_B\right)}{1 + \exp\left(\beta_{AA}X_A + \beta_{BB}X_B\right) + \exp\left(\beta_{BB}X_A + \beta_{AA}X_B\right)} \\ P_{BB} &= \frac{\exp\left(\beta_{BB}X_A + \beta_{AA}X_B\right)}{1 + \exp\left(\beta_{AA}X_A + \beta_{BB}X_B\right) + \exp\left(\beta_{BB}X_A + \beta_{AA}X_B\right)} \\ P_{AB} &= \frac{1}{1 + \exp\left(\beta_{AA}X_A + \beta_{BB}X_B\right) + \exp\left(\beta_{BB}X_A + \beta_{AA}X_B\right)}. \end{aligned} \tag{7.5}$$

Koefficienterne β_{AA} og β_{BB} kan estimeres i R med udgangspunkt i funktionen `optim()` fra pakken `stats`. Funktionen `optim()` er en general optimeringsalgoritme, som både anvendes til at finde minimum og maksimum. Som angivet i Afsnit 2.4, bestemmes de ukendte koefficienter $\beta = (\beta_{AA}, \beta_{BB})^T$ vha. MLE. Her kan der konstrueres en likelihood funktion, som er givet ved

$$\mathcal{L}(\beta_c) = \prod_{i=1}^{4400}\prod_{c=1}^{2} \frac{\exp\left(\beta_{1,c}X_{A,i} + \beta_{2,c}X_{B,i}\right)^{y_{i,c}}}{\left(1 + \exp\left(\beta_{1,j}X_{A,i} + \beta_{B,j}X_{2,i}\right) + \exp\left(\beta_{2,j}X_{A,i} + \beta_{1,j}X_{B,i}\right)\right)^{-n_i}}, \tag{7.6}$$

hvor 1 eksempelvis er AA, 2 eksempelvis er BB og $y_{i,c}$ er antallet af observationerne i den c'te gruppe. `optim()` kræver, at likelihood funktionen (7.6) defineres som en funktion af hhv. datasættet og en vektor med de ukendte koefficienter $\beta = (\beta_{AA}, \beta_{BB})^T$. Herefter vælges en initialværdi for $\hat{\beta}$ i `optim()`, og output af `optim()` vil blandt andet være de estimerede koefficienter til symmetrimodellen.

Efter koefficienten β er estimeret, kan posterior sandsynlighederne for hver observation findes ved (7.5). Her vælges det at klassificere observationen som den genotype med størst posterior sandsynlighed. Dette gør det muligt både at illustrere klassifikationerne med symmetrimodellen samt at udarbejde konfusionsmatricer med henblik på at undersøge andelen af fejlklassifikationer og modelpræcision.

Ydermere kan der for symmetrimodellen konstrueres samme grid, som i Afsnit 5.2, med henblik på at bestemme en tærskel ω for no-call samt beslutningsgrænser for et no-call bånd. Dette kan anvendes til at sammenligne MLR-modellen og symmetrimodellen til at vurdere, om der opnås en tilsvarende klassifikationsmodel med symmetrimodellen.

8 Diskussion

I Kapitel 5 observeres det, at MLR-modellen fortsat fejlklassificerer observationer efter introduktionen af no-call. Formålet med denne rapport er at konstruere en model, som kan klassificere så mange observationer korrekt som muligt. Da MLR-modellen klassificerer observationerne ud fra deres posterior sandsynlighed, er det næsten uundgåeligt ikke at lave fejlklassifikationer. MLR-modellen har til fordel, at den giver et indblik i posterior sandsynligheden for hver genotype for hver observation fremfor kun en endelig klassifikation. Dette kan være en ulempe, der kan medføre, at modellen bliver overtilpasset på træningsdataet, da modellen trænes til at kalde en genotype ud fra signalintensiteten. Dette kan betyde, at modellen overvurderer nøjagtigheden af prædikationer, hvis der eksempelvis mangler signaler. Dette kan ligge til grundlag for de fejlklassifikationer MLR-modellen foretager efter no-call, herunder klassifikationen af de *outliers*, som findes i datasættet, altså de observationer, hvis ene signal er nul. Hensigten ved at introducere no-call i denne rapport, var blandt andet at opfange disse outliers og reducerer fejlklassifikationer, men det viser sig, at MLR-modellen er meget sikker i sine klassifikationer heraf, til trods for, at det ikke er samme klassifikation som angivet af WGS.

En mulighed for at formindske antallet af fejlklassifikationer er ved at ændre tærsklen ω. Dette vil være på bekostning af, hvor mange observationer, som får kaldt en genotype, herunder også en større del af de korrekt klassificerede observationer. Det er dog ikke altid en mulighed at ændre på tærsklen, da den oftest er bestemt af udbyderen, hvor formålet med den statistiske analyse har stor indflydelse. I stedet kan det diskuteres, om der er andre forhold, som har indflydelse på, hvorfor nogle af MLR-modellens fejlklassifikationer har så høj en posterior sandsynlighed.

I Figur 4.2 visualiseres det, at nogle observationer, hvis genotype er bestemt med WGS, falder indenfor klyngerne tilhørende en af de andre genotyper. Dette kan være en indikation på manglende signaler. I Tabel 8.1 ses nogle eksempler på outliers og fejlklassificerede observationer efter no-call, hvor MPS står for maksimal posterior sandsynlighed.

Obs #	Person	SNP	Signal A	Signal B	WGS	MLR	MPS
901	1	226	0	436	AA	BB	1.00
2213	1	554	9901	0	AB	AA	1.00
12	4	3	1025	9344	AA	BB	0.96
175	3	44	8247	18921	BB	AB	0.86

Tabel 8.1: Eksempler: Observation 901 og 2213 er outliers, og observation 12 og 175 er fejlklassifikationer.

Her ses det blandt andet, at observation 901 og 2213 er outliers, da de direkte mangler et signal. Det er klart herfra, at MLR-modellen klassificerer outliers efter det signal, der er. MLR-modellen kan ikke lave samme klassifikation som WGS, når der sker aflæsningsfejl af signalintensiteter eller lignende fra enten snipchippen eller ved WGS. Lignende ses det for observation 12 og 175, at MLR-modellen klassificerer efter det signal, som er målt til at være det stærkeste, selvom det ikke er samme klassifikation, som opnås med responsvariablen WGS. Specifikt for observation 175 ses en alternativ høj intensitet af Signal A sammenlignet med, at observationen klassificeres som BB med WGS. Dette kan ligeledes indikere en fejlaflæsning af signalet, manglende signal eller støj. Disse faktorer kan MLR-modellen ikke tage højde for, men den kan trænes til at klassificere efter det eller de signaler, som er stærkest. Dette afspejles i særdeleshed også i den maksimale posterior sandsynlighed for hver observation, hvor disse observationer tydeligvis ikke vil blive opfanget af den tærskel, som er fastsat i denne rapport.

9 Konklusion

Denne rapport har haft til formål at træne en model på signalerne fra sekventeringsmetoden NGS, og undersøge, om den trænede model opnår klassifikationer tilsvarende sekventeringsmetoden WGS. I rapporten konstrueres en klassikationsmodel med udgangspunkt i supervised learning, hvor der fokuseres på en multinomial logistisk regressionsmodel. Det udleverede datasæt fra Retsgenetisk Afdeling, Retsmedicinsk Institut på Københavns Universitet består af 1100 SNPer fra fire anonymiserede personer. I datasættet findes signaler fra NGS, som er navngivet A- og B-signaler, samt en variabel WGS, som viser genotypen bestemt med helgenomssekventering. Klassifikationsmodellen MLR blev trænet på signalerne A og B med og uden log-transformation samt responsvariablen WGS. I denne rapport blev 4000 observationer udtaget til træningsdata, og 400 observationer til test af modellen.

MLR-modellen med signalerne A og B opnåede en modelpræcision målt på accuracy på 95.8%, hvor en MLR-model med log-transformerede signaler A og B opnåede en modelpræcision på 96.3%. Da der blev opnået en højere modelpræcision med log-transformeret signaler, arbejdede vi i rapporten videre med denne model. Vi observerede, at MLR-modellen fejlklassificerede observationer, hvorfor en ny gruppe, kaldt no-call, blev indført. Gruppen no-call fik tildelt observationer, hvor MLR-modellen var for usikker til at kalde en genotype, sådan at observationer, hvis maksimale posterior sandsynlighed var under en valgt tærskel, blev tildelt no-call. Derfor blev der foretaget analyse af forholdet mellem forskellige valg af tærskel og modelpræsicion, korrekte klassifikationer og fejlklassifikationer. Tærsklen i denne rapport blev fastsat til 70%. Efter introduktionen af no-call blev modelpræcision forbedret til 97.7%, men der tildeles 5.9% af observationerne til no-call.

Efterfølgende blev MLR-modellen testet på henholdsvis de resterende 400 observationer og ved 11-fold krydsvalidering. Ved test på 400 blev der opnået en modelpræcision 98.1%, og ved 11-fold krydsvalidering blev der opnået en gennemsnitslig modelpræcision 97.7%. Her blev omkring $5.5 - 6\%$ af observationerne tildelt no-call. Dermed blev MLR-modellen vurderet til at være en acceptabel model til klassifikation af signaler givet med NGS i forhold til klassifikationerne opnået med helgenomssekventering.

I visualisering af genotyperne antydes et symmetrisk forhold omkring identitetslinjen. Vi undersøgte derfor, om der var symmetri med henblik på at kunne simplificere MLR-modellen og efterfølgende teste, om denne klassifikationsmodel ville klassificere tilsvarende med MLR-modellen. Test for symmetri blev foretaget med hypotesestestene: korrelationstest, t-test, Box's M-test og Hotellings T^2-test. Resultaterne heraf viste, at der var symmetri iblandt genotyperne AA og BBs middelværdier, men der var ikke symmetri iblandt deres kovariansmatricer, og genotypen AB var ikke symmetrisk omkring identitetslinjen. Derfor var det ikke muligt at konstruere en simplere model, men rapporten blev afsluttet med at give en metode til, hvordan en symmetrimodel kunne konstrueres.

10 Perspektivering

Behandling af no-call

I Figur 4.2b visualiseres klyngerne for de tre genotyper bestemt med WGS, hvor det er illustreret, at flere observationer er placeret væk fra deres respektive klynge. Disse observationer kaldes *outliers*. Placering af disse observationer kan skyldes, at der helt eller delvist mangler et signal for den pågældende SNP. Dette kan ligge til grundlag for nogle af de fejlklassifikationer, som MLR-modellen laver. Iblandt disse outliers er det interessant at fire observationer direkte mangler ét signal. Disse ses i Tabel 10.1.

Obs #	SNP	Person	Signal A	Signal B	WGS	MLR
901	226	1	0	436	AA	BB
2947	737	3	8852	0	AA	AA
2213	554	1	9901	0	AB	AA
2214	554	2	11415	0	AB	AA

Tabel 10.1: Oversigt over de fire observationer med manglende signal.

For disse observationer, er det interessant, at der ifølge WGS ses henholdvis to observationer fra AA og to observationer fra AB, som MLR-modellen med sikkerhed klassificere som henholdvis tre observationer tilhørende AA og én observation tilhørende BB. Dette er visualiseret på Figur 5.2. Imidlertid illustreres det fra Figur 5.3, at dette er observationer med høj maksimal posterior sandsynlighed, hvorfor disse observationer ikke bliver tildelt no-call.

Da disse observationer afviger fra resten af datasættet, kunne det være belejligt at fjerne disse observationer fra datasættet, men dette er ikke muligt, da alle observationer med minimum ét signal beholdes. Hvis observationer skal fjernes fra datasættet, er det ofte forudbestemt fra udbyder. Da dette datasæt er et lille udrag af et større datasæt, er denne sortering allerede foretaget og kan findes i [10].

Det er derfor interessant, om denne problemstilling kan imødekommes ved brug af andre modeller indenfor supervised eller unsupervised learning. Et løsningseksempel kan være med udgangspunkt i unsupervised learning med anvendelse af mikstur modellen EM-algoritmen, som blev bearbejdet i [8, Kap. 3] og [9, Kap. 6]. I [9, kap. 6] fremlægges ligeledes en metode til at tildele blandt andet disse fire observationer til no-call.

Personspecifikke modeller

En alternativ måde at behandle datasættet på, er ved at opdele datasættet efter personnummer $\{1, 2, 3, 4\}$, og udarbejde en MLR-model for hver person. For hver model vil datasættet blive opdelt i 1000 observationer som træningsdata og 100 observationer som testdata. Med denne fremgangsmåde vil det være interessant at undersøge, om en personspecifik model ville kunne opnå en større modelpræcision målt på accuracy, og dermed mindske antallet af fejlklassifikationer. Dermed også undersøge, om MLR-modellens prædikationer opnår en højere maksimal posterior sandsynlighed. Dette vil kunne illustreres med en figur tilsvarende Figur 5.3. I den forbindelse kan det også undersøges, om forholdet mellem korrekte klassifikationer, fejlklassifikationer og tærsklen ω ændrer sig. Dette vil også tydeliggøre, om det vil være muligt at vælge en højere tærskel end 70% uden at forstørre den procentvise andel af observationer, som tildeles no-call, og derved konstruere en model, som er mere sikker i klassifikation af genotyper.

Afslutningsvis kunne det være interessant at undersøge, om der er et symmetrisk forhold mellem genotyperne omkring identitetslinjen for hver person. I denne rapport forkastes hypotesen om dette symmetriske forhold mellem AA og BB, selvom der er symmetri iblandt middelværdierne. For MLR-modellen i denne rapport ses det, at det ikke vil kræve en stor ændring i kovariansmatricerne før nulhypotesen for symmetriske kovariansmatricer vil være opfyldt. Ved at opdele datasættet vil det kunne undersøges, om denne forskel vil blive udlignet eller om denne påstand kan forkastes. Disse resultater kan med fordel sammenlignes med resultaterne opnået i denne rapport og undersøge, om der sker en forbedring eller forværring af modelpræcisionen i MLR-modellen.

Krydsvalidering

I denne rapport er der valgt 11-fold krydsvalidering, hvor R inddeler dataet i gennemsnit i 4000/400-inddelinger. Det kunne i den forbindelse være interessant at opdele dataet i præcise 4000/400-inddelinger. På den måde kan det undersøges, om antallet af fejlklassifikationer, antallet af no-call og modelpræcision ligger mere stabilt i hver fold. Dette vil kunne give et bedre indblik i, hvor god MLR-modellen er.

En anden måde at validere MLR-modellen er ved at vælge et andet antal folder. Normalvis betragtes enten 5-fold krydsvalidering eller 10-fold krydsvalidering. Valget om en 11-fold krydsvalidering i denne rapport adskiller sig fra normalen, idet det er den inddeling, som er tilsvarende valget af opdelingen i træningsdata og testdata, sådan at resultaterne fra testdata og krydsvalidering ville være sammenlignelige.

Yderligere kunne det være interessant at anvende 11-fold krydsvalidering for varierende tærskler ω. Dette vil give et indblik i, hvilken tærskel, der vil være optimal for det ønskede resultat af den statistiske analyse. Desuden vil varierende tærskler give et indblik i, om den procentvise andel af observationer, som tildeles no-call i testdata, er tilsvarende den procentvise andel af observationer, som tildeles no-call i træningsdata. Krydsvalideringen

vil være med til at understrege om den procentvise andel er forholdvis stabil eller ustabil.

Yderligere undersøgelse af symmetri

Det kunne ligeledes være interessant at undersøge, om det vil være muligt at opnå et symmetrisk forhold mellem genotyperne omkring identitetslinjen ved manipulation af gruppestørrelselserne.

For genotyperne AA og BB skal det undersøges, hvornår $\Sigma_{AA} = \Sigma'_{BB}$. Dette kan løses ved at betragte udtrykkende for kovariansellipserne. Lad

$$\Sigma_{AA} = \begin{bmatrix} \sigma_A^2 & \rho\sigma_B^2 \\ \rho\sigma_A^2 & \sigma_B^2 \end{bmatrix} \quad \text{og} \quad \Sigma'_{BB} = \begin{bmatrix} \sigma_B^2 & \rho\sigma_A^2 \\ \rho\sigma_B^2 & \sigma_A^2 \end{bmatrix},$$

hvor σ_A^2 angiver variansen for X_A, og σ_B^2 angiver variansen for X_B. Formlen for kovariansellipsen for AA givet ved

$$\begin{aligned} x_{AA}(t) &= \sqrt{\lambda_A}\cos(\theta_{AA})\cos(t) - \sqrt{\lambda_B}\sin(\theta_{AA})\sin(t) + \mu_{X_A}, \quad \text{og} \\ y_{AA}(t) &= \sqrt{\lambda_A}\cos(\theta_{AA})\cos(t) - \sqrt{\lambda_B}\sin(\theta_{AA})\sin(t) + \mu_{X_B}, \end{aligned}$$

og for BB, er kovariansellipsen givet ved

$$\begin{aligned} x_{BB}(t) &= \sqrt{\lambda_B}\cos(\theta_{BB})\cos(t) - \sqrt{\lambda_A}\sin(\theta_{BB})\sin(t) + \mu_{X_B}, \quad \text{og} \\ y_{BB}(t) &= \sqrt{\lambda_B}\cos(\theta_{BB})\cos(t) - \sqrt{\lambda_A}\sin(\theta_{BB})\sin(t) + \mu_{X_A}, \end{aligned}$$

hvor λ_A og λ_B angiver egenværdierne i kovariansmatricen, μ_{X_A} og μ_{X_B} angiver middelværdierne af henholdvis signal X_A og signal X_B, $t = 0, \ldots, 2\pi$ og

$$\theta_{AA} = \tan^{-1}\left(\frac{\lambda_A - \sigma_A^2}{\sigma_B^2}\right) \quad \text{og} \quad \theta_{BB} = \tan^{-1}\left(\frac{\lambda_B - \sigma_B^2}{\sigma_A^2}\right).$$

Derved skal ligningerne $x_{AA}(t) = x_{BB}(t)$ og $y_{AA}(t) = y_{BB}(t)$ løses for λ_A og λ_B med henblik på at undersøge, hvornår $\Sigma_{AA} = \Sigma'_{BB}$. Løsningen af disse ligninger vil kunne findes ved manipulation af gruppestørrelserne n_{AA} og n_{BB}.

For genotype AB kan det ligeledes undersøges, om der skal anvendes en manipulation af gruppens størrelse for at opnå et symmetrisk forhold i genotypen AB omkring identitetslinjen. Dette kan gøres ved at undersøge, hvad der skal til for, at de stokastiske variabler U og V bliver uafhængige samt at U opnår en middelværdi på 0.

Hvis det er muligt at opnå et symmetrisk forhold mellem genotyperne omkring identitetslinjen, kan der efterfølgende opstilles en symmetrimodel fra konstruktionen præsenteret i Afsnit 7.3, og undersøge om, der opnås en klassifikationsmodel tilsvarende MLR-modellen, som er konstrueret i denne rapport. At undersøge om der opnås en tilsvarende model, kan gøres ved at evaluere modellen ved konstruktion af grid, tildeling til no-call, test og krydsvalidering.

11 Litteratur

[1] Christiansen MV, Steen SE, Uhrbrand SL. *Master-scripts*. GitHub. 2022.

[2] Berg J, Tymoczko J, Gatto G, Stryer L. *Biochemistry*. 8th ed. Macmillan; 2015. ISBN: 978-1-464-12610-9.

[3] Illumina. *Microarray technology*. ̇2022. URL: https://www.illumina.com/science/technology/microarray.html.

[4] James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. 2nd ed. Springer, New York, NY; 2021. ISBN: 978-1-071-61417-4. URL: https://link.springer.com/content/pdf/10.1007%2F978-1-0716-1418-1.pdf.

[5] Madsen H, Thyregod P. *Introduction to General and Generalized Linear Models*. 5th ed. CRC Press/Chapmann & Hall; 2011. ISBN: 978-1-420-091557-0.

[6] Czepiel SA. *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*. . 2015. URL: https://czep.net/stat/mlelr.pdf.

[7] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. Springer, New York, NY; 2009. ISBN: 978-0-387-84858-7. URL: https://link.springer.com/content/pdf/10.1007%2F978-0-387-84858-7.pdf.

[8] Pedersen CM, Steen SE, Uhrbrand SL. *Klassificering af genotyper i retsgenetiske problemstillinger*. ̇2022.

[9] Christensen AB, Christensen JB, Østerby Jespersen L, Larsen MF, Jeppesen PL. *Bestemmelse af genotyper i retsgenetiske problemstillinger via klyngealgoritmer*. ̇2022.

[10] Andersen MM, Christiansen SN, Andersen JD, Eriksen S, Morling N. *SNP allele calling of Illumina Infinium Omni5-4 data using the butterfly method*. ̇2022. URL: https://doi.org/10.1038/s41598-022-22162-8.

[11] Seber GAF. *Multivariate Observations*. 2nd ed. John Wiley & Sons, Inc; 1984. ISBN: 0-471-88104-X.

[12] Wikipedia. *Fisher tranformation*. Wikipedia. 2022. URL: https://en.wikipedia.org/wiki/Fisher_transformation.

[13] Wikipedia. *Correlation*. Wikipedia. 2022. URL: https://en.wikipedia.org/wiki/Correlation.

[14] Wikipedia. *Student's t-test*. Wikipedia. 2022. URL: https://en.wikipedia.org/wiki/Student%27s_t-test.

[15] Jiamwattanapong K, Ingadapa N, Plubin B. *On Testing Homogeneity of Covariance Matrices with Box's M and the Approximate Tests for Multivariate Data*. Europe-

an Journal of Applied Science. 2022. URL: https://drive.google.com/file/d/1s062ekBDAM6V7dFWP25-K25IPH2D77F4/view.

[16] Box GEP. *A General Distribution Theory for a Class of Likelihood Criteria*. Oxford Journals, Oxford University Press. 1949. URL: https://www.jstor.org/stable/2332671#metadata_info_tab_contents.

[17] Zaiontz C. *Box's M test basic concepts*. ̇2022. URL: https://www.real-statistics.com/multivariate-statistics/boxs-test/boxs-test-basic-concepts/.

[18] Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. 4th ed. Academic Press; 1979 (1995). ISBN: 978-0-12-471250-8.

[19] McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. Chapman and Hall; 2009. ISBN: 978-1-4200-9155-7. URL: https://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/glmbook.pdf.

[20] Wikipedia. *Likelihood function*. Wikipedia. 2022. URL: https://en.wikipedia.org/wiki/Likelihood_function.

[21] Andersen LN. *Multivariat Statistisk Analyse I*. Institut for Matematik Fag, Aarhus Universitet. 2020.

Appendiks

A Appendiks A

A.1 Generaliserede lineære modeller (GLM)

Dette afsnit er baseret på [5, kap. 3] og [19, kap. 2] og er et supplement til Kapitel 2.

Definition A.1. Eksponentiel familie
Lad $Z = (Z_1, \ldots, Z_n)$ være en stokastisk variabel. En fordeling tilhører den eksponentielle familie, hvis dens tæthed kan skrives som

$$f(Z; \theta, \kappa) = \exp\Big(\kappa(\theta^T Z - b(\theta)) + c(Z, \kappa)\Big)$$

for en kanonisk parameter $\theta \in \Omega \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, dispersion parameter $\kappa > 0$ og funktioner $b : \Omega \to \mathbb{R}$, $c : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$.

Bemærkning. *Den underliggende familie, hvor $\kappa = 1$, kaldes den naturlige eksponentielle familie.*

Definition A.2. Generaliserede lineære modeller
Lad $Z = (Z_1, \ldots, Z_n)$ være en stokastisk variabel med forventede værdier $\mathbb{E}[Z_i] = \mu_i$ for $i = 1, \ldots, n$ og lad $Y \in \mathbb{R}^{n \times p}$ for $n, p \in \mathbb{N}$. Antag, at $\mu_i \in M \subseteq \mathbb{R}$ og lad $x_i = y_i^T \beta$ for $i = 1, \ldots, n$ og for et tilfældigt $\beta \in \mathbb{R}^p$. Lad $g : M \to \mathbb{R}$ være en invertibel funktion.

Så siges Z at følge en generaliseret lineær model med designmatrix Y, linkfunktion g og lineære prædiktore $x_1, \ldots, x_n$, hvis Z_i følger en fordeling fra den eksponentielle dispersion familie, hvor Z_i er uafhængig af Z_j og $x_i = g(\mu_i)$ for $i, j = 1, \ldots, n$, $i \neq j$.

Sætning A.3.
Logistisk regression defineret i Definition 2.1 er en GLM.

Bevis:
Lad $Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Bin}(n_i, h_i)$, lad x_i angive antallet af succeser og den lineære prædiktor for det i'te element for $i = 1, \ldots, n$, og lad $Z_i = \frac{Y_i}{n_i}$ sådan at $\mathbb{E}[Z_i] = h_i$, $Z_i \perp\!\!\!\perp Z_j$ for $i \neq j$, og lad

$$h_i = \frac{h_i}{1 - h_i}(1 - h_i) = \frac{h_i}{1 - h_i} \frac{1}{1 + \dfrac{h_i}{1 - h_i}} = \frac{\exp(x_i)}{1 + \exp(x_i)}$$

angive sandsynligheden for en succes for det i'te element, når modellen anvender en logistisk funktion på den lineære prædiktor på det i'te element. Så gælder det, at

$$
\begin{aligned}
f_{Z_i}(z) = \frac{d}{dz}F_{Z_i}(z) = \frac{d}{dz}F_{Y_i}(n_i z) = n_i f_{Y_i}(n_i z) = n_i \binom{n_i}{n_i z} h_i^{n_i z}(1-h_i)^{n_i - n_i z} \\
= n_i \binom{n_i}{n_i z} \exp\left(n_i z \log(h_i) + (n_i - n_i z)\log(1-h_i)\right) \\
= n_i \binom{n_i}{n_i z} \exp\left(n_i \left(z \log\left(\frac{h_i}{1-h_i}\right) + \log(1-h_i)\right)\right),
\end{aligned}
$$

som viser, at fordelingen af Z_i tilhører den eksponentielle familie med $\theta_i = L(h_i)$, $b(\theta) = \log(1+\exp(\theta))$ og $\kappa = n_i$. ■

A.2 Maksimum likelihood estimat (MLE)

Dette afsnit er baseret på [5, kap. 2] og [20] og er et supplement til Kapitel 2.

Definition A.4. Likelihood funktion
Lad $Z = (Z_1, \ldots, Z_n)^T$ være diskrete stokastiske variabler med frekvensfunktion afhængig af en parameter β. Så er likelihood funktionen defineret ved

$$\mathcal{L}(\beta; Z) = \prod_{i=1}^{n} P(Z = z_i|\beta).$$

Definition A.5. Log-likelihood funktion
Lad $Z = (Z_1, \ldots, Z_n)^T$ være diskrete stokastiske variabler. Så er log-likelihood funktionen defineret ved

$$\ell(\beta; Z) = \log(\mathcal{L}(\beta; Z)).$$

Definition A.6. Maksimum likelihood estimat (MLE)
Lad $Z = (Z_1, \ldots, Z_n)^T$, så siges et maksimum likelihood estimat (MLE) at eksistere, hvis $\ell(\hat{\beta}; Z)$ har et entydigt maksimum, og er i så fald det $\hat{\beta}(Z)$, der maksimerer $\ell(\beta; Z)$.

A.3 Fordelinger

Dette afsnit er baseret på [11, kap. 2] og [5, app. B] og er et supplement til Kapitel 6.

Definition A.7. χ^2-fordeling
Hvis $Z_1, \ldots, Z_n$ er uafhængigt standard normalfordelte stokastiske variabler $N(0,1)$, så følger deres kvadratsum

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n),$$

hvor n angiver antallet af frihedsgrader.

Definition A.8. t-fordeling
Lad $Z_1, \ldots, Z_n$ være uafhængige normalfordelte stokastiske variabler $N(\mu, \sigma^2)$. Lad

$$\bar{Z} = \frac{1}{n}\sum_{j=1}^{n} Z_j \quad \text{og} \quad S_i = \frac{1}{n-1}\sum_{j=1}^{n}(Z_j - \bar{Z})^2,$$

hvor $\bar{Z}$ angiver stikprøvemiddelværdien og S_i angiver stikprøvekovariansen. Så

$$X = \frac{\bar{Z} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{og} \quad Y = \frac{(n-1)S_i}{\sigma^2} \sim \chi^2(n-1).$$

Da de to stokastiske variabler X og Y er uafhængige, gælder det, at

$$t = (\bar{Z} - \mu)\sqrt{\frac{n}{S_i}} \sim t(n-1).$$

Definition A.9. F-fordeling
Lad Z_1 og Z_2 være uafhængige stokastiske variabler, og lad

$$Z = \frac{Z_1/df_1}{Z_2/df_2},$$

hvor $Z_1 \sim \chi^2(df_1)$ og $Z_2 \sim \chi^2(df_2)$, og hvor df_1 og df_2 angiver antallet af frihedsgrader. Så følger Z en F-fordeling med df_1 og df_2 frihedsgrader, noteret $Z \sim F(df_1, df_2)$.

Definition A.10. Wishart-fordeling
Antag, at $Z_1, Z_2, \ldots, Z_n$ er uafhængige og identisk fordelte stokastiske variabler, hvor Z_i følger en p-multivariat normalfordeling $N_p(0, \Sigma)$. Så siges

$$V = \sum_{i=1}^{n} Z_i Z_i^T$$

at have en Wishart fordeling med n frihedsgrader, noteret $V \sim W_p(n, \Sigma)$.

Sætning A.11.
Lad $Z = (Z_1, Z_2, \ldots, Z_n)^T$ være uafhængige og identisk fordelte stokastiske variabler, hvor $Z_i \sim N_p(\mu, \Sigma)$, lad $\bar{Z}$ angive stikprøvemiddelværdien af Z, og lad

$$S_i = \frac{1}{n-1} Q_i \quad \text{og} \quad Q_i = \sum_{i=1}^{n} (Z_i - \bar{Z})^2.$$

Så gælder det, at $\bar{Z} \sim N_p(\mu, \frac{1}{n}\Sigma)$, $Q_i \sim W_p(n-1, \Sigma)$ og $\bar{Z}$ og S_i er uafhængige.

Korollar A.12.
For $p = 1$ i Sætning A.11 så $\bar{Z} \sim N(\mu, \frac{\sigma^2}{n})$, $Q_i \sim \chi^2(n-1)$.

A.4 Multivariat normalfordeling

Dette afsnit er baseret på [21, kap. 2] og [11, kap. 3.6.1] og er et supplement til Kapitel 6.

Definition A.13. Multivariat normalfordeling
En p-dimensional stokastisk variabel Z er p-dimensionalt normalfordelt, noteret

$$Z \sim N_p(\mu, \Sigma),$$

hvis den én-dimensionale stokastiske variabel $a^T Z$ er normalfordelt for enhver fast vektor $a = (a_1, a_2, \ldots, a_p)^T \in \mathbb{R}^p$. Her angiver μ middelværdien, hvor $\mu \in \mathbb{R}^p$, og $\Sigma \in \mathbb{R}^{p \times p}$ angiver kovariansmatricen, hvor Σ er positiv semidefinit.

Sætning A.14.
Lad Z være en p-dimensional stokastisk variabel. Hvis $Z \sim N_p(\mu, \Sigma)$ er regulært normaltfordelt, så gælder det, at

$$(Z - \mu)^T \Sigma^{-1} (Z - \mu) \sim \chi^2(p).$$

Sætning A.15.
Lad Z være en p-dimensional stokastisk variabel. Antag, at $Z \sim N_p(\mu, \Sigma)$ og lad

$$Z = \begin{bmatrix} Z^{(1)} \\ Z^{(2)} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{A.1}$$

hvor $Z^{(i)}$ og $\mu^{(i)}$ er $p_i \times 1$ vektorer og Σ_{ij} er en $p_i \times p_j$ matrix, $p_1 + p_2 = p$. Da gælder

1. Enhver delmængde af elementerne i Z har en multivariat normalfordeling, specielt $Z^{(1)} \sim N_{p_1}(\mu^{(1)}, \Sigma_{11})$.
2. $Z^{(1)}$ og $Z^{(2)}$ er uafhængige hvis og kun hvis $\mathrm{Cov}(Z^{(1)}, Z^{(2)}) = 0$.
3. Den betingede fordeling af $Z^{(1)}$ givet $Z^{(2)} = z^{(2)}$ er

$$N_{p_2}\left(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}[Z^{(2)} - \mu^{(2)}], \Sigma_{11\cdot 2}\right), \quad \Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

Sætning A.16.
Lad $Z = (Z_1, Z_2, \ldots, Z_n)^T$ er uafhængige stokastiske variabler, hvor $Z_i \sim N_p(0, \Sigma)$, og lad $Y = X^T\psi$, hvor $\psi \neq 0$ og ψ er en $p \times 1$ vektor af konstanter. Lad A og B være $n \times n$ symmetriske matricer med henholdsvis rank r og s, og lad b være en $n \times 1$ vektor af konstanter. Lad W_p betegne Wishart-fordelingen. Da gælder det, at

1. $Z^TAZ \sim W_p(r, \Sigma)$ hvis og kun hvis $Y^TAY \sim \sigma_\psi^2\chi^2(r)$ for ethvert ψ, hvor $\sigma_\psi^2 = \psi^T\Sigma\psi$.
2. $Z^TAZ \sim W_p(r, \Sigma)$ og $Z^TBZ \sim W_p(s, \Sigma)$ har uafhængige Wishart fordelinger med henholdsvis r og s frihedsgrader, hvis og kun hvis $Y^TAY \sim \sigma_\psi^2\chi^2(r)$ og $Y^TBY \sim \sigma_\psi^2\chi^2(s)$ for ethvert ψ.
3. $Z^Tb \sim N_p(0, \Sigma)$ og $Z^TAZ \sim W_p(r, \Sigma)$ er uafhængige fordelte hvis og kun hvis $Y^Tb \sim N(0, \sigma^2)$ og $Y^TAY/\sigma_\psi^2 \sim \chi^2(r)$ er uafhængige fordelte for ethvert ψ.

Proposition A.17.
Lad $Z_1, Z_2, \ldots, Z_p$ være uafhængige og identisk fordelte $N(\mu, \sigma^2)$, og lad

$$Q_i = \sum_{j=1}^{p}(Z_j - \bar{Z})^2,$$

hvor $\bar{Z}$ angiver stikprøvemiddelværdien af Z. Så er $\bar{Z}$ uafhængig af Q_i og

$$Q_i/\sigma^2 \sim \chi^2(n-1).$$

A.4.1 MLE for multivariat normalfordeling

Sætning A.18. MLE for multivariat normalfordeling
Antag, at $Z_1, \ldots, Z_n$ er uafhængig og identisk fordelt stokastiske variabler, hvor $Z_i \sim N_p(\mu, \Sigma)$ og $n - 1 \geq p$. Så er maksimum likelihood estimatet for μ, noteret $\hat{\mu}$, givet ved stikprøvemiddelværdien $\bar{Z}$, og maksimum likelihood estimatet for Σ, noteret $\hat{\Sigma}$, er stikprøvekovariansen givet ved

$$\frac{1}{n}Q = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^T$$

Da er likelihood funktionen for en multivariat normalfordeling givet ved

$$\mathcal{L}(\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Z_i - \mu)^T\Sigma^{-1}(Z_i - \mu)\right), \tag{A.2}$$

og maksimum likelihood estimatet er udtrykt ved

$$\mathcal{L}(\hat{\mu}, \hat{\Sigma}) = (2\pi)^{-np/2}|\hat{\Sigma}|^{-n/2}\exp(-np/2). \tag{A.3}$$

A.4.2 Likelihood ratio test for multivariat normalfordeling

Lad $\mathcal{L}_i(\mu_i, \Sigma_i)$ antage formen af (A.2), sådan at likelihood funktionen

$$\mathcal{L}_{12}(\mu_1, \mu_2, \Sigma_1, \Sigma_2) = \mathcal{L}_1(\mu_1, \Sigma_1)\mathcal{L}_2(\mu_2, \Sigma_2).$$

At maksimere $\mathcal{L}_{12}$ er ækvivalent med at maksimere alle $\mathcal{L}_i$ samtidig, sådan at $\mathcal{L}_{12}$ maksimeres ved

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\Sigma}_1 = Q_1/n_1 = \sum_{i=1}^{n_i}\frac{(x_i - \bar{x})(x_i - \bar{x})^T}{n_i}$$
$$\hat{\mu}_2 = \bar{y}, \quad \hat{\Sigma}_2 = Q_2/n_2 = \sum_{i=1}^{n_i}\frac{(y_i - \bar{y})(y_i - \bar{y})^T}{n_i}$$

Fra (A.3) er maksimum af $\mathcal{L}_{12}$ givet ved

$$\begin{aligned}\mathcal{L}_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2) &= \mathcal{L}_1(\hat{\mu}_1, \hat{\Sigma}_1)\mathcal{L}_2(\hat{\mu}_2, \hat{\Sigma}_2) \\ &= (2\pi)^{-np/2}|\hat{\Sigma}_1|^{-n/2}|\hat{\Sigma}_2|^{-n/2}\exp(-np/2),\end{aligned}$$

hvor $n = n_1 + n_2$. Sæt nu $\Sigma_1 = \Sigma_2 = \Sigma$, hvorefter det ønskes at maksimere

$$\begin{aligned}\ell_{12}(\mu_1, \mu_2, \Sigma, \Sigma) =& \ell_1(\mu_1, \Sigma) + \ell_2(\mu_2, \Sigma) \\ =& c - \frac{n\log(|\Sigma|)}{2} - \frac{\operatorname{tr}(\Sigma^{-1}Q)}{2} \\ &+ \frac{\operatorname{tr}\left(\Sigma^{-1}\left[n_1(\bar{x} - \mu_1)(\bar{x} - \mu_1)^T + n_2(\bar{y} - \mu_2)(\bar{y} - \mu_2)^T\right]\right)}{2},\end{aligned}$$

hvor $c = \frac{np\log(2\pi)}{2}$ og $Q = Q_1 + Q_2$. Siden

$$\begin{aligned}\operatorname{tr}\left(\Sigma^{-1}\left[n_1(\bar{x}-\mu_1)(\bar{x}-\mu_1)^T\right]\right) &= n_1(\bar{x}-\mu_1)^T\Sigma^{-1}(\bar{x}-\mu_1) \geq 0\\ \operatorname{tr}\left(\Sigma^{-1}\left[n_2(\bar{y}-\mu_2)(\bar{y}-\mu_2)^T\right]\right) &= n_2(\bar{y}-\mu_2)^T\Sigma^{-1}(\bar{y}-\mu_2) \geq 0,\end{aligned}$$

er ℓ_{12} maksimeret for ethvert positiv definit Σ, hvor $\mu_1 = \bar{x} = \hat{\mu}_1$ og $\mu_2 = \bar{y} = \hat{\mu}_2$. Dermed er

$$\ell_{12}(\hat{\mu}_1, \hat{\mu}_2, \Sigma, \Sigma) \geq \ell_{12}(\mu_1, \mu_2, \Sigma, \Sigma).$$

I næste trin maksimeres

$$\ell_{12}(\mu_1, \mu_2, \Sigma, \Sigma) = c - \frac{n\log(|\Sigma|)}{2} - \frac{\operatorname{tr}(\Sigma^{-1}Q/n)}{2} \tag{A.4}$$

med hensyn til en positiv definit Σ. Siden hvert Q_i er positiv definit med sandsynlighed 1, så er Q dermed positiv definit, og (A.4) er maksimeret, når $\Sigma = \hat{\Sigma} = Q/n$. Dermed er

$$\ell_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}, \hat{\Sigma}) \geq \ell_{12}(\hat{\mu}_1, \hat{\mu}_2, \Sigma, \Sigma) \geq \ell_{12}(\mu_1, \mu_2, \Sigma, \Sigma),$$

sådan at $\hat{\mu}_1, \hat{\mu}_2$ og $\hat{\Sigma}$ er maksimum likelihood estimater af μ_1, μ_2, og Σ under nulhypotesen $\mathcal{H}_0$. Dermed kan (A.4) omskrives til

$$\ell_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}, \hat{\Sigma}) = c - \frac{n\log(|\hat{\Sigma}|)}{2} - \frac{np}{2}, \tag{A.5}$$

sådan at

$$\mathcal{L}_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2) = (2\pi)^{-np/2}|\hat{\Sigma}|^{-n/2}\exp(-np/2).$$

Proposition A.19. Likelihood ratio teststørrelse
Likelihood ratio teststørrelsen er givet ved

$$\lambda = \frac{\mathcal{L}_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}, \hat{\Sigma})}{\mathcal{L}_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_1, \hat{\Sigma}_2)} = \frac{|\hat{\Sigma}|^{-n/2}}{|\hat{\Sigma}_1|^{-n_1/2}|\hat{\Sigma}_2|^{-n_2/2}} = c_{12}\frac{|Q_1|^{n_1/2}|Q_2|^{n_2/2}}{|Q_1+Q_2|^{(n_1+n_2)/2}}, \tag{A.6}$$

hvor

$$c_{12} = \frac{n^{np/2}}{n_1^{n_1p/2}n_2^{n_2p/2}}.$$

Som (A.6) angiver sammenlignes to modeller i likelihood ratio testen med udgangspunkt i at finde den bedste model. Den ene model vil være et specialtilfælde af den anden model. Størrelsen af modellen kan eksempelvis bestemmes af antallet af frihedsgrader. Nulhypotesen $\mathcal{H}_0$ angiver, at den mindste model er den bedste model, hvor den alternative hypotese

$\mathcal{H}_1$ angiver, at den største model er den bedste model. Altså, $\mathcal{H}_0$ forkastes, når den største model er en forbedring af den mindste model.

Likelihood ratio testen angiver derfor den relative forskel mellem de to modeller likelihood vha. frihedsgrader, hvorfor en likelihood ratio test er $-2\log(\lambda) \sim \chi^2(df)$, $df = \frac{1}{2}p(p+1)$ når nulhypotesen $\mathcal{H}_0$ er sand.

B Appendiks B

B.1 Kritiske værdier

De følgende tabeller er uddrag af kritiske værdier for forskellige statistiske tests. Tabellerne er konstrueret ved hjælp af funktionerne `qchisq()` (χ^2_{krit}), `qf()` (F_{krit}), `qt()` (T_{krit}) og `qnorm()` (Z_{krit}) i R-studio og R (version 4.2.1).

Frihedsgrader	Sandsynlighed										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
	Ikke signifikant								Signifikant		

Tabel B.1: Kritiske værdier for χ^2-test.

$df_2 \backslash df_1$	1	2	3	4	5	6
1	4052.19	4999.52	5403.34	5624.63	5763.65	5858.97
2	98.50	99.00	99.17	99.25	99.30	99.33
3	34.12	30.82	29.46	28.71	28.24	27.91
4	21.20	18.00	16.69	15.98	15.52	15.21
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3716	6.64	4.61	3.79	3.32	3.02	2.81

Tabel B.2: Kritiske værdier for F-test med signifikans $\alpha = 0.01$.

$df_2 \backslash df_1$	1	2	3	4	5	6
1	161.45	199.5	215.71	224.58	230.16	233.99
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3716	3.84	3.00	2.61	2.37	2.22	2.10

Tabel B.3: Kritiske værdier for F-test med signifikans $\alpha = 0.05$.

Signifikansniveau	$\lvert Z_{krit} \rvert$
0.1	1.6449
0.05	1.9600
0.01	2.5758

Tabel B.4: Kritiske værdier for to halet Z-test med 3 forskellige signifikansniveauer.

Frihedsgrader	0.05
1	$\pm$12.706205
2	$\pm$4.302653
3	$\pm$3.182446
4	$\pm$2.776445
5	$\pm$2.570582
$\vdots$	$\vdots$
679	$\pm$1.963464
680	$\pm$1.963459

Tabel B.5: Kritiske værdier for to halet t-test med signifikans $\alpha = 0.05$.

C Appendiks C

C.1 Beregninger til Box's M-test i Maple (version 2021.2)

```
S1 := Matrix(2,2,[[0.3377493, 0.0586135],[0.0586135, 0.5338129]]):     (1)
Determinant(S1)
                          0.1768593909                                  (2)

S2 := Matrix(2,2,[[0.28428990, 0.07066685],[0.07066685, 0.51651521]]): (3)
Determinant(S2)
                          0.1418462537                                  (4)

S := Matrix(2,2,[[0.30770446, 0.06538763],[0.06538763, 0.52409141]]):  (5)
Determinant(S)
                          0.1569897221                                  (6)

(2)^((1629 - 1)/2);
                       3.709315158*10^(-613)                            (7)

(4)^((2090 - 1)/2);
                       1.185128130*10^(-886)                            (8)

(6)^((3719 - 2)/2);
                      3.362511303*10^(-1495)                            (9)

M := ((7)*(8))/(9);
                          0.0001307360285                              (10)

c := (2*2^2+3*2-1)/(6*(2+1)*(2-1))*((1/(1629-1)-1/(3719-2))
       +(1/(2090-1)-1/(3719-2)));
                             0.00040                                   (11)

-2*log(M)*(1 - c);
                          17.87749341                                  (12)
```