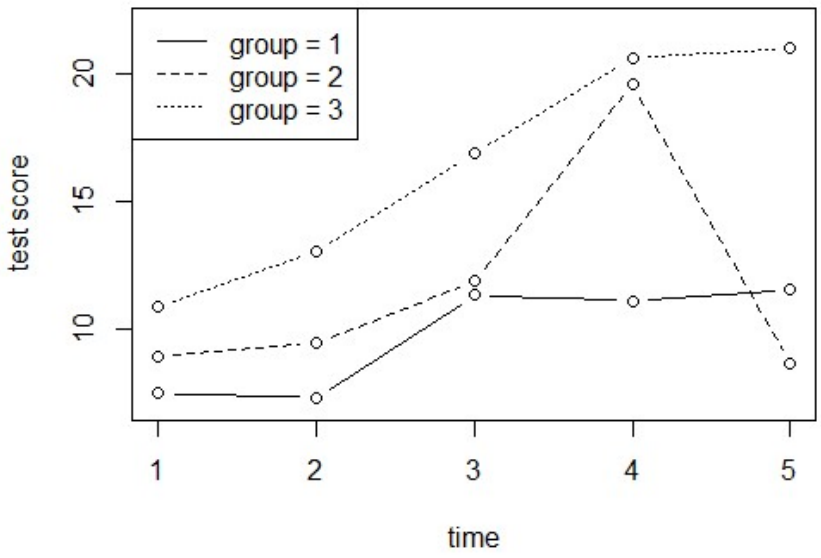
# Growth Mixture modeller

Longitudinelle data med latente grupperinger

Kandidatspeciale




Frederik Grove Hougaard
Tea Caroline Konnerup
Anders Lindgaard Sørensen

Aalborg Universitet
Institut for Matematiske Fag
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

**Institut for Matematiske Fag**
Aalborg Universitet
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

**Titel:**

Growth Mixture modeller - Longitudinelle data med latente grupperinger

**Projekt:**

Kandidatspeciale

**Projektperiode:**

1. februar 2023 - 2. juni 2023

**Projektgruppe:**

STAT 10 4.111d

**Deltagere:**

Frederik Grove Hougaard
Tea Caroline Konnerup
Anders Lindgaard Sørensen

**Vejleder:**

Rasmus Waagepetersen

**Sider:** 61

**Afsluttet:** 2. juni 2023

Synopsis:

Dette kandidatspeciale tager udgangspunkt i et datasæt bestående af folkeskoleelever, der over fire peroder er blevet undervist i forskellige emner inden for matematik. Specialet indeholder en analyse af, hvordan disse elever har udviklet deres brøkregningskompetencer over disse perioder, og yderligere hvordan eleverne kan grupperes samt den relevante underliggende teori.
Datsættet indeholder dog manglende værdier i responsvariablen, hvorfor vi indledningsvist undersøger, om disse værdier er *missing at random*, således at de kan udelades fra analysen.
Ved at inddele eleverne i højt, lavt og middel præsterende grupper, baseret på hvordan de har præsteret i de danske nationale tests, udføres en analyse af om elevgrupperne udvikler deres brøkregningskompetencer forskelligt i de fire perioder, og om dette stemmer overens med, hvad de er blevet undervist i. Dette gør vi ved at benytte en *mixed model*, hvorfor den underbyggende teori herfor beskrives.
En anden inddeling kan ske ved brug af *Growth Mixture modeller (GMM)*, som udover at modellere elevernes *vækstkurver* også finder en passende gruppeinddeling af eleverne for det antal *latente* grupper, der ønskes. Parameterestimation for en GMM kan dog være vanskeligt, hvorfor vi introducerer numeriske metoder til dette som *Marquardt-* og *EM-algoritmen*. Brugen af en GMM muliggør yderligere undersøgelsen af, hvilket antal grupper der er mest passende til data, hvortil informationskriterier og en likelihood ratio-test beskrives.

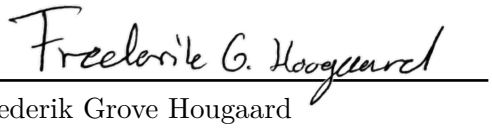
# Forord

Det indeværende kandidatspeciale er udarbejdet af studerende i STAT-gruppen 4.111d på tiende og afsluttende semester, på matematikuddannelsen ved det Ingeniør- og Naturvidenskabelige Fakultet på Aalborg Universitet, i foråret 2023. Projektets omfang er afgrænset af, at det henvender sig til matematikstuderende med godt kendskab til sandsynlighedsteori og statistik inferens.

Projektet følger konventionerne for Vancouver-formatet til referering af kilder. Alle kilder nummereres med et tal, [n], som da kan findes i litteraturlisten under dette nummer. Yderligere nummereres eksempler, figurer og tabeller løbende efter kapitel. Ydermere vil det statistiske programmeringssprog R blive brugt til at udføre simuleringer og statistiske analyser.

Slutteligt ønsker projektets forfattere at rette en stor tak til Rasmus Waagepetersen for behjælpelige råd og vejledning.

**Underskrifter**



Frederik Grove Hougaard

Tea Caroline Konnerup

Anders Lindgaard Sørensen

# Indholdsfortegnelse

# 1 | Introduktion

Vi tager i det indeværende kandidatspeciale udgangspunkt i datasættet benyttet i studiet *Differences in High- and Low-Performing Students' Fraction Learning in the Fourth Grade* [24]. Dette studie omhandler en undersøgelse af, hvordan elever i fjerde klasse har udviklet sig inden for brøkregning over fire perioder. Eleverne er i disse fire perioder blevet undervist i emner som multiplikation, division, brøkregning og ligninger med henblik på at teste om højt og lavt præsterende elever har forskellige udviklingsmønstre. En oversigt over, hvilke emner eleverne er undervist i, og i hvilke perioder, kan findes i [24, s. 5]. Før og efter hver af disse perioder har hver elev deltaget i en såkaldt *curriculum-based measurement test (CBM-test)*, som er en kort test, der kun vedrører deres brøkregningskompetencer.

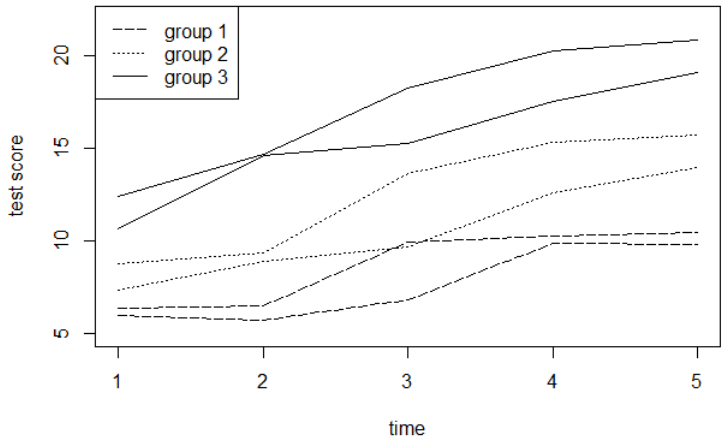
| **variabel** | **forklaring** | **værdi** | **antal elever** |
|---|---|---|---|
| `id` | elev-id | $[1, 587]$ | 431 |
| `test` | CBM-score | $[0, 35]$ | |
| `time` | nummer på test | $[1, 5]$ | |
| `gender` | køn | 1 = dreng<br>0 = pige | 206<br>225 |
| `age` | alder i 2018 | [9,12] | |
| `class_id` | klasse-id | [1,22] | $[13, 26]$ |
| `school_id` | skole-id | [1,11] | $[13, 95]$ |
| `intervention` | interventionsgruppe | 1 = undervisningsforløb 1<br>2 = undervisningsforløb 2 | 218<br>213 |
| `group` | præstationsgruppe | 1 = lavt præsterende elever<br>2 = medium præsterende elever<br>3 = højt præsterende elever | 100<br>99<br>232 |

***Tabel 1.1:*** Forklaring af data.

Dette datasæt indeholder i alt 431 elever, hver med op til fem testresultater (CBM-score), hvis tider definerer de fire perioder, hvormed vi har i alt 2155 observationer, hvoraf der er observeret 1912 CBM-scorer. Det gælder yderligere, at $41,5\%$ af eleverne er udeblevet fra minimum én af testene, samt at fem elever er udeblevet fra fire ud af de fem tests. En forklaring af data kan ses i Tabel 1.1.

I artiklen, Differences in High- and Low-Performing Students' Fraction Learning in the Fourth Grade [24], konkluderes det, at de højt præsterende elever i interventionsgruppe 1, som star-

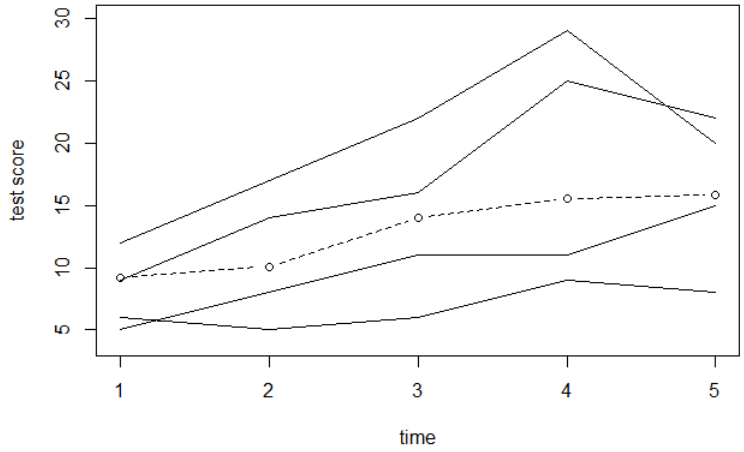
tede på forløbet først, oplevede en signifikant udvikling af deres brøkregningskompetencer i de første tre ud af fire forløbsperioder. Derimod udviklede de lavt præsterende elever i samme interventionsgruppe kun deres brøkregningskompetencer i den periode, hvor de blev undervist i brøkregning. Lignende konkluderes for eleverne i interventionsgruppe 2. Denne opdeling baseret på interventionsgrupperne skyldes, at de to grupper ikke har haft samme undervisningsforløb, hvormed de kan være besværlige at sammenligne. Dette indikeres også af de gennemsnitlige testscorer, inddelt efter de tre grupper over de to forskellige undervisningsforløb, vist på Figur 1.1. Det ene undervisningsforløb introducerer eksempelvis brøkregning i anden periode, hvor det andet undervisningsforløb først introducerer dette i tredje periode. Vi vil derfor omtale data for elever, som har modtaget undervisningsforløb 1 som *datasæt 1*, og data for de som modtog det andet undervisningsforløb som *datasæt 2*.



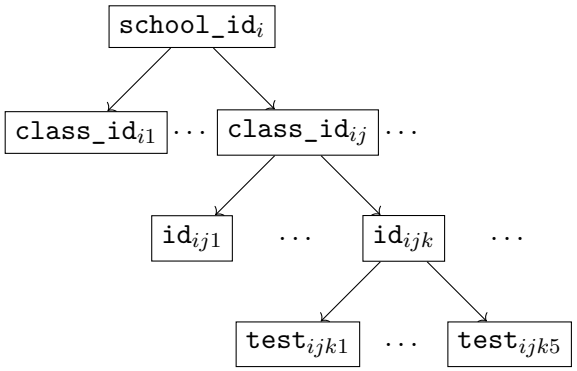

***Figur 1.1:*** Gennemsnitlige testscorer inddelt efter de tre grupper over de to forskellige undervisningsforløb.

Hvis en elev eksempelvis har fået en god testscore ved den første test, er det forventeligt, at den samme elev har større sandsynlighed for igen at få en god testscore ved en efterfølgende test. Dette er illustreret i Figur 1.2, hvor fire tilfældige elevers observerede testscorer er sammenlignet med de gennemsnitlige testscorer for elever med samme undervisningsforløb givet ved den stiplede linje. Dette skaber derfor en korrelation mellem observationer inden for samme elev. Derudover kunne vi også forestille os, at lærernes evne til at undervise varierer eller at skolerne har forskellige trivselsniveauer, som påvirker elevernes evne til at tilegne sig ny viden. Dette kunne derfor være med til yderligere at skabe korrelation mellem observationer inden for samme skole og samme klasse. Dette giver altså data den hierarkiske struktur, som kan ses på træet i Figur 1.3.

Vi skal derfor i projektet benytte *hierarkiske modeller* til at modellere data, herunder modeller der indeholder disse underliggende effekter, som der skal tages højde for. En speciel fremgangsmåde til at analysere data med sådanne effekter, som dem vi betragter igennem projektet, er *variansanalyse (ANOVA)*, som benyttes til at udregne og sammenligne variansestimater imellem og inden for forskellige grupper i data. Disse variansestimater er især relevante, idet de blandt andet kan indikere, hvor spredte de forskellige grupper er.



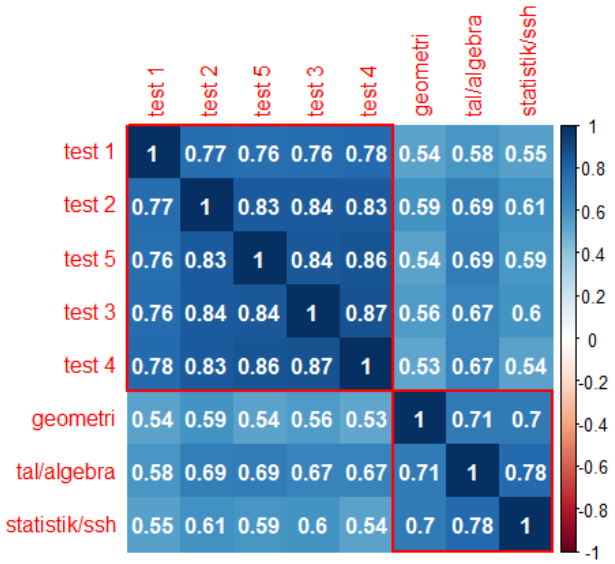

***Figur 1.2:*** Observerede testscorer for fire tilfældige elever, som alle har modtaget undervisningsforløb 1, samt de gennemsnitlige testscorer for eleverne der modtog dette forløb givet ved den stiplede linje.




***Figur 1.3:*** Den hierarkiske inddeling af data.

Variansanalyse benytter eksakte resultater til at bestemme forskellene mellem grupperne, når vi har lige mange observationer i hver af disse. Dette er dog ikke tilfældet for vores data, hvor vi ikke har målinger for alle fem tests for alle eleverne. I sådan et tilfælde kan vi ikke være sikre på, hvor troværdige disse resultater er.

De nationale tests for matematiske kompetencer er opdelt i tre underkategorier; *tal og algebra*, *geometri* og *statistik og sandsynlighed*. Da et supplerende datasæt indeholder testscorerne for disse tre kategorier for eleverne i gruppe 1 og 3, kan vi undersøge korrelationerne mellem disse testscorer og CBM-scorerne for eleverne, der tilhører disse grupper, som kan ses på Figur 1.4. Vi ser her, at de nationale tests er højt positivt korrelerede med hinanden, samt at de fem CBM-scorer ligeledes er indbyrdes højt positivt korrelerede. De nationale tests er moderat positivt korrelerede med CBM-scorerne, og specielt er tal og algebra mest korrelerede med CBM-scoren som forventet. Der er altså grundlag for at kunne inddele eleverne baseret på CBM-scorerne i stedet for testscorerne for tal og algebra, som eleverne allerede er opdelt efter i datasættet.




***Figur 1.4:*** Korrelationer mellem testscorerne for de nationale tests og CBM-scorerne.

Vi bemærker, at denne inddeling i præstationsgrupperne, baseret på de danske nationale tests, muligvis kan være misvisende for nogle elever, idet disse er lange og omhandler mange emner inden for matematik og ikke bare brøkregning. Yderligere kan nogle elever have svært ved at gennemføre de nationale tests i skolen grundet koncentrationsbesvær, eller fordi eleven blot havde en dårlig dag, men have nemmere ved de kortere CBM-tests. Vi vil derfor i dette projekt se nærmere på, hvordan vi kan inddele disse elever i grupper, alt efter hvordan de har klaret sig til disse CBM-tests, samtidig med at vi gerne vil undersøge, hvordan eleverne i de pågældende grupper udvikler deres kompetencer. Dette vil vi gøre ved at betragte en *Growth Mixture model (GMM)*, som er en model, der beskriver tilstedeværelsen af grupper i en population samt individernes udvikling inden for hver af disse, uden at grupperne for hvert individ er observerede. Yderligere vil vi undersøge, om der er overensstemmelse mellem gruppeinddelingen baseret på de nationale tests og gruppeinddeling baseret på en GMM.

Vi vil i projektet benytte en generisk notation for tæthedsfunktioner, således at vi benytter $f$ som en fælles betegnelse for alle tætheder og skal derfor forstås ud fra sammenhængen. Når vi ønsker at evaluere tætheden i et givet punkt, vil vi være nødt til at benytte et logisk udtryk som input som eksempelvis $f(x = 7)$, idet $f(7)$ ikke vil bevare den fornødne information.

# 2 | Manglende data

Som nævnt i tidligere afsnit indeholder datasættet en del manglende værdier, som skyldes at nogle elever er udeblevet fra en eller flere tests. Der mangler i alt 243 CBM-testscorer, hvilket svarer til 11% af de samlede observationer. Vi vil derfor i dette kapitel beskrive, hvordan vi skal håndtere disse manglede værdier.

## 2.1 Typer af manglende data

Manglende data kan kategoriseres i forskellige typer alt efter grunden til, at de ikke forekommer. Lad $y_i = (y_{i1}, \ldots, y_{iT})^T$ være testscorerne for den $i$'te elev, hvormed $y = (y_1^T, \ldots, y_d^T)^T$ er datavektoren for testscorerne. Vi noterer de manglede værdier i $y$ ved $y_m$ og de observerede ved $y_o$. Yderligere lader vi $X$ være matricen med observerede kovariater for hver elev og $R$ en matrix, hvor den $it$'te indgang er 1, hvis $y_{it}$ er observeret og 0 ellers, for $i = 1, \ldots, d$ og $t = 1, \ldots, T$. Hver række i $R$ beskriver altså et mønster for de manglende værdier for hver elev. Vi siger, at data er *missing at random (MAR)* hvis

$$\mathbb{P}(R \mid y_o, y_m, X) = \mathbb{P}(R \mid y_o, X), \tag{2.1}$$

hvormed fordelingen af $R$ kun afhænger af det observerede data [13, s. 796]. Hvis dette ikke er opfyldt, siger vi, at data er *missing not at random (MNAR)*.

Betragt en simultan parametrisk model for $R$ og $y$ givet ved

$$f(R, y; \phi, \theta) = f(R \mid y; \phi) f(y; \theta), \tag{2.2}$$

hvor $(\phi^T, \theta^T)^T \in \Phi \times \Theta$. Bemærk, at det ikke er muligt at observere $R$ og $y$ på samme tid, medmindre alle indgange i $R$ er lig 1, hvorfor vi integrerer $y_m$ ud, sådan at vi opnår den simultane tæthedsfunktion for det observerede data givet ved

$$f(R, y_o; \phi, \theta) = \int f(R, y; \phi, \theta) \mathrm{d}y_m = \int f(R \mid y; \phi) f(y; \theta) \mathrm{d}y_m. \tag{2.3}$$

Da vi kun er interesserede i fordelingen af $y$, ønsker vi blot at maksimere likelihoodfunktionen for $\theta$ baseret på $y$, men ved at betragte ovenstående tæthed bliver vi nødt til også at tage højde for $\phi$. Ved da at antage at data er MAR, er (2.1) opfyldt, hvormed vi får at

$$f(R, y_o; \phi, \theta) = f(R \mid y_o; \phi) \int f(y; \theta) \mathrm{d}y_m = f(R \mid y_o; \phi) f(y_o; \theta). \tag{2.4}$$

Vi kan derfor bestemme et estimat for $\theta$ uden at tage højde for $\phi$ ved blot at maksimere $\log f(y_o; \theta)$. Hvis den observerede datamængde er tilstrækkelig stor, kan vi altså håndtere

manglede data ved blot at fjerne observationerne med de manglende værdier fra analysen [2, s. 62].

Vi kan dog ikke undersøge, om data er MAR, idet vi i så fald skal kende $y_m$, og vi definerer derfor den stærkere antagelse *missing completely at random (MCAR)*, som er givet ved

$$\mathbb{P}(R \mid y_o, y_m, X) = \mathbb{P}(R) \tag{2.5}$$

og som betyder, at $y_o$ eller $X$ heller ikke har indflydelse på mønsteret af de manglede værdier [13, s. 796].

## 2.2 Littles MCAR-test

Vi vil nu beskrive *Littles MCAR-test*, som vi skal benytte til at undersøge, om data er MCAR. Lad $J$ være antallet af entydige mønstre af manglende værdier og lad $o_j$ og $m_j$ være indeksmængderne for henholdsvis de observerede og manglende komponenter i det $j$'te mønster, samt $T_{o_j} = |o_j|$, for $j = 1, \ldots, J$. Antag, at $y_i$ er modelleret ved en $T$-dimensional normalfordeling med middelværdivektor $\mu$ og kovariansmatrix $\Sigma$ og lad $\mu_{o_j}$ og $\Sigma_{o_j}$ svare til den observerede del af disse for hvert entydigt mønster. Lad yderligere $r_i$ være mønsteret for den $i$'te elev og $S_j \subseteq \{1, \ldots, d\}$ være mængden af indekser for individerne, hvor $r_i$ er lig det $j$'te entydige mønster samt $d_j = |S_j|$, hvormed $\sum_{j=1}^{J} d_j = d$. For at teste om data er MCAR, lader vi $y_{i,o}$ være den observerede del af $y_i$ og betragter hypotesen

$$\mathcal{H}_0 : Y_{i,o} \mid r_i \sim N_{T_{o_j}}(\mu_{o_j}, \Sigma_{o_j}), \quad i \in S_j, \quad j = 1, \ldots, J, \tag{2.6}$$

mod den alternative hypotese

$$\mathcal{H}_1 : Y_{i,o} \mid r_i \sim N_{T_{o_j}}(\nu_j, \Sigma_{o_j}), \quad i \in S_j, \quad j = 1, \ldots, J, \tag{2.7}$$

hvor $\nu_j$ er en $T_{o_j}$-dimensional middelværdivektor, som kan være distinkt for $j = 1, \ldots, J$. Det betyder altså, at under den alternative hypotese, varierer middelværdien af $y_{i,o}$ i forhold til, hvilken $S_j$ individet tilhører. Det gælder derfor, at forkastelse af (2.6) er tilstrækkeligt for at forkaste MCAR-antagelsen.

Lad nu $\bar{y}_{o_j}$ være gennemsnittet af $y_{i,o}$ over alle $i \in S_j$. Da kan vi teste nulhypotesen med teststørrelsen

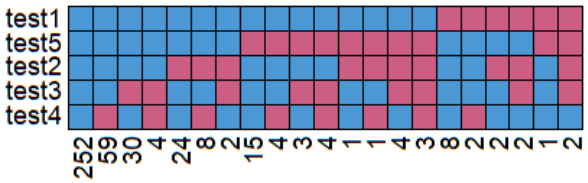
$$\delta_0^2 = \sum_{j=1}^{J} d_j (\bar{Y}_{o_j} - \mu_{o_j})^T \Sigma_{o_j}^{-1} (\bar{Y}_{o_j} - \mu_{o_j}), \tag{2.8}$$

som under $\mathcal{H}_0$ asymptotisk følger en $\chi^2$-fordeling med $\sum_{j=1}^{J} T_{o_j} - T$ frihedsgrader [14, s. 1200]. Bemærk, at man kan undlade antagelsen om normalitet ved tilpas stor datamængde, idet det af den centrale grænseværdisætning gælder, at $\bar{Y}_{o_j}$ går mod en normalfordeling, når $d$ går mod uendelig. Da vi i praksis ikke kender $\mu$ og $\Sigma$, erstattes disse blot med $\hat{\mu}$ og $\widetilde{\Sigma} = n\hat{\Sigma}/(n-1)$, hvor $\hat{\mu}$ og $\hat{\Sigma}$ er MLE under nulhypotesen. Dermed benytter vi teststørrelsen

$$\delta^2 = \sum_{j=1}^{J} d_j (\bar{Y}_{o_j} - \hat{\mu}_{o_j})^T \widetilde{\Sigma}_{o_j}^{-1} (\bar{Y}_{o_j} - \hat{\mu}_{o_j}). \tag{2.9}$$

## 2.3 Analyse af manglende data

For at undersøge om vores data er MCAR, undersøger vi først strukturen af $R$, og en illustration af disse mønstre ses i Figur 2.1. Vi ser eksempelvis, at $J = 21$ og yderligere observerer vi, at der er $d_1 = 252$ elever med alle testscorer observeret, altså $r_1 = (1, 1, 1, 1, 1)$. Derudover er der $d_2 = 59$ elever, der kun mangler den fjerde test, altså $r_2 = (1, 1, 1, 0, 1)$. Vi bruger nu




***Figur 2.1:*** Mønstre af manglende værdier af CBM-scoren, hvor et rødt felt illustrerer en manglende værdi til den pågældende test for et antal elever.

Littles MCAR-test for at undersøge, om de manglende værdier opfylder MCAR-antagelsen. Dette gør vi ved R-funktionen `mcar_test` fra pakken `naniar` [30], som giver $\delta^2 = 66$ med 57 frihedsgrader, hvormed vi opnår en $p$-værdi på 0.193, og vi ikke kan forkaste antagelsen om MCAR. Vi kan derfor blot fjerne de manglede værdier fra analysen.

# 3 | Modeller med stokastiske effekter

Under regression, ved eksempelvis at benytte normale eller generaliserede lineære modeller, ønsker vi ofte at estimere den ukendte parametervektor $\beta$. Parametrene i denne vektor kaldes *systematiske effekter*, hvorfor en sådan model hører under kategorien *systematisk effekt-modeller* [16, s. 7]. Betragt i stedet tilfældet, hvor vi ikke er interesserede i de enkelte systematiske effekter, men derimod en underliggende effekt som tillader variation mellem grupper af data. Dette kunne eksempelvis være hver elevs effekt på testscoren, altså en eleveffekt. I så fald kan en simpel model, som indeholder denne tilfældige effekt samt den fælles middelværdi $\mu \in \mathbb{R}$, være givet ved

$$Y_{ij} = \mu + U_i + \varepsilon_{ij}, \quad i = 1, \dots, d, \quad j = 1, \dots, n_i, \tag{3.1}$$

hvor $Y_{ij}$ er variablen for den $j$'te CBM-score til den $i$'te elev, og hvor $n_i$ er antallet af observerede CBM-scorer for denne elev. Yderligere er $U_i \sim N(0, \sigma_u^2)$ den stokastiske effekt på CBM-scoren, det har at være den $i$'te elev, som er uafhængig af $\varepsilon_{ij} \sim N(0, \sigma^2)$. En sådan model kaldes for en *stokastisk effekt-model*, som hører under kategorien *hierakiske modeller*, idet både $Y_{ij}$ og $U_i$ er stokastiske, og $U_i$ indgår i modelspecifikationen for $Y_{ij}$ [16, s. 162-163]. Bemærk, at det for to forskellige CBM-scorer for samme elev gælder, at

$$\begin{aligned} \mathrm{Cov}(Y_{ij}, Y_{ij'}) &= \mathrm{Cov}(U_i + \varepsilon_{ij}, U_i + \varepsilon_{ij'}) \\ &= \mathrm{Cov}(U_i, U_i) + \mathrm{Cov}(U_i, \varepsilon_{ij'}) + \mathrm{Cov}(\varepsilon_{ij}, U_i) + \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) \\ &= \mathrm{Var}(U_i) = \sigma_u^2, \quad j \neq j', \end{aligned} \tag{3.2}$$

hvormed

$$\mathrm{Corr}(Y_{ij}, Y_{ij'}) = \frac{\mathrm{Var}(U_i)}{\sqrt{\mathrm{Var}(Y_{ij})\mathrm{Var}(Y_{ij'})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}. \tag{3.3}$$

Altså gælder det, at observationer af CBM-scoren for den samme elev er korrelerede. Vi ser også, at jo større elevvarians der er, des større er denne korrelation. Stokastisk effekt-modeller kan derfor tage højde for elevernes individuelle heterogenitet, som ellers ikke er muligt i almindelige systematisk effekt-modeller. Betragt eksempelvis tilfældet, hvor vi ignorerer variationen mellem eleverne. Da vil vi få, at $Y_{ij} = \mu + \varepsilon_{ij}$, hvormed $\varepsilon_{ij} \sim N(0, \sigma_u^2 + \sigma^2)$. Hvis vi har estimeret middelværdien $\mu$ ved gennemsnittet af testscorerne, vil det da gælde, at

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n_i}\varepsilon_{ij}\right) = \frac{\sigma_u^2 + \sigma^2}{n}, \tag{3.4}$$

hvor $n = \sum_{i=1}^{d} n_i$ er antallet af observationer. Denne varians er mindre end i tilfældet, hvor vi har inkluderet variationen mellem eleverne, hvor vi ville få

$$\begin{aligned} \mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n_i}(U_i + \varepsilon_{ij})\right) = \frac{1}{n^2}\left(\sum_{i=1}^{d} n_i^2 \mathrm{Var}(U_i) + \sum_{i=1}^{d}\sum_{j=1}^{n_i}\mathrm{Var}(\varepsilon_{ij})\right) \\ &= \frac{1}{n^2}\left(\sigma_u^2 \sum_{i=1}^{d} n_i^2 + n\sigma^2\right) = \frac{\sigma_u^2 \sum_{i=1}^{d} n_i^2/n + \sigma^2}{n}, \end{aligned} \tag{3.5}$$

idet $\sum_{i=1}^{d} n_i^2/n \geq 1$. Bemærk, at den stokastiske effekt svarer til effekten af en variabel som inddeler observationerne i elevspecifikke grupper. Sådan en variabel kaldes for en *faktor*, $F$ defineret ved en surjektiv funktion

$$\varphi_F : \mathcal{I} \to \mathcal{F}, \tag{3.6}$$

hvor $\mathcal{F}$ er en endelig mængde, og $\mathcal{I}$ er indeksmængden for alle observationerne [31, s. 36]. Det vil sige, at faktoren $F$ inddeler observationerne i $|\mathcal{F}| = d_F$ grupper. Ligeledes kan vi definere *enhedsfaktoren*, $I$, ved $\varphi_I : \mathcal{I} \to \mathcal{I}$, som tildeler hver observation til en entydig gruppe. Tilsvarende kan vi også definere faktoren 0, som tildeler alle observationer til den samme gruppe.

Hvis vi yderligere ønsker at inkludere effekten af forskellige kovariater, såsom køn og præstationsgruppe, og samtidig tage højde for variationen mellem eleverne, kan vi specificere en såkaldt *mixed model*, som består af både stokastiske og systematiske effekter [16, s. 158]. Betragt derfor

$$Y_{ij} = \beta_0 + x_{ij}^T \beta_1 + U_i + \varepsilon_{ij}, \quad i = 1, \ldots, d, \quad j = 1, \ldots, n_i, \tag{3.7}$$

hvor $x_{ij}$ er en vektor af kovariater. Denne model kaldes lineær, idet $Y_{ij}$ givet $U_i = u_i$ har middelværdi $\mu_i = \beta_0 + x_{ij}\beta_1^T + u_i$. Vi kan da specificere denne lineære mixed model ved

$$Y = X\beta + ZU + \varepsilon, \tag{3.8}$$

hvor $X \in \mathbb{R}^{n \times p}$ og $Z \in \mathbb{R}^{n \times d}$ er passende designmatricer. Designmatricen $Z$ består af rækkerne $z_{ij}$, hvor den $q$'te indgang er 1, hvis $i = q$ og 0 ellers for $i = 1, \ldots, d$ og $j = 1, \ldots, n_i$. Ovenstående kan udvides til en model med $K$ stokastiske effekter, hvor vi for nemhedes skyld antager, at disse er uafhængige. Dette er ikke nødvendigvis en korrekt antagelse i praksis, men hvis to stokastiske effekter er afhængige, er det oftest nemmest at betragte dem som én stokastisk effekt [23, s. 452]. Da kan modellen i (3.8) betragtes som en generel specifikation for en mixed model, hvor $Z = [Z_1 \cdots Z_K]$ er designmatricen for den stokastiske effekt $U = (U_1^T, \ldots, U_K^T)^T$, hvor $U_k \sim N_{d_k}(0, R_k)$ og $\varepsilon \sim N_n(0, \Sigma)$ er uafhængige for $k = 1, \ldots, K$ [16, s. 179]. Det gælder altså, at

$$Y \sim N_n(X\beta, ZRZ^T + \Sigma), \tag{3.9}$$

hvor $R$ er en blokdiagonal matrix, som består af blokkene $R_k$ for $k = 1, \ldots, K$, og vi vil fremover kalde kovariansmatricen i ovenstående ligning for $V$. Fra ovenstående gælder det altså,

at $(U^T, \varepsilon^T)^T$ følger en multivariat normalfordeling med middelværdi 0 og en kovariansmatrix, som består af henholdvis $R$ og $\Sigma$ som blokdiagonaler. Dette medfører, at

$$\begin{bmatrix} Y \\ U \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \end{bmatrix} + \begin{bmatrix} Z & I_n \\ I_D & 0 \end{bmatrix} \begin{bmatrix} U \\ \varepsilon \end{bmatrix}, \tag{3.10}$$

hvor $D = \sum_{k=1}^{K} d_k$, følger en normalfordeling med kovariansmatrix

$$\begin{bmatrix} Z & I_n \\ I_D & 0 \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & \Sigma \end{bmatrix} \begin{bmatrix} Z^T & I_D \\ I_n & 0 \end{bmatrix} = \begin{bmatrix} V & ZR \\ RZ^T & R \end{bmatrix}. \tag{3.11}$$

## 3.1 Parameterestimation

Det gælder ofte, at kovariansmatricerne for $U$ og $\varepsilon$ afhænger af $q$ ukendte parametre, $\psi = (\psi_1, \ldots, \psi_q)^T$. Derfor skal vi udlede profil-likelihoodfunktionen for $\psi$, for at kunne estimere parameterne i modellen i (3.8). Ved at betegne kovariansmatricen for $Y$ ved $V(\psi)$, kan vi opskrive log-likelihoodfunktionen baseret på (3.9) som

$$\ell(\beta, \psi; y) = -\frac{1}{2} \log |V(\psi)| - \frac{1}{2} (y - X\beta)^T V(\psi)^{-1} (y - X\beta). \tag{3.12}$$

Ved at maksimere denne i forhold til $\beta$ får vi

$$\hat{\beta}(\psi) = (X^T V(\psi)^{-1} X)^{-1} X^T V(\psi)^{-1} y, \tag{3.13}$$

under antagelsen af, at $(X^T V(\psi)^{-1} X)^{-1}$ eksisterer, hvilket vi genkender som GLS-estimatet. Denne estimator har følgende kovariansmatrix

$$\begin{aligned} \mathrm{Var}(\hat{\beta}(\psi)) &= (X^T V(\psi)^{-1} X)^{-1} X^T V(\psi)^{-1} \mathrm{Var}(Y) (V(\psi)^{-1})^T X ((X^T V(\psi)^{-1} X)^{-1})^T \\ &= (X^T V(\psi)^{-1} X)^{-1} (X^T V(\psi)^{-1} X)^T ((X^T V(\psi)^{-1} X)^{-1})^T \\ &= (X^T V(\psi)^{-1} X)^{-1}. \end{aligned} \tag{3.14}$$

Da kan vi opskrive profil-likelihoodfunktionen for $\psi$ ved at indsætte $\hat{\beta}(\psi)$ i (3.12), hvormed vi får

$$\ell(\psi; y) = -\frac{1}{2} \log |V(\psi)| - \frac{1}{2} \left(y - X\hat{\beta}(\psi)\right)^T V(\psi)^{-1} \left(y - X\hat{\beta}(\psi)\right). \tag{3.15}$$

Vi kan nu bestemme maksimum likelihood-estimatet $\hat{\psi}$ ved at maksimere ovenstående log-likelihoodfunktion i forhold til det $r$'te komponent i $\psi$, hvormed $\hat{\psi}$ er givet ved en løsning til ligningssystemet

$$\mathrm{tr}\left(V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_r}\right) = \left(y - X\hat{\beta}(\psi)\right)^T V(\psi)^{-1} \frac{\partial V(\psi)}{\partial \psi_r} V(\psi)^{-1} \left(y - X\hat{\beta}(\psi)\right), \tag{3.16}$$

for $r = 1, \ldots, q$, hvor $\mathrm{tr}(\cdot)$ angiver sporet af en matrix [10, s. 11]. Ved at indsætte estimatet for $\psi$ kan vi estimere den systematiske effekt ved

$$\hat{\beta} = (X^T V(\hat{\psi})^{-1} X)^{-1} X^T V(\hat{\psi})^{-1} y. \tag{3.17}$$

### 3.1.1 REML-estimation

Vi vil eksempelvis for en normal lineær model opleve, at estimatet for $\psi$, der løser (3.16), ikke er middelværdiret. Vi vil nu beskrive et konkret eksempel, hvor dette også er tilfældet.

**Eksempel 3.1** *(Neyman-Scott-problemet)*
Antag at $i = 1, \ldots, d$ og $j = 1, 2$, hvormed $n = 2d$ samt at $Y_{ij} \sim N(\mu_i, \sigma^2)$, sådan at hver gruppe har forskellige middelværdier men samme varians, hvormed vi har $d + 1$ ukendte parametre. Da kan vi opskrive følgende model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{3.18}$$

hvor $\varepsilon_{ij} \sim N(0, \sigma^2)$. I så fald vil MLE for $\mu_i$ og $\sigma^2$ være givet ved henholdsvis $\bar{y}_{i\cdot} = (y_{i1}+y_{i2})/2$ og

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{2}(y_{ij} - \bar{y}_{i\cdot})^2 \\
&= \frac{1}{2d}\sum_{i=1}^{d}\left(\frac{y_{i1} - y_{i2}}{2}\right)^2 + \left(\frac{y_{i2} - y_{i1}}{2}\right)^2 \\
&= \frac{1}{4d}\sum_{i=1}^{d}(y_{i1} - y_{i2})^2,
\end{aligned} \tag{3.19}$$

hvor det gælder, at $\sum_{i=1}^{d}(Y_{i1} - Y_{i2})^2 \sim 2\sigma^2\chi^2(d)$. Dermed er $\hat{\sigma}^2$ hverken middelværdiret eller konsistent, idet

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{4d}2\sigma^2 d = \frac{1}{2}\sigma^2. \tag{3.20}$$

Yderligere er variansen givet ved

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{(2\sigma^2)^2}{(4d)^2}2d = \frac{1}{2d}\sigma^4. \tag{3.21}$$

Denne bias opstår, da vi har benyttet MLE for $\mu_i$, hvilket vi ser, da $\sum_{i=1}^{d}\sum_{j=1}^{2}(Y_{ij} - \mu_i)^2 \sim \sigma^2\chi^2(n)$, hvormed

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{2}(Y_{ij} - \mu_i)^2\right] = \sigma^2. \tag{3.22}$$

Vi får dermed, at MSE af $\hat{\sigma}^2$ er

$$\begin{aligned}
\mathrm{MSE} &= \mathbb{E}[(\hat{\sigma}^2 - \sigma^2)^2] = \mathbb{E}[\hat{\sigma}^4 + \sigma^4 - 2\hat{\sigma}^2\sigma^2] \\
&= \frac{1}{2d}\sigma^4 + \frac{1}{4}\sigma^4 + \sigma^4 - \sigma^4 = \frac{1}{2d}\sigma^4 + \frac{1}{4}\sigma^4.
\end{aligned} \tag{3.23}$$

◀

Jævnfør eksemplet er det fordelagtigt at estimere $\psi$ uden først at estimere middelværdien, hvorfor vi transformerer data, så middelværdien fjernes. Vi betragter derfor en $n \times (n-p)$-matrix, $A$, hvis søjler udspænder det ortogonale komplement af span$(X)$, hvormed $A^T X = 0$. Vi får dermed, at

$$\widetilde{y} = A^T y = A^T X\beta + A^T ZU + A^T \varepsilon = A^T ZU + A^T \varepsilon, \tag{3.24}$$

hvorfor det gælder, at $\widetilde{Y}$ er normalfordelt med middelværdi 0 og kovariansmatrix $\widetilde{V}(\psi) = A^T V(\psi) A$, hvormed log-likelihoodfunktionen er givet ved

$$\ell(\psi; y) = -\frac{1}{2}\log|\widetilde{V}(\psi)| - \frac{1}{2}\widetilde{y}^T \widetilde{V}(\psi)^{-1}\widetilde{y}. \tag{3.25}$$

Vi kan da udlede maksimum likelihood-estimatet for $\psi$ baseret på fordelingen af $\widetilde{Y}$, hvor $\beta$ ikke indgår, hvormed vi har mere tiltro til, at estimatet er middelværdiret. Dette estimat kalder vi for *restricted/residual maximum likelihood-estimatet (REML-estimatet)* [10, s. 14] og er, jævnfør (3.16), opnået ved at løse ligningssystemet

$$\operatorname{tr}\left(A\widetilde{V}(\psi)^{-1}A^T \frac{\partial V(\psi)}{\partial \psi_r}\right) = y^T A\widetilde{V}(\psi)^{-1}A^T \frac{\partial V(\psi)}{\partial \psi_r} A\widetilde{V}(\psi)^{-1}A^T y, \tag{3.26}$$

for $r = 1, \ldots, q$. Bemærk, at REML-estimatet ikke afhænger af transformationen $A$. Hvis vi eksempelvis erstatter $A$ med en anden matrix $B = AC$, hvor $C$ er en invertibel matrix, så gælder det, at

$$B(B^T V(\psi) B)^{-1} B^T = AC(C^T A^T V(\psi) AC)^{-1} C^T A^T = A(A^T V(\psi) A)^{-1} A^T, \tag{3.27}$$

hvormed log-likelihoodfunktion for $\psi$ baseret på fordelingen af $BY^T$ er

$$\begin{aligned} &-\frac{1}{2}\log|B^T V(\psi) B| - \frac{1}{2} y^T B(B^T V(\psi) B)^{-1} B^T y \\ &= -\frac{1}{2}\log|C^T C| - \frac{1}{2}\log|A^T V(\psi) A| - \frac{1}{2} y^T A(A^T V(\psi) A)^{-1} A^T y \\ &\propto -\frac{1}{2}\log|\widetilde{V}(\psi)| - \frac{1}{2}\widetilde{y}^T \widetilde{V}(\psi)^{-1}\widetilde{y}. \end{aligned} \tag{3.28}$$

Dermed opnår vi de samme estimater, hvilket er fordelagtigt, idet $A$ ikke er entydigt valgt.

**Eksempel 3.1** *(fortsat)*
Betragter vi igen Neyman-Scott-problemet, så vil det gælde, at

$$\widetilde{Y}_i = Y_{i1} - Y_{i2} \sim N(0, 2\sigma^2), \tag{3.29}$$

hvilket svarer til transformationen $\widetilde{Y} = A^T Y$, hvor $A$ er en $n \times d$-matrix givet ved

$$A^T = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -1 \end{bmatrix}. \tag{3.30}$$

Det gælder heraf, at $A^TA = 2I_d$, samt at $AA^T$ er en $n \times n$ blok-diagonal matrix bestående af $2 \times 2$-dimensionale blokke med 1 på diagonalen og $-1$ på de resterende indgange. Ved at løse (3.26) for $\sigma^2$ får vi at

$$\begin{aligned} \operatorname{tr}\left(\frac{1}{2\sigma^2}AA^T\right) &= \frac{1}{4\sigma^4}y^TAA^TAA^Ty \Leftrightarrow \\ \frac{2d}{2\sigma^2} &= \frac{1}{2\sigma^4}y^TAA^Ty \Leftrightarrow \\ \hat{\sigma}^2 &= \frac{1}{2d}\sum_{i=1}^{d}\widetilde{y}_i^2. \end{aligned} \tag{3.31}$$

Da det gælder, at $\sum_{i=1}^{d}\widetilde{Y}_i^2 \sim 2\sigma^2\chi^2(d)$, får vi at

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{2d}2\sigma^2 d = \sigma^2, \tag{3.32}$$

samt at

$$\operatorname{Var}(\hat{\sigma}^2) = \frac{1}{4d^2}(2\sigma^2)^2 2d = \frac{2}{d}\sigma^4, \tag{3.33}$$

hvormed REML-estimatoren altså er middelværdiret og konsistent, modsat maksimum likelihood-estimatoren. Vi får derfor, at MSE af REML-estimatoren er givet ved $\operatorname{Var}(\hat{\sigma}^2)$. Det gælder da, at

$$\frac{1}{2d}\sigma^4 + \frac{1}{4}\sigma^4 > \frac{2}{d}\sigma^4 \Leftrightarrow \frac{1}{2d} + \frac{1}{4} > \frac{2}{d} \Leftrightarrow 2 + d > 8 \Leftrightarrow d > 6, \tag{3.34}$$

hvormed MSE af ML-estimatoren er større end MSE af REML-estimatoren for $d > 6$. ◀

Modsat hvad vi ser i ovenstående eksempel, har vi ofte størst tiltro til at MSE for ML-estimatoren er mindre end for REML. Hvis vi dog vælger at bruge REML-estimation, opnår vi en estimator, som vi har større tiltro til er middelværdiret, end vi har til ML-estimatoren. Derfor bruger vi REML-estimatet som standardmetode, medmindre nøjagtigheden af estimatet er vigtigere end et middelværdiret estimat [16, s. 182].

## 3.2 Modeller med krydsfaktorer

Vi siger, at en faktor $F$ er *finere* end en faktor $G$, eller at $G$ er *grovere* end $F$, hvis der findes en funktion $\varphi_{FG} : \mathcal{F} \to \mathcal{G}$, sådan at

$$\varphi_{FG} \circ \varphi_F = \varphi_G \quad [31, \text{ s. } 36]. \tag{3.35}$$

Det betyder altså, at grupper af $G$ kan opnås ved at sammensætte grupper af $F$, og vi noterer dette med $G \preceq F$. Dette kan illustreres ved *strukturdiagrammet*

$$I \to F \to G \to 0. \tag{3.36}$$

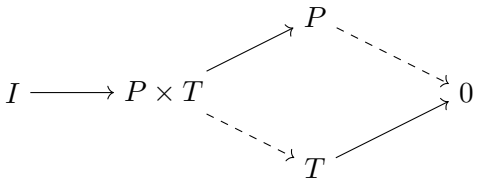
Antag nu, at vi har to faktorer, $P$ og $T$, hvor grupperne i hver faktor ikke kan opnås ved sammensætninger af den anden. Da definerer vi *krydsfaktoren* $P \times T$ ved

$$\varphi_{P\times T} : \mathcal{I} \to \mathcal{P} \times \mathcal{T}, \tag{3.37}$$

hvor $\times$ er det kartesiske produkt [31, s. 37]. Det gælder, at $P \preceq P \times T$, samt at krydsfaktoren inddeler observationerne i $d_P d_T$ grupper. Betragt nu en mixed model med stokastiske effekter $U_P$ og $U_{P\times T}$ givet ved

$$Y = \mathbf{1}_n \beta_0 + X_T \beta_T + Z_P U_P + Z_{P\times T} U_{P\times T} + \varepsilon, \tag{3.38}$$

hvor $U_P \sim N_{d_P}(0, R_P)$, $U_{P\times T} \sim N_{d_P d_T}(0, R_{P\times T})$ og $\varepsilon \sim N_n(0, \Sigma)$. Vi kalder en faktor for systematisk, hvis effekten af denne, i modellen, er systematisk og tilsvarende for en stokastisk faktor. Bemærk, at krydsfaktoren $P \times T$ ikke kan være systematisk, idet $P$ er stokastisk. Det gælder, at en systematisk faktor $G$ tilhører $F$'s *stratum*, hvis $F$ er den groveste stokastiske faktor, som er finere end $G$. Vi kan derfor betragte følgende strukturdiagram for modellen, hvor en stiplet pil indikerer, at faktorerne tilhører samme stratum.




Altså gælder det, at 0 tilhører $P$'s stratum, $T$ tilhører $P \times T$'s stratum og $I$ er den eneste faktor i sit stratum. I tilfældet hvor både $P$ og $T$ er stokastiske, og 0 er systematisk, kan vi ikke bestemme et entydigt stratum, som 0 tilhører, hvorfor både $P$ og $T$ ikke begge kan være stokastiske, medmindre 0 er det.

En model med flere faktorer skal altid respektere det *hierarkiske princip*, som betyder, at hvis den indeholder krydsfaktoren $F \times G$, skal faktorerne $F$ og $G$ også være indeholdt i denne model [19]. Dette skyldes, at vi ikke vil have en entydig fortolkning af effekterne opnået ved brug af forskellige parametriseringer af designmatricen, når sådanne vekselvirkninger også indgår i modellen. Vi bør derfor ikke teste for, om vi kan fjerne hverken $F$ eller $G$, hvis krydsfaktoren indgår.

## 3.3 Prædiktion af stokastiske effekter

Den stokastiske effekt $U$ er ikke en parameter i modellen og blev derfor ikke estimeret i Afsnit 3.1. Dog gælder det, at for at lave prædiktioner ved brug af en mixed model, skal vi bruge et estimat af både de stokastiske og de systematiske effekter, som vi indsætter i (3.8) på følgende måde

$$\hat{Y} = X\hat{\beta} + Z\hat{U}. \tag{3.39}$$

For at udlede en estimator for $U$ bemærker vi, at den simultane tæthed for $(Y^T, U^T)^T$ er givet ved

$$f(y, u; \beta, \psi) = f(y \mid u; \beta) f(u; \psi), \tag{3.40}$$

og da $Y \mid U = u$ og $U$ begge er normalfordelte, er log-likelihoodfunktionen givet ved

$$\begin{aligned} \ell(\beta, \psi; y, u) = &-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta - Zu)^T \Sigma^{-1}(y - X\beta - Zu) \\ &- \frac{1}{2}\log|R| - \frac{1}{2}u^T R^{-1} u. \end{aligned} \tag{3.41}$$

Ved at differentiere denne i forhold til $u$ og sætte lig nul får vi estimatet

$$\begin{aligned}\hat{u} &= (Z^T\Sigma^{-1}Z + R^{-1})^{-1}Z^T\Sigma^{-1}(y - X\beta) \\ &= RZ^TV^{-1}(y - X\beta),\end{aligned} \tag{3.42}$$

givet at $(Z^T\Sigma^{-1}Z+R^{-1})^{-1}$ eksisterer, hvor sidste lighed kommer af Woodburys identitet [21, s. 614]. I praksis vil vi i ovenstående erstatte de ukendte parametre med estimater heraf. Det er klart, at $\hat{U}$ er lineær og middelværdiret, idet (3.42) er en lineær funktion af $y$, samt at

$$\mathbb{E}[U - \hat{U}] = -RZ^TV^{-1}(\mathbb{E}(Y) - X\beta) = 0. \tag{3.43}$$

Estimatoren $\hat{U}$ kaldes *best linear unbiased predictor (BLUP)* [23, s. 441], idet der for enhver anden lineær middelværdiret prædiktor $\widetilde{U}$ gælder, at $\mathrm{Var}(U - \widetilde{U}) - \mathrm{Var}(U - \hat{U})$ er positiv semi-definit. Dette ser vi, da der for en vilkårlig lineær funktion $\alpha + CY$ gælder, at

$$\begin{aligned}\mathrm{Cov}(U - \hat{U}, \alpha + CY) &= (\mathrm{Cov}(U, Y) - RZ^TV^{-1}\mathrm{Cov}(Y, Y))C^T \\ &= (RZ^T - RZ^TV^{-1}V)C^T = 0,\end{aligned} \tag{3.44}$$

jævnfør (3.11), hvormed vi får, at

$$\begin{aligned}\mathrm{Var}(U - \widetilde{U}) &= \mathrm{Var}(U - \hat{U} + \hat{U} - \widetilde{U}) \\ &= \mathrm{Var}(U - \hat{U}) + \mathrm{Var}(\hat{U} - \widetilde{U}) + 2\mathrm{Cov}(U - \hat{U}, \hat{U} - \widetilde{U}) \\ &= \mathrm{Var}(U - \hat{U}) + \mathrm{Var}(\hat{U} - \widetilde{U}),\end{aligned} \tag{3.45}$$

idet $\hat{U} - \widetilde{U}$ også er en lineær funktion af $Y$. Det gælder altså, at $\mathrm{Var}(U - \widetilde{U}) - \mathrm{Var}(U - \hat{U}) = \mathrm{Var}(\hat{U} - \widetilde{U})$, som er positiv semi-definit, hvormed $\hat{U}$ er den bedste prædiktor.

Yderligere gælder det jævnfør (3.44), at $\mathrm{Cov}(U - \hat{U}, \hat{U}) = 0$, hvilket medfører, at

$$\begin{aligned}\mathrm{Var}(\hat{U}) &= \mathrm{Cov}(\hat{U}, \hat{U}) + \mathrm{Cov}(U - \hat{U}, \hat{U}) \\ &= \mathrm{Cov}(U, \hat{U}) = ZRV^{-1}RZ^T.\end{aligned} \tag{3.46}$$

Vi definerer nu $W = U - \hat{U}$, således at $U = \hat{U} + W$. Da gælder det, at $W$ er normalfordelt med middelværdi 0 og

$$\mathrm{Var}(W) = \mathrm{Var}(U) + \mathrm{Var}(\hat{U}) - 2\mathrm{Cov}(U, \hat{U}) = R - ZRV^{-1}RZ^T, \tag{3.47}$$

samt $\mathrm{Cov}(W, Y) = 0$, hvormed $W$ og $Y$ er uafhængige. Givet $Y = y$, så er $\hat{U} = \hat{u}$ en konstant, hvormed det følger, at

$$U \mid Y = y \sim N_d(\hat{u}, R - ZRV^{-1}RZ^T), \tag{3.48}$$

hvilket viser, at $\mathbb{E}[U \mid Y = y] = \hat{u}$. BLUP er altså en optimal prædiktor blandt alle lineære middelværdirette estimatorer, da den minimerer den kvadrerede prædiktionsfejl [16, s. 183].

## 3.4 F-test

For at undersøge om vi kan reducere middelværdistrukturen i en mixed model, vil det være oplagt at benytte en F-test. Mere generelt kunne vi teste for nulhypotesen

$$\mathcal{H}_0: \quad K\beta = b, \tag{3.49}$$

hvor $K$ er en passende $f \times p$-matrix og $b \in \mathbb{R}^f$. Eksakte F-tests er ofte ikke mulige i praksis, hvorfor man i stedet benytter asymptotiske $\chi^2$-tests, såsom den følgende, baseret på Walds teststørrelse. Denne er givet ved

$$F = \frac{1}{f}(K\hat{\beta} - b)^T(K(X^T V(\hat{\psi})^{-1}X)^{-1}K^T)^{-1}(K\hat{\beta} - b), \tag{3.50}$$

hvor $\hat{\psi}$ er REML-estimatet, og $\hat{\beta}$ er givet som i (3.17) [8, s. 10]. Det gælder, at $\hat{\beta}$ er asymptotisk normalfordelt med middelværdi $\beta$ og kovariansmatrix $(X^T V(\hat{\psi})^{-1}X)^{-1}$, hvormed det under nulhypotesen gælder, at teststørrelsen $fF$ asymptotisk følger en $\chi^2(f)$-fordeling. Vi ved dog ikke, hvor stort datasættet skal være, for at denne approksimation er god. Kenward og Roger foreslog i stedet en modificeret teststørrelse, med en approksimeret $F(f, m)$-fordeling, for et $m > 0$ [8, s. 10]. Det kan vises, at

$$\mathrm{Var}(\hat{\beta}) = (X^T V^{-1} X)^{-1} + A \tag{3.51}$$

hvor $A$ svarer til den bias, som den asymptotiske kovariansmatrix underestimerer $\mathrm{Var}(\hat{\beta})$ [11, s. 985]. Så i stedet for at benytte $(X^T V(\hat{\psi})^{-1}X)^{-1}$ som estimat for $(X^T V^{-1}X)^{-1}$ i (3.50), benytter vi et estimat for hele udtrykket i (3.51), hvilket vil forbedre den asymptotiske fordeling af F-teststørrelsen.

Yderligere definerer vi $\lambda \geq 0$, sådan at en approksimation af den forventede værdi og variansen af $\lambda F$, svarer til henholdsvis middelværdien og variansen af en $F(f, m)$-fordelt variabel. Det vil sige, at $\lambda$ og $m$ kan bestemmes ved at løse

$$\lambda \mathrm{E}^* = \frac{m}{f-2}, \quad \lambda^2 \mathrm{V}^* = \frac{2m^2(f+m-2)}{f(m-2)^2(m-4)}, \tag{3.52}$$

hvor $\mathrm{E}^*$ og $\mathrm{V}^*$ er approksimationer af henholdsvis $\mathbb{E}[F]$ og $\mathrm{Var}(F)$, baseret på en første ordens Taylor-udvidelse af $F$ [8, s. 31]. Dermed vil vi opnå

$$\lambda = \frac{m}{\mathrm{E}^*(m-2)}, \quad m = 4 + \frac{(f+2)2(\mathrm{E}^*)^2}{f\mathrm{V}^* - 2(\mathrm{E}^*)^2}, \tag{3.53}$$

hvormed $\lambda F$ asymptotisk følger en $F(f, m)$-fordeling, og vi kan derfor lave F-test, baseret på denne teststørrelse [11, s. 986-988].

## 3.5 Dataanalyse ved brug af mixed modeller

Vi vil i dette afsnit tilpasse en mixed model på begge datasæt og beskrive de forskellige præstationsgruppers udvikling over tid. I denne sammenhæng vil vi betragte parameterestimater, estimater af kurvernes hældninger samt lave forskellige hypotesetests for disse. Yderligere vil vi undersøge sammenhængen mellem prædiktioner af eleveffekterne og gruppeinddelingen baseret på de nationale tests. R-koden til dette afsnit kan findes i Appendiks B.1.

### 3.5.1 Mixed model med gruppefaktor

For at tilpasse en mixed model skal vi overveje, hvilke variable der er relevante for analysen, og hvilke effekter der skal være henholdsvis stokastiske eller systematiske. Effekterne af `group` og `time` samt deres vekselvirkning `group:time`, skal indgå i modellen som systematiske, da vi ønsker at undersøge de forskellige gruppers udviklinger over tid. Vi vurderer, at `age` ikke er nødvendig, da alle indvider er fjerdeklasseselever. Yderligere tilføjer vi også effekten af `gender` som en systematisk effekt. Som stokastiske effekter er det oplagt at vælge effekten af `id`, `class_id` og `school_id`, idet vi dermed tillader individuel variation for eleverne, klasserne og skolerne. Dermed er modellen givet ved

$$Y = X \begin{bmatrix} \beta_0 \\ \beta_{\texttt{time}} \\ \beta_{\texttt{gender}} \\ \beta_{\texttt{group}} \\ \beta_{\texttt{group:time}} \end{bmatrix} + Z \begin{bmatrix} U_{\texttt{id}} \\ U_{\texttt{class\_id}} \\ U_{\texttt{school\_id}} \end{bmatrix} + \varepsilon, \tag{3.54}$$

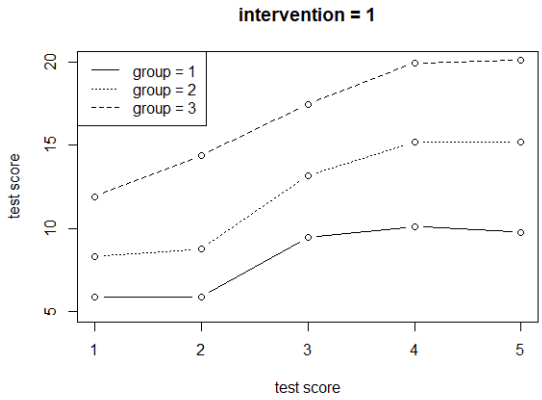
hvor $Y$ er variablen for alle testresultaterne for alle elever. Benytter vi os af `lmer`-funktionen fra `lme4`-pakken [1] til at tilpasse denne model for datasæt 1, får vi REML-estimaterne af parametrene i Tabel 3.1. Vi ser, at drenge i gennemsnit scorer $\hat{\beta}_{\texttt{gender1}} = 2.8264$ point højere

`intervention` = 1:

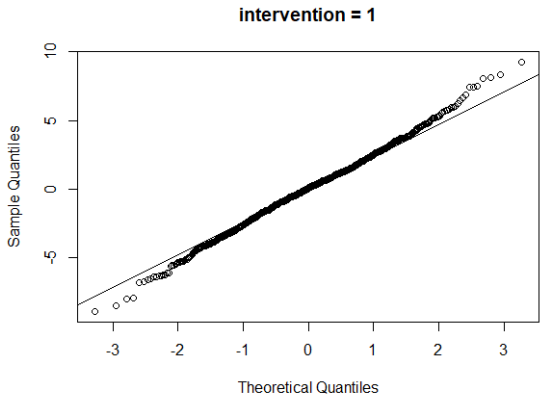
| | estimat | standardfejl | t-test | $p$-værdi | |
|---|---|---|---|---|---|
| $\beta_0$ | 4.7295 | 1.0615 | 4.456 | 0.00178 | ** |
| $\beta_{\texttt{time2}}$ | -0.0242 | 0.6319 | -0.038 | 0.96942 | |
| $\beta_{\texttt{time3}}$ | 3.5708 | 0.6345 | 5.628 | 2.61e-08 | *** |
| $\beta_{\texttt{time4}}$ | 4.1993 | 0.6738 | 6.233 | 7.77e-10 | *** |
| $\beta_{\texttt{time5}}$ | 3.8752 | 0.6351 | 6.101 | 1.71e-09 | *** |
| $\beta_{\texttt{group2}}$ | 2.3233 | 0.8606 | 2.700 | 0.00724 | ** |
| $\beta_{\texttt{group2:time2}}$ | 0.5022 | 0.7567 | 0.663 | 0.50723 | |
| $\beta_{\texttt{group2:time3}}$ | 1.2991 | 0.7614 | 1.706 | 0.08840 | . |
| $\beta_{\texttt{group2:time4}}$ | 2.7008 | 0.8055 | 3.353 | 0.00084 | *** |
| $\beta_{\texttt{group2:time5}}$ | 2.9807 | 0.7579 | 3.933 | 9.20e-05 | *** |
| $\beta_{\texttt{group3}}$ | 5.4178 | 0.9983 | 5.427 | 1.03e-07 | *** |
| $\beta_{\texttt{group3:time2}}$ | 2.4759 | 0.8673 | 2.855 | 0.00443 | ** |
| $\beta_{\texttt{group3:time3}}$ | 1.9992 | 0.8719 | 2.293 | 0.02214 | * |
| $\beta_{\texttt{group3:time4}}$ | 3.8280 | 0.9126 | 4.195 | 3.07e-05 | *** |
| $\beta_{\texttt{group3:time5}}$ | 4.3119 | 0.8700 | 4.956 | 8.98e-07 | *** |
| $\beta_{\texttt{gender1}}$ | 2.8264 | 0.5966 | 4.738 | 4.00e-06 | *** |

***Tabel 3.1:*** REML-estimater for datasæt 1. Signifikanskoder: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

end piger, og at testscoren for de højt præsterende elever forventes at være $\hat{\beta}_{\texttt{group3}} = 5.4178$ point højere end for de lavt præsterende elever til første test, hvis vi fastholder `gender`. Middelvækstkurverne for de tre grupper kan ses i Figur 3.1.




***Figur 3.1:*** Middelvækstkurverne for gruppe 1, 2 og 3 blandt elever i datasæt 1.




***Figur 3.2:*** Q-Q-plot af residualerne for modellen tilpasset på datasæt 1.

Udregner vi residualvektoren $y - \hat{y}$, hvor $\hat{y}$ er givet som i (3.39), ser vi på Figur 3.2, at disse ser ud til at være realiseringer af en normalfordeling. Dette øger altså plausibiliteten om antagelsen om normalitet af $Y$ og dermed vores tiltro til modellen.

Vi kan estimere ændringerne af CBM-scoren i hver af de fire perioder for hver gruppe ved at betragte de lineære transformationer i Appendiks A af parameterestimaterne i Tabel 3.1. Dermed kan vi også udregne standardfejlene for disse ændringer ved brug af standardfejlene fra tabellen og de tilhørende kovarianser. Bemærk, at disse ændringer svarer til hældningerne på Figur 3.1 og kan ses i Tabel 3.2. For at undersøge om hældningerne er signifikante, udregner vi t-teststørrelsen for hver af disse estimater, givet ved

$$t = \frac{\hat{\delta}^{t(t+1)}_{\texttt{group}k}}{\sqrt{\mathrm{Var}\left(\hat{\delta}^{t(t+1)}_{\texttt{group}k}\right)}}, \quad k = 1, 2, 3, \quad t = 1, \ldots, 4, \tag{3.55}$$

som under nulhypotesen, $\mathcal{H}_0 : \delta^{t(t+1)}_{\texttt{group}k} = 0$, følger en $t$-fordeling med $n - 1$ frihedsgrader [16, s. 70-71]. Vi ser i Tabel 3.2, at det kun er ændringen af CBM-scoren i anden periode, der er signifikant for de lavt præsterende elever. Denne ændring i testresultatet er estimeret ved

$$\hat{\delta}^{23}_{\texttt{group1}} = \hat{\beta}_{\texttt{time3}} - \hat{\beta}_{\texttt{time2}} = 3.5708 + 0.0242 = 3.5950. \tag{3.56}$$

For de højt præsterende elever, er det kun den sidste periode, som ikke er signifikant, og for de middel præsterende elever er anden og tredje periode signifikante, og disse ændringer er estimeret ved

$$\begin{aligned} \hat{\delta}^{23}_{\texttt{group2}} &= \hat{\beta}_{\texttt{group2:time3}} + \hat{\beta}_{\texttt{time3}} - (\hat{\beta}_{\texttt{group2:time2}} + \hat{\beta}_{\texttt{time2}}) = 4.3919, \\ \hat{\delta}^{34}_{\texttt{group2}} &= \hat{\beta}_{\texttt{group2:time4}} + \hat{\beta}_{\texttt{time4}} - (\hat{\beta}_{\texttt{group2:time3}} + \hat{\beta}_{\texttt{time3}}) = 2.0302. \end{aligned} \tag{3.57}$$

Bemærk, at eleverne blev undervist i brøkregning i anden og tredje periode. Derudover kan vi også estimere tre kontraster mellem grupperne for hver af de fire perioder. Her er blandt andet kontrasten i CBM-scoren mellem højt og lavt præsterende elever og mellem højt og

middel præsterende elever i første periode signifikante og er estimeret ved

$$\begin{aligned}\hat{\delta}^{12}_{\texttt{group3}} - \hat{\delta}^{12}_{\texttt{group1}} &= \hat{\beta}_{\texttt{group3:time2}} = 2.4759,\\ \hat{\delta}^{12}_{\texttt{group3}} - \hat{\delta}^{12}_{\texttt{group2}} &= \hat{\beta}_{\texttt{group3:time2}} - \hat{\beta}_{\texttt{group2:time2}} = 1.9737.\end{aligned} \tag{3.58}$$

Hvis eksempelvis $\hat{\delta}^{12}_{\texttt{group3}} - \hat{\delta}^{12}_{\texttt{group2}} = 0$, ville kurven for den gennemsnitlige udvikling i første periode for de to grupper blot være forskydninger af hinanden. For at teste om den gennemsnitlige udvikling i hver af de tre grupper er signifikant forskellige, kan vi benytte en F-test til at teste for hypotesen om, at $\beta_{\texttt{group}k\texttt{:time}t} = 0$ for $k = 1, 2, 3$ og $t = 2, \ldots, 5$. Ved at gøre dette med funktionen `KRmodcomp` fra R-pakken `pbkrtest` [8], får vi en $p$-værdi på 1.38e-05, hvormed nogle af hældningerne er signifikant forskellige.

`intervention` = 1:

| | estimat | standardfejl | t-test | $p$-værdi | |
|---|---|---|---|---|---|
| $\delta^{12}_{\texttt{group1}}$ | -0.0242 | 0.6319 | -0.038 | 0.96940 | |
| $\delta^{23}_{\texttt{group1}}$ | 3.5951 | 0.6518 | 5.516 | 4.48e-08 | *** |
| $\delta^{34}_{\texttt{group1}}$ | 0.6284 | 0.6956 | 0.904 | 0.36650 | |
| $\delta^{45}_{\texttt{group1}}$ | -0.3240 | 0.6926 | -0.468 | 0.63998 | |
| $\delta^{12}_{\texttt{group2}}$ | 0.4779 | 0.4166 | 1.147 | 0.25160 | |
| $\delta^{23}_{\texttt{group2}}$ | 4.3920 | 0.4297 | 10.221 | 2.48e-23 | *** |
| $\delta^{34}_{\texttt{group2}}$ | 2.0302 | 0.4530 | 4.482 | 8.30e-06 | *** |
| $\delta^{45}_{\texttt{group2}}$ | -0.0443 | 0.4457 | -0.099 | 0.92110 | |
| $\delta^{12}_{\texttt{group3}}$ | 2.4517 | 0.5941 | 4.127 | 4.00e-05 | *** |
| $\delta^{23}_{\texttt{group3}}$ | 3.1184 | 0.6137 | 5.081 | 4.52e-07 | *** |
| $\delta^{34}_{\texttt{group3}}$ | 2.4572 | 0.6311 | 3.894 | 0.00010 | *** |
| $\delta^{45}_{\texttt{group3}}$ | 0.1598 | 0.6269 | 0.255 | 0.79880 | |

***Tabel 3.2:*** Estimater af hældningerne i datasæt 1. Signifikanskoder: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

Fra modellen får vi også, at estimatet for henholdsvis elev-, klasse- og skolevariansen er givet ved

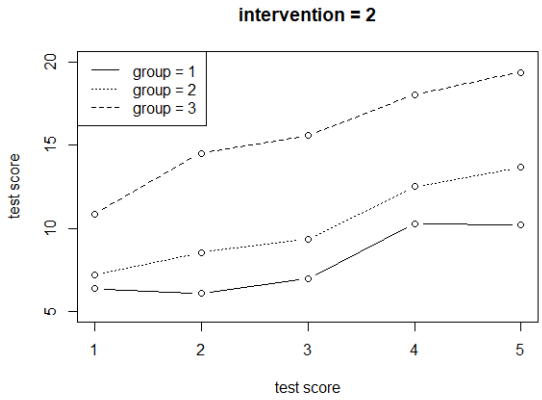
$$\hat{\sigma}^2_{\texttt{id}} = 16.23, \quad \hat{\sigma}^2_{\texttt{class\_id}} = 0.30, \quad \hat{\sigma}^2_{\texttt{school\_id}} = 2.06, \tag{3.59}$$

samt at variansestimatet af residualet er $\hat{\sigma}^2 = 8.80$.

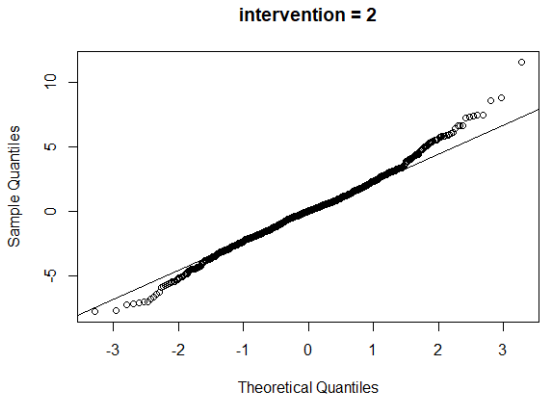
For datasæt 2 får vi REML-estimaterne, som kan ses i Tabel 3.3. Her ser vi igen, at de højt præsterende elever i gennemsnit har en højere testscore end de lavt præsterende elever. Derudover er testscoren for drenge nu kun estimeret til at være $\hat{\beta}_{\texttt{gender1}} = 0.8093$ point højere end testscoren for piger og er ikke længere signifikant. Middelvækstkurverne for denne model kan ses på Figur 3.3 og et Q-Q-plot af residualerne på Figur 3.4. Her ser vi igen, at residualerne ser ud til at være realiseringer af en normalfordeling, hvormed vi slutter samme konklusion som tidligere.

| `intervention` = 2: | | | | | |
|---|---|---|---|---|---|
| | estimat | standardfejl | t-test | $p$-værdi | |
| $\beta_0$ | 6.1343 | 0.8802 | 6.969 | 1.24e-07 | *** |
| $\beta_{\texttt{time2}}$ | -0.2742 | 0.5967 | -0.460 | 0.64598 | |
| $\beta_{\texttt{time3}}$ | 0.6036 | 0.6138 | 0.983 | 0.32577 | |
| $\beta_{\texttt{time4}}$ | 3.8975 | 0.6219 | 6.267 | 6.20e-10 | *** |
| $\beta_{\texttt{time5}}$ | 3.8492 | 0.6006 | 6.409 | 2.60e-10 | *** |
| $\beta_{\texttt{group2}}$ | 0.6687 | 0.8362 | 0.800 | 0.42436 | |
| $\beta_{\texttt{group2:time2}}$ | 1.6062 | 0.7122 | 2.255 | 0.02440 | * |
| $\beta_{\texttt{group2:time3}}$ | 1.5131 | 0.7243 | 2.089 | 0.03704 | * |
| $\beta_{\texttt{group2:time4}}$ | 1.3766 | 0.7311 | 1.883 | 0.06010 | . |
| $\beta_{\texttt{group2:time5}}$ | 2.5764 | 0.7102 | 3.628 | 0.00031 | *** |
| $\beta_{\texttt{group3}}$ | 4.3126 | 1.0002 | 4.312 | 2.06e-05 | *** |
| $\beta_{\texttt{group3:time2}}$ | 3.9199 | 0.8627 | 4.544 | 6.45e-06 | *** |
| $\beta_{\texttt{group3:time3}}$ | 4.1035 | 0.8780 | 4.674 | 3.51e-06 | *** |
| $\beta_{\texttt{group3:time4}}$ | 3.2805 | 0.9113 | 3.600 | 0.00034 | *** |
| $\beta_{\texttt{group3:time5}}$ | 4.6613 | 0.8653 | 5.387 | 9.62e-08 | *** |
| $\beta_{\texttt{gender1}}$ | 0.8093 | 0.5742 | 1.409 | 0.16020 | |

***Tabel 3.3:*** REML-estimater for datasæt 2. Signifikanskoder: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'




***Figur 3.3:*** Middelvækstkurverne for gruppe 1, 2 og 3 blandt elever i datasæt 2.




***Figur 3.4:*** Q-Q-plot af residualerne for modellen tilpasset på datasæt 2.

Vi får yderligere variansestimaterne

$$\begin{aligned} \hat{\sigma}^2_{\texttt{id}} = 13.948, \quad \hat{\sigma}^2_{\texttt{class\_id}} = 0, \\ \hat{\sigma}^2_{\texttt{school\_id}} = 1.687, \quad \hat{\sigma}^2 = 8.101. \end{aligned} \tag{3.60}$$

Der er altså ingen evidens er for variation mellem klasserne, hvorfor denne effekt kunne udelades fra modellen. Vi bemærker, at elevvariationen udgør 59% af den totale varians,

som også er tilfældet for datasæt 1. Klassevariationen er ligeledes lille i begge datsæt, og skolevariationen svarer til henholdsvis 8% og 7% af den totale varians. Der er altså overordnet set stor konsistens mellem variansestimaterne for de to uafhængige datasæt. Dette bekræfter os i, at elevvariansen udgør den største del af den totale varians, samt at klassevariansen næsten er uden betydning.

Ved igen at benytte t-tests for ændringerne i CBM-scoren for hver af de tre grupper ser vi, at ændringen i første, anden og fjerde periode er signifikante for både de højt og middel præsterende elever. For de lavt præsterende elever er det kun ændringen i tredje periode, som er signifikant. Bemærk, at eleverne i dette datasæt blev undervist i brøkregning i tredje og fjerde periode. Benytter vi nu F-test for nulhypotesen om, at $\beta_{\texttt{group}k\texttt{:time}t} = 0$ for $k = 1, 2, 3$ og $t = 2, \ldots, 5$, får vi en $p$-værdi på 6.62e-06, hvorfor vi igen kan konkludere, at nogle af hældningerne er signifikant forskellige.

### 3.5.2 Vurdering af gruppeinddelingen

For at undersøge om gruppeinddelingen i datasættet, baseret på de nationale tests, stemmer overens med eleveffekterne, tilpasser vi samme mixed model som i (3.54), men hvor `group`, og dermed også krydsfaktoren `group:time`, ikke indgår i modellen. Vi får da estimaterne i Tabel 3.4 af de systematiske effekter for de to datasæt.

`intervention` = 1:

| | estimat | standardfejl | t-test | $p$-værdi | |
|---|---|---|---|---|---|
| $\beta_0$ | 6.8612 | 1.0751 | 6.382 | 0.00453 | ** |
| $\beta_{\texttt{time2}}$ | 0.8504 | 0.3063 | 2.776 | 0.00564 | ** |
| $\beta_{\texttt{time3}}$ | 4.7500 | 0.3088 | 15.380 | <2.00e-16 | *** |
| $\beta_{\texttt{time4}}$ | 6.5958 | 0.3235 | 20.388 | <2.00e-16 | *** |
| $\beta_{\texttt{time5}}$ | 6.5233 | 0.3056 | 21.346 | <2.00e-16 | *** |
| $\beta_{\texttt{gender}}$ | 3.5340 | 0.6929 | 5.100 | 7.58e-07 | *** |

`intervention` = 2:

| | estimat | standardfejl | t-test | $p$-værdi | |
|---|---|---|---|---|---|
| $\beta_0$ | 7.1904 | 0.6500 | 11.062 | 8.66e-08 | *** |
| $\beta_{\texttt{time2}}$ | 1.4586 | 0.2945 | 4.952 | 9.05e-07 | *** |
| $\beta_{\texttt{time3}}$ | 2.3259 | 0.2952 | 7.880 | 1.15e-14 | *** |
| $\beta_{\texttt{time4}}$ | 5.3644 | 0.2995 | 17.914 | <2.00e-16 | *** |
| $\beta_{\texttt{time5}}$ | 6.2828 | 0.2909 | 21.598 | <2.00e-16 | *** |
| $\beta_{\texttt{gender}}$ | 1.2511 | 0.6658 | 1.879 | 0.06160 | . |

***Tabel 3.4:*** REML-estimater for de to datasæt. Signifikanskoder: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

For datasæt 1 får vi variansestimaterne

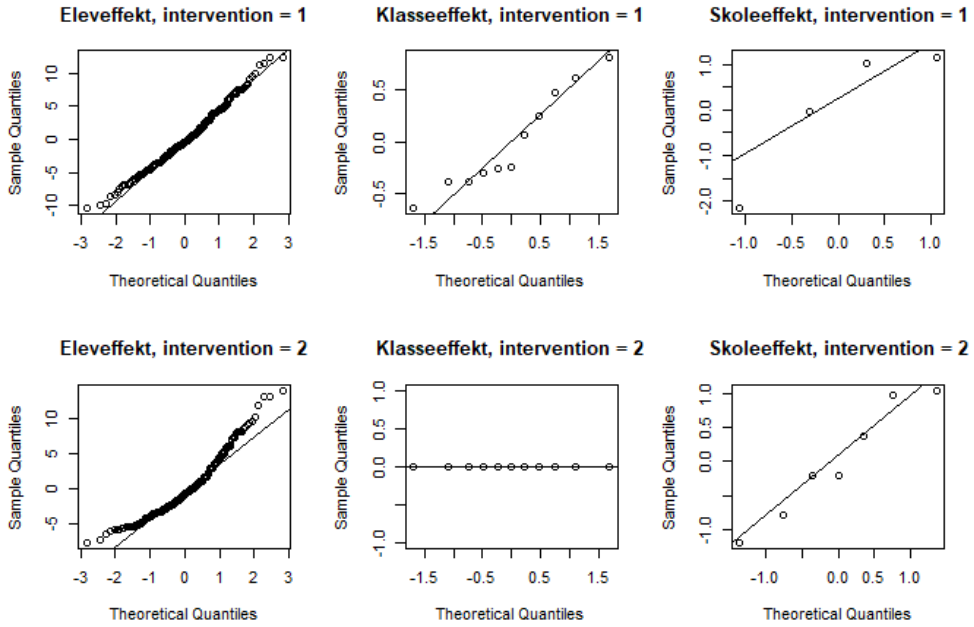
$$\begin{aligned} \hat{\sigma}^2_{\texttt{id}} = 23.05, \quad \hat{\sigma}^2_{\texttt{class\_id}} = 0.64, \\ \hat{\sigma}^2_{\texttt{school\_id}} = 3.3, \quad \hat{\sigma}^2 = 9.16, \end{aligned} \tag{3.61}$$

og for datasæt 2 får vi

$$\begin{aligned} \hat{\sigma}^2_{\texttt{id}} = 20.4, \quad \hat{\sigma}^2_{\texttt{class\_id}} = 0, \\ \hat{\sigma}^2_{\texttt{school\_id}} = 1.2, \quad \hat{\sigma}^2 = 8.43. \end{aligned} \tag{3.62}$$

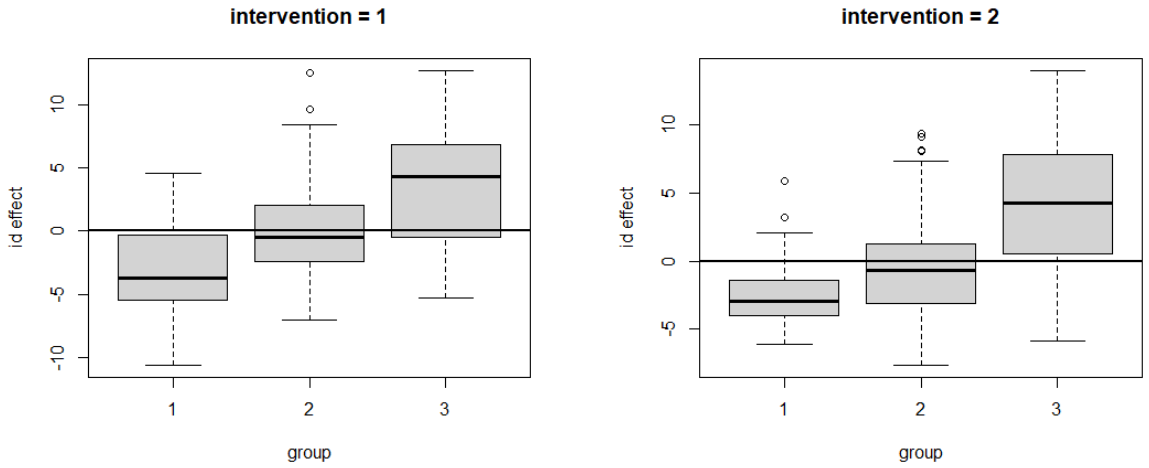
Vi bemærker altså, at alle varianserne er højere end varianserne i modellen med gruppefaktoren med undtagelse af skolevariansen for datasæt 2. Det giver mening, at det særligt er elevvarianserne, der er blevet højere, idet vi har fjernet gruppefaktoren fra middelværdistrukturen.

Ved at benytte BLUP kan vi prædiktere de forskellige stokastiske effekter. På Q-Q-plottene på Figur 3.5 ser vi, at det er rimeligt at antage, at prædiktionerne er realiseringer af en normalfordeling.




***Figur 3.5:*** Q-Q-plots for de tre forskellige stokastiske effekter for begge datasæt.

Vi laver nu en opdeling af prædiktionerne for eleveffekterne baseret på præstationsgrupperne. På Figur 3.6 ser vi for begge datasæt, at de højt præsterende elever generelt har positive stokastiske effekter, hvorimod det omvendte er gældende for de lavt præsterende elever. Yderligere er effekterne af de middel præsterende eleverne centreret omkring 0. Dette tyder altså på, at den oprindelige gruppeinddeling, baseret på de nationale tests, er overvejende acceptabel.




***Figur 3.6:*** Boksplots af de prædikterede eleveffekter fordelt over de tre grupper, baseret på de nationale tests.

# 4 | Latente vækstmodeller

Idet vores datasæt består af *longitudinelle* data, altså gentagne målinger af testscoren for hver af de $d$ elever, er vi i stand til at estimere en uobserveret underliggende matematisk kompetence, som udvikler sig som en kontinuert funktion af tiden. Ved at betragte en såkaldt *latent vækstmodel*, som indeholder stokastiske skæringer og stokastiske hældninger, er det muligt, at eleverne kan have forskellige vækstkurver over tid [2, s. xi]. En latent vækstmodel er specificeret ved to niveauer af ligninger. Det første niveau består af en vækstkurvemodel for $Y$, som for det $i$'te individ til tidspunkt $t \in \{1, \dots, T\}$ er givet ved

$$Y_{it} = \eta_{i1} + \sum_{s=2}^{T} \lambda_{t(s-1)} \eta_{is} + \varepsilon_{it}, \tag{4.1}$$

hvor $\eta_{i1}$ er den stokastiske skæring for det $i$'te individ, $\sum_{s=2}^{m+1} \eta_{is}$ er den stokastiske hældning i den $m$'te periode for det $i$'te individ og $\lambda_{ts} = (t-s)\mathbf{1}[t > s]$ for alle $t$. Vi antager, at $\varepsilon_{it} \sim N(0, \sigma_{it}^2)$, hvor $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{jt'}) = 0$ for $i \neq j$ og $t, t' = 1, \dots, T$. Altså er fejlleddene for forskellige elever ukorrelerede til de forskellige tider [2, s. 127]. Det andet niveau består af ligninger for de stokastiske latente variable

$$\eta_{is} = \alpha_s + \sum_{j=1}^{p} \gamma_{sj} x_{ij} + \zeta_{is}, \quad s = 1, \dots, T, \tag{4.2}$$

hvor $\alpha_s$ er skæringen for $\eta_{is}$ på tværs af alle elever for alle $s = 1, \dots, T$. Derudover er $x_{ij}$, for $j = 1, \dots, p$, kovariater som påvirker de latente variable og $\gamma_{sj}$, for $s = 1, \dots, T$, er de tilhørende parametre. Vi antager, at $\zeta_{is} \sim N(0, \psi_s)$, samt at $\mathrm{Cov}(\zeta_{is}, \zeta_{is'}) = \psi_{ss'}$ for $s \neq s'$. Derudover er $\mathrm{Cov}(\zeta_{is}, \zeta_{js'}) = \mathrm{Cov}(\zeta_{is}, \varepsilon_{it}) = 0$ for $i \neq j$ og $t, s, s' = 1, \dots, T$ [2, s. 127].

Modellen kan også udtrykkes på matrix-vektorform ved

$$Y_i = \Lambda \eta_i + \varepsilon_i, \tag{4.3}$$

hvor $Y_i = (Y_{i1}, \dots, Y_{iT})^T$, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT})^T$ og $\Lambda$ er en $T \times T$-matrix bestående af $\mathbf{1}_T$ som første søjle og $\lambda_{ts}$ på den $t(s+1)$'te indgang, sådan at

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & T-1 & T-2 & \cdots & 1 \end{bmatrix}. \tag{4.4}$$

Yderligere er $\eta_i$ vektoren bestående af de latente variable for det $i$'te individ og er givet ved

$$\eta_i = \alpha + \Gamma x_i + \zeta_i, \tag{4.5}$$

hvor $\alpha = (\alpha_1, \ldots, \alpha_T)^T$, $\zeta_i = (\zeta_{i1}, \ldots, \zeta_{iT})^T$ og

$$\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1p} \\ \vdots & \ddots & \vdots \\ \gamma_{T1} & \cdots & \gamma_{Tp} \end{bmatrix}. \tag{4.6}$$

Vi lader yderligere $\mathrm{Var}(\varepsilon_i) = \Sigma$ og $\mathrm{Var}(\zeta_i) = \Psi$. Ved at opskrive den latente vækstmodel som en kombineret model givet ved

$$Y_i = \Lambda(\alpha + \Gamma x_i + \zeta_i) + \varepsilon_i, \tag{4.7}$$

bemærker vi, at

$$Y_i \mid x_i \sim N(\Lambda(\alpha + \Gamma x_i), \Lambda\Psi\Lambda^T + \Sigma), \tag{4.8}$$

hvormed modellen er på samme form som en mixed model, og en latent vækstmodel er altså et specialtilfælde heraf.

**Eksempel 4.1**
For simpelhedens skyld betragter vi nu en delmængde af datasættet, hvor vi kun medtager observationerne fra de første tre tests. Vi lader nu $\zeta_i = e_1\zeta_{i1}$, hvor $\zeta_{i1} \sim N(0, \psi_1)$. Ved da at betinge med kovariaterne `gender` og `age`, betegnet ved henholdsvis $x_1$ og $x_2$, kan vi opskrive den kombinerede model som i (4.7) for den $i$'te elev ved

$$\begin{aligned}
\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \left( \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \\ \gamma_{31} & \gamma_{32} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} + \begin{bmatrix} \zeta_{i1} \\ 0 \\ 0 \end{bmatrix} \right) + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 + \gamma_{11}x_{i1} + \gamma_{12}x_{i2} \\ \alpha_1 + \alpha_2 + (\gamma_{11} + \gamma_{21})x_{i1} + (\gamma_{12} + \gamma_{22})x_{i2} \\ \alpha_1 + 2\alpha_2 + \alpha_3 + (\gamma_{11} + 2\gamma_{21} + \gamma_{31})x_{i1} + (\gamma_{12} + 2\gamma_{22} + \gamma_{32})x_{i2} \end{bmatrix} \\
&\quad + \begin{bmatrix} \zeta_{i1} \\ \zeta_{i1} \\ \zeta_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} x_{i1} \\ x_{i1} \\ x_{i1} \end{bmatrix} \gamma_{11} + \begin{bmatrix} 0 & 0 \\ x_{i1} & 0 \\ 2 & x_{i1} \end{bmatrix} \begin{bmatrix} \gamma_{21} \\ \gamma_{31} \end{bmatrix} \\
&\quad + \begin{bmatrix} x_{i2} \\ x_{i2} \\ x_{i2} \end{bmatrix} \gamma_{12} + \begin{bmatrix} 0 & 0 \\ x_{i2} & 0 \\ 2 & x_{i2} \end{bmatrix} \begin{bmatrix} \gamma_{22} \\ \gamma_{32} \end{bmatrix} + \begin{bmatrix} \zeta_{i1} \\ \zeta_{i1} \\ \zeta_{i1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix}.
\end{aligned} \tag{4.9}$$

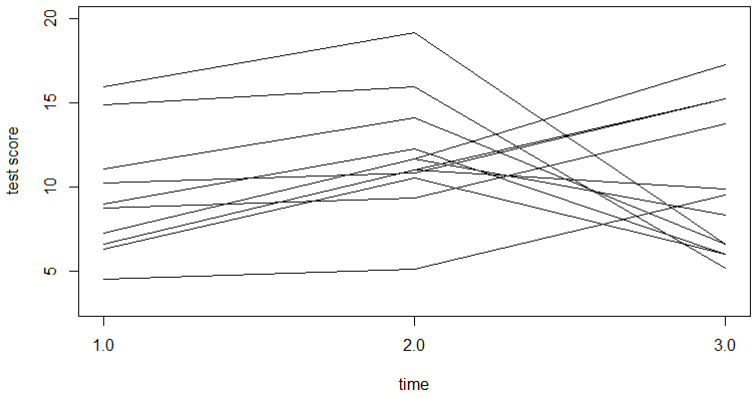
Vi ser altså, at dette svarer til en mixed model på formen

$$\begin{aligned} y_i = \mathbf{1}_3\beta_0 + (X_{\texttt{time}})_i\beta_{\texttt{time}} + (X_{\texttt{gender}})_i\beta_{\texttt{gender}} + (X_{\texttt{gender:time}})_i\beta_{\texttt{gender:time}} \\ + (X_{\texttt{age}})_i\beta_{\texttt{age}} + (X_{\texttt{age:time}})_i\beta_{\texttt{age:time}} + \mathbf{1}_3(U_{\texttt{id}})_i + \varepsilon_i, \end{aligned} \tag{4.10}$$

hvor eksempelvis

$$(X_{\texttt{gender}})_i = \mathbf{1}_3 x_{i1} \quad \text{og} \quad \beta_{\texttt{gender}} = \gamma_{11}. \tag{4.11}$$

Eksempler på vækstkurver for denne model kan ses på Figur 4.1.




***Figur 4.1:*** Individuelle vækstkurver for 12 forskellige elever baseret på modellen i (4.10).

◀

Da vi er interesserede i at undersøge udviklingen over forskellige undervisningsforløb, har vi i dette afsnit begrænset vores model til kun at omfatte stykkevise lineære vækstkurver, idet vi har valgt $\Lambda$ som i (4.4). På denne måde opnår vi vækstkurver beskrevet med lineære splines, hvilket ofte er anvendeligt i praksis. Hvis vi eksempelvis tilpasser en model med kun én stokastisk hældning, altså vælger $\Lambda$ som en $T \times 2$-matrix, bestående af de to første søjler i (4.4), vil væksten i elevernes testresultater være konstant over alle tidsperioder. En sådan model kaldes en lineær latent vækstmodel og vil i dette tilfælde ikke være særlig retvisende grundet elevernes undervisningsforløb. Antallet af lineære splines afhænger altså af data, og hvad vi ønsker at modellere. Yderligere kunne vi også estimere de $T^2$ indegange i $\Lambda$ ud fra data, men i så fald kan vi risikere at få for mange parametre i modellen, hvormed vi ikke kan identificere denne [20, s. 465]. Alternativt kan vækstkurverne også modelleres ved at tilføje kvadratiske eller kubiske led af tiden til en lineær latent vækstmodel, hvormed vi vil opnå ikke-lineære vækstkurver [2, s. 89].

## 4.1 Growth Mixture modeller

Hvis data indeholder latente grupperinger, kan vi udvide en latent vækstmodel til en *Growth Mixture model (GMM)*. Denne model kan identificere de uobserverede grupper, givet at vi specificerer antallet af disse. Vi lader $K$ være antallet af latente grupper og definerer den stokastiske variabel $C_i$, hvor $C_i = k$, hvis og kun hvis det $i$'te individ tilhører den $k$'te latente gruppe. Dermed gælder det altså, at $C_i$ følger en multinomialfordeling med antalsparameter 1 og er uafhængig af $C_j$ for $j \neq i$. Vi opskriver da en GMM ved

$$Y_i = \Lambda \eta_i + \varepsilon_i, \tag{4.12}$$

hvor $\text{Var}(\varepsilon_i) = \Sigma$ er en diagonalmatrix [20, s. 464]. Dermed er dette niveau af modelspecifikationen ens med modelspecifikationen for latente vækstmodeller. Forskellen opstår dog ved specifikationen af $\eta_i$, som for en GMM er givet ved

$$\eta_i = \sum_{k=1}^{K} (\alpha_k + \Gamma_k x_i)\, \mathbf{1}[c_i = k] + \zeta_i, \tag{4.13}$$

hvor $\alpha_k$ og $\Gamma_k$ er henholdsvis skæringsvektor og parametermatrix for den $k$'te gruppe [20, s. 464].

Lad nu $\pi_i$ være en $K$-dimensional vektor med indgangene $\pi_{ik} = \mathbb{P}(C_i = k \mid x_i)$ for $k = 1, \ldots, K$, hvor $\sum_{k=1}^{K} \pi_{ik} = 1$, hvormed tætheden for $C_i \mid x_i$ kan udtrykkes som

$$f(c_i \mid x_i) = \prod_{k=1}^{K} \pi_{ik}^{\mathbf{1}[c_i = k]}. \tag{4.14}$$

Vi kan modellere $\pi_i$ ved en multivariat logistisk regression, hvor vi bruger den $K$'te gruppe som reference. Det gælder derfor, at

$$\begin{bmatrix} \text{logit}(\pi_{i1}) \\ \vdots \\ \text{logit}(\pi_{i(K-1)}) \end{bmatrix} = \alpha_c + \Gamma_c x_i, \tag{4.15}$$

med skæringsvektor $\alpha_c$ og parametermatrix $\Gamma_c$.

Af ovenstående modelspefikation gælder det, at $\eta_i \mid C_i = k, x_i \sim N_T(\alpha_k + \Gamma_k x_i, \Psi)$ og $Y_i \mid \eta_i \sim N_T(\Lambda \eta_i, \Sigma)$, hvormed vi får, at

$$Y_i \mid C_i = k, x_i \sim N_T\left(\Lambda(\alpha_k + \Gamma_k x_i), \Lambda \Psi \Lambda^T + \Sigma\right). \tag{4.16}$$

Vi kan nu udregne tætheden for $Y_i \mid x_i$ ved

$$\begin{aligned} f(y_i \mid x_i) &= \sum_{c_i} f(c_i \mid x_i) f(y_i \mid c_i, x_i) \\ &= \sum_{c_i} \left( \prod_{k=1}^{K} \pi_{ik}^{\mathbf{1}[c_i = k]} \right) f(y_i \mid c_i, x_i) \\ &= \sum_{k=1}^{K} \pi_{ik} f(y_i \mid c_i = k, x_i). \end{aligned} \tag{4.17}$$

Tætheden er altså en sum af $K$ normalfordelingstætheder vægtet med sandsynligheden for, at individet tilhører den pågældende gruppe, givet $x_i$, hvilket er en såkaldt *mixturtæthed* [20, s. 464]. Dermed er log-likelihoodfunktionen for $\theta$, som indeholder alle de ukendte parametre i $\alpha_1, \ldots, \alpha_K$, $\Gamma_1, \ldots, \Gamma_K$, $\alpha_c$, $\Gamma_c$, $\Psi$ og $\Sigma$, givet ved

$$\ell(\theta; y) = \sum_{i=1}^{d} \log \sum_{k=1}^{K} \pi_{ik}(\theta) f(y_i \mid c_i = k, x_i; \theta). \tag{4.18}$$

Bemærk dog, at denne log-likelihoodfunktion er besværlig at maksimere, idet den består af logaritmen af en sum af tæthederne, og vi er ikke garanteret at kunne finde et udtryk for MLE på lukket form. Vi ønsker derfor i stedet at benytte numeriske optimeringsmetoder, som vi vil beskrive i Afsnit 4.3.

Tilsvarende Afsnit 3.3 kan vi prædiktere $\zeta_i$ ved

$$\hat{\zeta}_i = \Psi\Lambda^T(\Lambda\Psi\Lambda^T + \Sigma)^{-1}(y_i - \Lambda(\alpha_k + \Gamma_k x_i)), \tag{4.19}$$

givet at $C_i = k$, såfremt at $(\Lambda^T\Sigma^{-1}\Lambda + \Psi^{-1})^{-1}$ eksisterer, hvor vi i praksis vil indsætte estimater for de ukendte parametre.

## 4.2 Prædiktion af klassifikation

Vi kan prædiktere, hvilken gruppe det $i$'te individ tilhører ved at finde det $k$, som opfylder, at

$$k = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \mathbb{P}(C_i = k \mid y_i, x_i). \tag{4.20}$$

Da $C_i$ er en diskret variabel, skal vi blot finde det $k$, som maksimerer

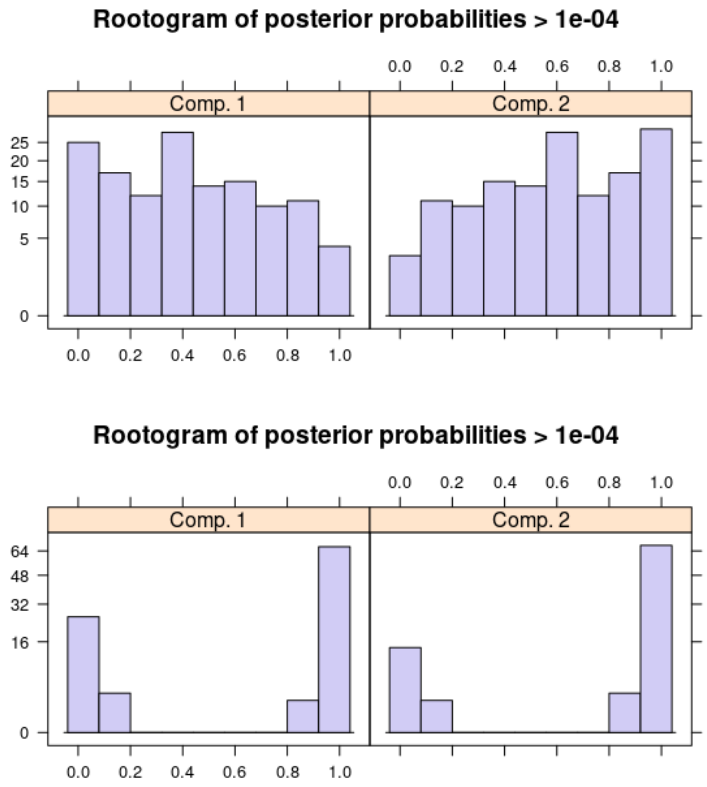
$$f(c_i = k \mid y_i, x_i) = \frac{f(y_i \mid c_i = k, x_i) f(c_i = k \mid x_i)}{f(y_i \mid x_i)} = \frac{\pi_{ik} f(y_i \mid c_i = k, x_i)}{\sum_{k'=1}^{K} \pi_{ik'} f(y_i \mid c_i = k', x_i)}. \tag{4.21}$$

Denne tæthed kalder vi en *posterior* sandsynlighed, og vi kan bestemme et estimat for denne, såfremt vi har maksimeret log-likelihoodfunktionen i (4.18) [12, s. 3].

Hvis det $i$'te individ bliver prædikteret til den $k$'te gruppe, hvor $f(c_i = k \mid y_i, x_i)$ er tæt på 1, og $f(c_i = l \mid y_i, x_i)$ er tæt på 0 for $l \neq k$, siger vi, at gruppe $k$ er godt separeret fra de resterende grupper. Omvendt er der mere overlap i den $k$'te gruppe, hvis $f(c_i = k \mid y_i, x_i)$ er tættere på $f(c_i = l \mid y_i, x_i)$ for et $l \neq k$. For hvert $k$ kan vi udregne den gennemsnitlige posterior sandsynlighed for at tilhøre denne gruppe over individerne, der er prædikteret til at tilhøre gruppe $k$. Hvis dette gennemsnit er tæt på 1, indikerer det, at gruppen er separeret fra de resterende. Omvendt betyder et mindre gennemsnit, at der er mere overlap i grupperne.

Bemærk, at idet vi betragter longitudinelle data, kan vi tildele det enkelte elevs posterior sandsynlighed til hver af elevens observationer. Disse posterior sandsynligheder kan vi benytte til at konstruere et *rootogram* for hver gruppe, som blot er et histogram, hvor højden af blokkene er kvadratroden af antallet af observationer, hvis posterior sandsynlighed ligger i det pågældende interval. Rootogrammet for den $k$'te gruppe indeholder altså posterior sandsynlighederne for at tilhøre denne gruppe, for alle observationerne i datasættet. Vi frasorterer

observationer med en posterior sandsynlighed mindre end et $\delta > 0$, da det ellers kan være besværligt at vurdere rootogrammet for de observationer, som har en betydelig sandsynlighed for at tilhøre den pågældende gruppe, hvis mange af disse er tæt på 0. Ofte vælges $\delta =$ 1e-04. To eksempler på rootogrammer kan ses på Figur 4.2 for $K = 2$. Hvis rootogrammet viser en




***Figur 4.2:*** Eksempler på rootogrammer.

betydelig høj frekvens omkring 1, indikerer dette, at gruppen er separeret fra de resterende. Derimod indikerer det, at gruppen overlapper med en eller flere af de andre grupper, hvis der er meget masse i midten af intervallet. Vi ser altså en indikation på, at modellen tilhørende de øverste rootogrammer har overlappende grupper, og at modellen tilhørende de nederste rootogrammer har separerede grupper

## 4.3 Metoder til parameterestimation

I dette afsnit ønsker vi at undersøge, hvordan vi maksimerer log-likelihoodfunktionen i (4.18). Som nævnt tidligere kan vi oftest ikke gøre dette direkte, hvorfor vi i dette afsnit vil undersøge nogle numeriske metoder til at maksimere denne. For sådanne numeriske metoder skal vi vælge en startværdi, $\theta^{(0)}$, for parametervektoren $\theta \in \mathbb{R}^{n_\theta}$, hvorefter denne iterativt opdateres, således at den konvergerer mod det $\theta$, der maksimerer log-likelihoodfunktionen.

### 4.3.1 Quasi Newton Raphson-algoritmen

Vi kan eksempelvis opdatere $\theta$ iterativt ved at benytte *Newton Raphson-algoritmen* til at finde et nulpunkt for scorefunktionen, $S(\theta; y) = \frac{\partial}{\partial\theta}\ell(\theta; y)$. Af Taylors Sætning får vi, at denne kan approksimeres ved

$$S(\theta; y) \approx S(\theta^{(r)}; y) + H(\theta^{(r)}; y)(\theta - \theta^{(r)}), \tag{4.22}$$

hvor $H(\theta; y) = \frac{\partial}{\partial\theta^T}S(\theta; y)$ er Hessematricen for log-likelihoodfunktionen, og hvor $\theta^{(r)}$ er den approksimation af $\theta$, som er opnået ved den $r$'te iteration. Nulpunktet for denne approksimation er

$$\theta = \theta^{(r)} - H(\theta^{(r)}; y)^{-1}S(\theta^{(r)}; y), \tag{4.23}$$

hvorfor vi for $r = 0, 1, \ldots$ får, at en opdatering af $\theta$ er givet ved

$$\theta^{(r+1)} = \theta^{(r)} - H(\theta^{(r)}; y)^{-1}S(\theta^{(r)}; y), \tag{4.24}$$

som fortsættes indtil den ønskede konvergens er opnået. Her er det nødvendigt, at $H(\theta^{(r)}; y)$ er negativ definit, for at opdateringerne af $\theta$ for hver iteration går mod et maksimum [21, s. 23]. For at opstille Hessematricen for hver iteration skal vi dog udlede de dobbeltafledede af log-likelihoodfunktionen, hvilket kan være besværligt. Vi kan i stedet benytte en såkaldt *quasi Newton Raphson-algoritme*, hvor $H(\theta^{(r)}; y)$ approksimeres ved en *finite differences-metode* for hver iteration, således at vi ikke behøver at specificere denne. Vi betegner denne approksimerede Hessematrix ved en symmetrisk matrix $H^{(r)}$, som opfylder *sekantligningen*

$$H^{(r)}(\theta^{(r)} - \theta^{(r-1)}) = S(\theta^{(r)}; y) - S(\theta^{(r-1)}; y), \tag{4.25}$$

for $r = 1, 2, \ldots$, idet den i så fald er en approksimation af Jacobimatricen af $S(\theta^{(r)}; y)$. I stedet for at approksimere Hessematricen for derefter at bestemme den inverse, er det fordelagtigt direkte at benytte en approksimeret invers Hessematrix, $B^{(r)}$. Denne skal da opfylde, at

$$B^{(r)}s^{(r-1)} = \Delta^{(r-1)}, \tag{4.26}$$

hvor $\Delta^{(r-1)} = \theta^{(r)} - \theta^{(r-1)}$ og $s^{(r-1)} = S(\theta^{(r)}; y) - S(\theta^{(r-1)}; y)$ [21, s. 137]. Vi vælger indledningsvist en negativ definit og symmetrisk startværdi for denne approksimation, som eksempelvis $B^{(0)} = -I_{n_\theta}$, hvormed vi sikrer os, at første iteration bevæger sig i retning af gradienten for log-likelihoodfunktionen i $\theta^{(0)}$. For at bestemme $B^{(r)}$ skal vi vælge den entydige symmetriske matrix, som opfylder (4.26), og som er tættest på $B^{(r-1)}$, altså den matrix som løser

$$\min_B \|B - B^{(r-1)}\| \quad \text{s.t.} \quad B^T = B, \quad Bs^{(r-1)} = \Delta^{(r-1)}, \tag{4.27}$$

hvor $\|\cdot\|$ betegner Frobenius-normen [21, s. 139-140]. Det kan vises, at matricen som opfylder dette, kan opnås ved at korrigere den forrige værdi af approksimationen af den inverse Hessematrix, sådan at

$$\begin{aligned} B^{(r)} = B^{(r-1)} &+ \left(1 + \frac{(s^{(r-1)})^T B^{(r-1)} s^{(r-1)}}{(s^{(r-1)})^T \Delta^{(r-1)}}\right) \frac{\Delta^{(r-1)}(\Delta^{(r-1)})^T}{(s^{(r-1)})^T \Delta^{(r-1)}} \\ &- \frac{\Delta^{(r-1)}(s^{(r-1)})^T B^{(r-1)} + B^{(r-1)} s^{(r-1)}(\Delta^{(r-1)})^T}{(s^{(r-1)})^T \Delta^{(r-1)}}. \end{aligned} \tag{4.28}$$

Det gælder yderligere, at den tilsvarende approksimation af Hessematricen er givet ved

$$H^{(r)} = H^{(r-1)} - \frac{H^{(r-1)}\Delta^{(r-1)}(\Delta^{(r-1)})^T H^{(r-1)}}{(\Delta^{(r-1)})^T H^{(r-1)}\Delta^{(r-1)}} + \frac{s^{(r-1)}(s^{(r-1)})^T}{(s^{(r-1)})^T\Delta^{(r-1)}}, \tag{4.29}$$

samt at disse approksimationer er hinandens inverse [21, s. 140].

Antag nu, at $H^{(r-1)}$ er negativ definit for et $r \geq 1$. Da $H^{(r-1)}$ yderligere er symmetrisk, findes der en entydig Cholesky-dekomposition, $H^{(r-1)} = -LL^T$, for en nedre trekantsmatrix, $L$, med positive diagonalindgange [21, s. 608]. Dermed gælder det for et vilkårligt $z \neq 0$ med dimension $n_\theta$, at

$$\begin{aligned} &z^T\left(H^{(r-1)} - \frac{H^{(r-1)}\Delta^{(r-1)}(\Delta^{(r-1)})^T H^{(r-1)}}{(\Delta^{(r-1)})^T H^{(r-1)}\Delta^{(r-1)}}\right) z \\ &= z^T\left(\frac{LL^T\Delta^{(r-1)}(\Delta^{(r-1)})^T LL^T}{(\Delta^{(r-1)})^T LL^T\Delta^{(r-1)}} - LL^T\right) z \\ &= \frac{((L^T\Delta^{(r-1)})^T L^T z)^2}{(L^T\Delta^{(r-1)})^T L^T\Delta^{(r-1)}} - (L^T z)^T L^T z \leq 0, \end{aligned} \tag{4.30}$$

hvor uligheden kommer af Cauchy-Schwarz-uligheden [21, s. 600]. Derudover gælder det, at hvis $(s^{(r-1)})^T\Delta^{(r-1)} < 0$, så er

$$z^T\left(\frac{s^{(r-1)}(s^{(r-1)})^T}{(s^{(r-1)})^T\Delta^{(r-1)}}\right) z = \frac{((s^{(r-1)})^T z)^2}{(s^{(r-1)})^T\Delta^{(r-1)}} \leq 0. \tag{4.31}$$

Da $z \neq 0$, er (4.30) kun lig 0, hvis $L^T z \propto L^T\Delta^{(r-1)}$. I så fald er $z \propto \Delta^{(r-1)}$, hvormed (4.31) er strengt mindre end 0. Altså gælder det, at

$$z^T H^{(r)} z = z^T\left(H^{(r-1)} - \frac{H^{(r-1)}\Delta^{(r-1)}(\Delta^{(r-1)})^T H^{(r-1)}}{(\Delta^{(r-1)})^T H^{(r-1)}\Delta^{(r-1)}}\right) z + z^T\left(\frac{s^{(r-1)}(s^{(r-1)})^T}{(s^{(r-1)})^T\Delta^{(r-1)}}\right) z < 0, \tag{4.32}$$

hvormed $H^{(r)}$, og dermed også $B^{(r)}$, er negativ definit for alle $r$, hvis $(s^{(r-1)})^T\Delta^{(r-1)} < 0$ for alle $r \geq 1$. Vi lader da den $r+1$'te opdatering af $\theta$ være givet ved

$$\theta^{(r+1)} = \theta^{(r)} - B^{(r)} S(\theta^{(r)}; y). \tag{4.33}$$

Denne metode, hvor vi opdaterer den inverse Hessematrix ved (4.28), er kendt som *Broyden-Fletcher-Goldfarb-Shanno-algoritmen (BFGS-algoritmen)* [6, s. 55] og er implementeret i R-funktionen `optim` [28], som dermed kan benyttes til at maksimere log-likelihoodfunktionen i (4.18). Hvis vi ikke specificerer scorefunktionen, approksimeres denne ved brug af endnu en finite differences-metode. Yderligere er standarden for opnået konvergens i `optim`, at ændringen i værdien af objektfunktionen er mindre end $\epsilon(|\ell(\theta^{(r)}; y)| + \epsilon)$ for et $\epsilon > 0$ [28].

Idet log-likelihoodfunktioner baseret på mixturtætheder ofte kan være multimodale, kan der eksistere flere nulpunkter for scorefunktionen, hvormed vi kan risikere blot at finde frem til et lokalt maksimum for $\ell(\theta; y)$. Det kan derfor være nødvendigt at benytte flere forskellige startværdier, $\theta^{(0)}$, for algoritmen, hvoraf vi slutteligt bør udvælge det $\theta$, som giver den største

værdi af log-likelihoodfunktionen. Dog er vi ikke sikret, at $(s^{(r)})^T\Delta^{(r)} < 0$ for alle $r$, hvormed vi kan risikere, at en iteration af BFGS-algoritmen ikke er veldefineret, eller at nogle af approksimationerne af den inverse Hessematrix ikke er negativ definitte, hvorfor en opdatering af $\theta$ ikke resulterer i en større værdi af log-likelihoodfunktionen.

### 4.3.2 Marquardt-algoritmen

Vi kan i stedet betragte en udvidelse af quasi Newton Raphson-algoritmen, nærmere bestemt en såkaldt *modificeret Marquardt-algoritme* [26, s. 167]. Denne algoritme opdaterer $\theta$ ved

$$\theta^{(r+1)} = \theta^{(r)} - (\tilde{H}^{(r)})^{-1}S^{(r)}, \tag{4.34}$$

hvor $S^{(r)}$ er en finite differences-approksimation af $S(\theta^{(r)}; y)$, og $\tilde{H}^{(r)}$ er givet som $H^{(r)}$ i (4.29) med undtagelse af diagonalindgangene, som i stedet er givet ved

$$\tilde{H}^{(r)}_{jj} = H^{(r)}_{jj} - \lambda_1\left((1-\lambda_2)|H^{(r)}_{jj}| - \lambda_2 \mathrm{tr}(H^{(r)})\right), \tag{4.35}$$

for $j = 1, \ldots, n_\theta$ og $\lambda_1, \lambda_2 \in \mathbb{R}$ [26, s. 167]. Yderligere vælges $\lambda_1$ og $\lambda_2$ som udgangspunkt begge til 0.01 i hver iteration. Disse værdier kan da justeres, hvis $\tilde{H}^{(r)}$ ikke er negativ definit, således at vi opnår en negativ definit matrix, som har en tilpas lille afvigelse fra $H^{(r)}$.

Denne metode sikrer os, at approksimationen af Hessematricen bliver negativ definit, uanset om $(s^{(r)})^T\Delta^{(r)} < 0$ for alle $r$ eller ej. Den ønskede konvergens er da opnået, når de følgende tre trin er opfyldt på samme tid for tilpas små $\epsilon_b, \epsilon_l, \epsilon_g > 0$.

(i) $(\theta^{(r)} - \theta^{(r-1)})^T(\theta^{(r)} - \theta^{(r-1)}) \leq \epsilon_b$,

(ii) $|\ell(\theta^{(r)}; y) - \ell(\theta^{(r-1)}; y)| \leq \epsilon_l$,

(iii) $\frac{1}{n_\theta} S(\theta^{(r)}; y)^T (H^{(r)})^{-1} S(\theta^{(r)}; y) \leq \epsilon_g$ [27, s. 10].

Bemærk, at vi dog stadig bør benytte algoritmen for et gitter af startværdier.

For at sikre os at approksimationerne af $\Sigma$ og $\Psi$ i (4.18) er positiv definitte i hver iteration, når vi bruger en quasi Newton Raphoson-algoritme, kan vi benytte en omparametrisering. Hvis eksempelvis både $\Sigma$ og $\Psi$ er diagonalmatricer, kan vi omparametrisere deres diagonalelementer, således at vi, i stedet for at tilpasse en model med hensyn til parameteren $\theta_i$, tilpasser med hensyn til parameteren $s = \log(\theta_i)$. Dette gør vi ved at erstatte $\theta_i$ i log-likelihoodfunktionen med $\exp(s)$, som altid er positiv for den nye parameter $s$.

Marquardt-algoritmen for tilpasning af en GMM er i R implementeret i funktionen `hlme` fra `lcmm`-pakken [27].

**Eksempel 4.2**
Vi vil i dette eksempel betragte et simulationsstudie, hvor vi ønsker at estimere parametrene i en GMM ved brug af både `optim`- og `hlme`-funktionen i R, som benytter henholdsvis BFGS-algoritmen og den modificerede Marquardt-algoritme. Når vi benytter BFGS-algoritmen, estimerer vi yderligere de førsteafledede ved en finite differences-metode. Vi lader den sande model, som består af to grupper med fem datapunkter for hvert individ, være givet som i (4.12), (4.13) og (4.15), hvor

$$\alpha_1 = (8, 2, 4, -8, 10)^T, \quad \alpha_2 = (10, -2, 4, -2, 5)^T, \quad \Gamma_1 = \Gamma_2 = 0, \quad \Gamma_c = 0 \tag{4.36}$$

og hvor $\zeta_i \sim N_5(0, \Psi)$, $\varepsilon_i \sim N_5(0, \Sigma)$, $\Psi = \text{diag}(2, 0, 0, 0, 0)$ og $\Sigma = I_5$. Yderligere er $\Lambda$ valgt som i (4.4), hvormed

$$\Lambda\alpha_1 = (8, 10, 16, 14, 22)^T \quad \text{og} \quad \Lambda\alpha_2 = (10, 8, 10, 10, 15)^T. \tag{4.37}$$

Dette betyder, at skæringen for gruppe 1 er $\alpha_{11} = 8$, samt at hældningen for den $m$'te periode er givet ved $\sum_{s=2}^{m+1} \alpha_{1s} = (\Lambda\alpha_1)_{m+1} - (\Lambda\alpha_1)_m$, for $m = 1, \ldots, 4$.

Vi simulerer data fra denne model, hvor vi lader stikprøven bestå af $d = 100$ individer, som tilhører hver gruppe med sandsynlighed 0.5. Dette giver en gruppeinddeling på henholdsvis 57 og 43 individer i de to grupper, hvormed

$$\alpha_c = \log\left(\frac{0.43}{0.57}\right) = -0.28. \tag{4.38}$$

For at vælge de samme startværdier til begge funktioner benytter vi en metode implementeret i `hlme`, hvor vi først tilpasser en model med én latent gruppe og får parameterestimaterne $\hat{\theta}^{K=1}$. Herefter lader vi $\alpha_c^{(0)} = 0$, $\Sigma^{(0)} = \hat{\Sigma}^{K=1}$ og $\Psi^{(0)} = \hat{\Psi}^{K=1}$, samt de resterende startværdier være givet ved

$$\begin{aligned} \alpha_1^{(0)} &= \hat{\alpha}^{K=1} + \left(1 - \frac{3}{2}\right)\sqrt{\text{Var}(\hat{\alpha}^{K=1})}, \\ \alpha_2^{(0)} &= \hat{\alpha}^{K=1} + \left(2 - \frac{3}{2}\right)\sqrt{\text{Var}(\hat{\alpha}^{K=1})} \quad \text{[27, s. 17]}. \end{aligned} \tag{4.39}$$

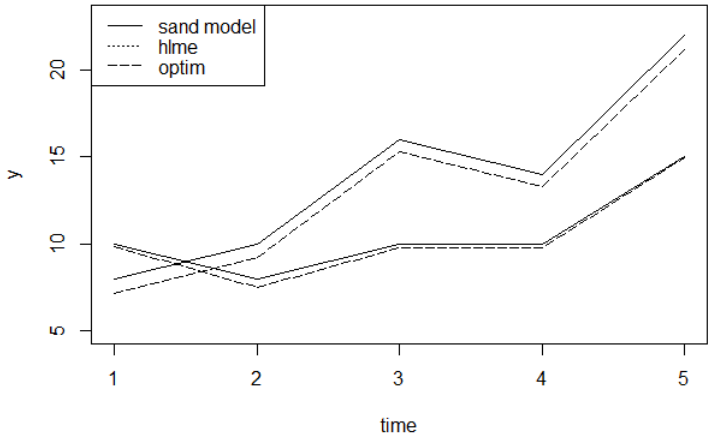
Da kan vi efterfølgende tilpasse en GMM ved brug af de to R-funktioner. Resultaterne af dette kan ses i Tabel 4.1, hvor fejlen beskriver afstanden $\|\theta - \hat{\theta}\|$.

| | iterationer | tid i ms | log-likelihood | gruppestørrelser | fejl |
|---|---|---|---|---|---|
| `hmle` | 11 | 373 | -926.95 | 43, 57 | 4.92 |
| `optim` | 21 | 17125 | -926.96 | 43, 57 | 4.91 |

***Tabel 4.1:*** Tabel over resultater for de to R-funktioner.

Vi ser, at `hlme`-funktionen er en del hurtigere end `optim`-funktionen, men at de to funktioner stort set har fundet samme maksimum for log-likelihoodfunktionen. Det er vigtigt at bemærke, at algoritmerne har forskellige konvergenskriterier, samt at vi har brugt funktionernes standardtolerancer, hvilket kan have påvirket resultaterne. Vi ser dog på de prædikterede gruppestørrelser, samt på Figur 4.3, at begge metoder giver gode resultater. Bemærk, at kurverne for de to tilpassede modeller ligger oven i hinanden, hvorfor de er svære at adskille på Figuren.

Hvis vi i stedet vælger at simulere indgangene i $\alpha_1$ og $\alpha_2$ som uafhængige realiseringer af en uniform fordeling på intervallet $[-10, 10]$, kan vi benytte disse til at simulere $N = 100$ forskellige datasæt. Da vil vi få resultaterne i Tabel 4.2, hvor log-likelihood og standardafvigelse angiver gennemsnittet af maksimum af log-likelihoodfunktionen samt dennes standardafvigelse over de $N$ datasæt.




***Figur 4.3:*** Vækstkurver for den sande model og de to modeller, opnået ved henholdsvis `hlme` og `optim`.

| | iterationer | tid i ms | log-likelihood | standardafvigelse |
|---|---|---|---|---|
| `hmle` | 16.86 | 342 | -926.01 | 14.54 |
| `optim` | 27.33 | 22325 | -941.09 | 61.50 |

***Tabel 4.2:*** Tabel over resultater for de to R-funktioner for $N = 100$ forskellige datasæt.

Her ser vi, at `hlme` klart fungerer bedre end `optim` både med hensyn til tid og maksimering af log-likelihoodfunktionen. Årsagen til, at `optim` ikke finder samme maksimum, kan være at den ikke fungerer lige så godt med det valg af startværdier, som `hlme` benytter. Dertil skal det nævnes, at de to funktioners maksimum af log-likelihoodfunktionen var ens for 79 af tilfældene ud til fjerde decimal. Vi kan altså konkludere, at det er hurtigere at benytte `hlme`, men at vi har tiltro til, at `optim` vil give acceptable resultater, hvis vi benytter flere forskellige startværdier. Det er yderligere nemmere at bruge `hlme`, idet vi hverken selv skal vælge startværdier eller kode log-likelihoodfunktionen i (4.18) eller tætheden i (4.21). Koden til dette eksempel kan ses i Appendiks B.2. ◀

### 4.3.3 EM-algoritmen

Vi vil i dette afsnit beskrive *Expectation-Maximization-algoritmen (EM-algoritmen)*, som er en iterativ metode, der ofte benyttes til at approksimere et lokalt maksimum for en log-likelihoodfunktion i tilfældet, hvor vi har latente variable [29].

Vi lader nu $\eta = (\eta_1^T, \ldots, \eta_d^T)$, $C = (C_1, \ldots, C_d)$ og $x = [x_1 \cdots x_d]^T$. For at kunne estimere parametervektoren $\theta$ i log-likelihoodfunktionen i (4.18), skal vi benytte den *komplette likeli-*

*hoodfunktion* givet ved

$$\begin{aligned} L_{kom}(\theta; y, \eta, c) = f(y, \eta, c \mid x; \theta) &= \prod_{i=1}^{d} f(y_i, \eta_i, c_i \mid x_i; \theta) \\ &= \prod_{i=1}^{d} f(y_i \mid \eta_i, c_i, x_i; \theta) f(\eta_i \mid c_i, x_i; \theta) f(c_i \mid x_i; \theta) \\ &= \prod_{i=1}^{d} f(y_i \mid \eta_i; \theta) f(\eta_i \mid c_i, x_i; \theta) f(c_i \mid x_i; \theta). \end{aligned} \tag{4.40}$$

Dermed bliver den komplette log-likelihoodfunktion

$$\ell_{kom}(\theta; y, \eta, c) = \sum_{i=1}^{d} \left( \log f(y_i \mid \eta_i; \theta) + \log f(\eta_i \mid c_i, x_i; \theta) + \sum_{k=1}^{K} \mathbf{1}[c_i = k] \log \pi_{ik}(\theta) \right). \tag{4.41}$$

EM-algoritmen benytter denne log-likelihoodfunktion til at finde et estimat af $\theta$, som øger værdien af log-likelihoodfunktionen i (4.18) i hver iteration. Vi vælger indledningsvist en startværdi, $\theta^{(0)}$, og gennemfører herefter de følgende to trin for $r = 0, 1, 2, \ldots$, indtil den ønskede konvergens er opnået [29].

(i) Udregn den forventede værdi af $\ell_{kom}(\theta; y, \eta, c)$, baseret på $(\eta, C) \mid y, x \sim f(\cdot, \cdot \mid y, x; \theta^{(r)})$, givet ved

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \mathbb{E}_{(\eta, C)|y,x}[\ell_{kom}(\theta; y, \eta, C)] \\ &= \int \sum_{c} f(\eta, c \mid y, x; \theta^{(r)}) \ell_{kom}(\theta; y, \eta, c) \mathrm{d}\eta. \end{aligned} \tag{4.42}$$

(ii) Lad $\theta^{(r+1)}$ være det $\theta$ der maksimerer $Q(\theta, \theta^{(r)})$, altså

$$\theta^{(r+1)} = \arg\max_{\theta} Q(\theta, \theta^{(r)}). \tag{4.43}$$

Det næste parameterestimat findes altså ved at maksimere den forventede værdi af den komplette log-likelihoodfunktion, hvoraf navnet på algoritmen opstår. Bemærk, at vi kan maksimere (4.42) for de tre led i (4.41) enkeltvist, idet de ikke deler nogle fælles parametre i $\theta$. Derudover ser vi, at

$$\begin{aligned} f(\eta, c \mid y, x; \theta) = \frac{f(y, \eta, c \mid x; \theta)}{f(y \mid x; \theta)} &= \frac{f(y \mid \eta, c, x; \theta) f(\eta, c \mid x; \theta)}{\int \sum_{c'} f(y, \eta', c' \mid x; \theta) \mathrm{d}\eta'} \\ &= \frac{f(y \mid \eta; \theta) f(\eta, c \mid x; \theta)}{\int \sum_{c'} f(y \mid \eta'; \theta) f(\eta', c' \mid x; \theta) \mathrm{d}\eta'}, \end{aligned} \tag{4.44}$$

hvor

$$f(\eta, c \mid x; \theta) = \prod_{i=1}^{d} f(\eta_i, c_i \mid x_i; \theta) = \prod_{i=1}^{d} f(\eta_i \mid c_i, x_i; \theta) f(c_i \mid x_i; \theta), \tag{4.45}$$

hvormed $f(\eta, c \mid y, x; \theta)$ er en kendt tæthed givet $\theta$.

For at se at EM-algoritmen finder et estimat som øger værdien af log-likelihoodfunktionen i (4.18) i hver iteration, betragter vi det følgende resultat.

**Sætning**
For maksimering af log-likelihoodfunktionen i (4.18) ved brug af EM-algoritmen, gælder det for $r = 0, 1, \ldots,$ at

$$\ell(\theta^{(r+1)}; y) \geq \ell(\theta^{(r)}; y). \tag{4.46}$$

***Bevis:***
Det gælder, at log-likelihoodfunktionen baseret på $Y$ er givet ved

$$\ell(\theta; y) = \log f(y \mid x; \theta) = \log f(y, \eta, c \mid x; \theta) - \log f(\eta, c \mid y, x; \theta), \tag{4.47}$$

hvor sidste lighed kommer af Bayes' Sætning, hvormed

$$\begin{aligned} &\int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \ell(\theta; y) \mathrm{d}\eta \\ &= \int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \left( \log f(y, \eta, c \mid x; \theta) - \log f(\eta, c \mid y, x; \theta) \right) \mathrm{d}\eta. \end{aligned} \tag{4.48}$$

Idet $\int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \mathrm{d}\eta = 1$ får vi, at ovenstående kan skrives som

$$\begin{aligned} \ell(\theta; y) &= \int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \left( \log f(y, \eta, c \mid x; \theta) - \log f(\eta, c \mid y, x; \theta) \right) \mathrm{d}\eta \\ &= Q(\theta, \theta^{(r)}) - \int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \log f(\eta, c \mid y, x; \theta) \mathrm{d}\eta. \end{aligned} \tag{4.49}$$

Det gælder derfor, at

$$\begin{aligned} \ell(\theta; y) - \ell(\theta^{(r)}; y) = Q(\theta, \theta^{(r)}) &- Q(\theta^{(r)}, \theta^{(r)}) \\ &- \int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \log \left( \frac{f(\eta, c \mid y, x; \theta)}{f(\eta, c \mid y, x; \theta^{(r)})} \right) \mathrm{d}\eta, \end{aligned} \tag{4.50}$$

hvor vi ifølge Jensens Ulighed [7, s. 561] kan omskrive sidste led til

$$\begin{aligned} &\int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \log \left( \frac{f(\eta, c \mid y, x; \theta)}{f(\eta, c \mid y, x; \theta^{(r)})} \right) \mathrm{d}\eta \\ &\leq \log \left( \int \sum_c f(\eta, c \mid y, x; \theta^{(r)}) \frac{f(\eta, c \mid y, x; \theta)}{f(\eta, c \mid y, x; \theta^{(r)})} \mathrm{d}\eta \right) \\ &= \log \left( \int \sum_c f(\eta, c \mid y, x; \theta) \mathrm{d}\eta \right) = \log(1) = 0, \end{aligned} \tag{4.51}$$

idet logaritmefunktionen er konkav. Dermed gælder det for $\theta = \theta^{(r+1)}$, at

$$\ell(\theta^{(r+1)}; y) - \ell(\theta^{(r)}; y) \geq Q(\theta^{(r+1)}, \theta^{(r)}) - Q(\theta^{(r)}, \theta^{(r)}), \tag{4.52}$$

som er ikke-negativ, idet $\theta^{(r+1)}$ maksimerer $Q(\theta, \theta^{(r)})$ jævnfør (4.43). Altså er

$$\ell(\theta^{(r+1)}; y) \geq \ell(\theta^{(r)}; y). \tag{4.53}$$

■

Som konsekvens af dette resultat, er EM-algoritmen numerisk stabil, hvilket betyder, at en iteration i algoritmen ikke vil gøre log-likelihoodfunktionen mindre. Vi er dog ikke garanteret at finde frem til et globalt maksimum, hvorfor det tilsvarende quasi Newton Raphson-algoritmerne kan være nødvendigt at benytte algoritmen gentagne gange for forskellige startværdier, $\theta^{(0)}$ [29].

Vi er heller ikke garanteret, at maksimering af $Q(\theta, \theta^{(r)})$ i (4.42) er mindre tungt end maksimeringen af log-likelihoodfunktionen i (4.18). Vi ser dog, at hvis vi begrænser en GMM ved at fjerne $\zeta_i$, så bliver EM-algoritmen forholdsvis simpel, hvilket vi illustrerer i det følgende eksempel.

**Eksempel 4.3** *(Gaussian Mixture model)*
Antag, at $\eta_i = \sum_{k=1}^{K} \alpha_k \mathbf{1}[c_i = k]$, samt at vi for simpelthedens skyld også antager, at vi ikke har kovariater i data. Da vil det jævnfør (4.17) gælde, at

$$f(y_i; \theta) = \sum_{k=1}^{K} \pi_{ik} f(y_i \mid c_i = k; \theta), \tag{4.54}$$

hvor $f(y_i \mid c_i = k; \theta)$ er en normalfordelingstæthed med ukendte parametre $\alpha_k$ og $\Sigma$, og da vi ikke har nogle kovariater, gælder det, at $\pi_{ik}$ er ens for alle $i$, og vi noterer dem blot ved $\pi_{1k}$ for hvert $k$. I dette tilfælde reduceres en GMM altså til en *Gaussian Mixture model*, som ofte benyttes som en unsupervised klassificeringsmetode [3]. Parametrene i modellen består altså af $\alpha_1, \ldots \alpha_K, \pi_{11}, \ldots \pi_{1K}$ og $\Sigma$, og vi opfatter ikke længere $\eta_i$ som latent. Dermed gælder det, at

$$\begin{aligned} Q(\theta, \theta^{(r)}) &= \sum_{c} f(c \mid y; \theta^{(r)}) \ell_{kom}(\theta; y, c) \\ &= \sum_{c} f(c \mid y; \theta^{(r)}) \sum_{i=1}^{d} \log \big( f(y_i \mid c_i; \theta) f(c_i; \theta) \big) \\ &= \sum_{i=1}^{d} \sum_{c_i} f(c_i \mid y_i; \theta^{(r)}) \log \big( f(y_i \mid c_i; \theta) f(c_i; \theta) \big) \\ &= \sum_{i=1}^{d} \sum_{k=1}^{K} f(c_i = k \mid y_i; \theta^{(r)}) \left( \log f(y_i \mid c_i = k; \theta) + \log \pi_{1k} \right), \end{aligned} \tag{4.55}$$

hvor

$$f(c_i = k \mid y_i; \theta^{(r)}) = \frac{f(y_i \mid c_i = k; \theta^{(r)}) \pi_{1k}^{(r)}}{f(y_i; \theta^{(r)})}, \tag{4.56}$$

hvormed vi har udført det første trin i EM-algoritmen. I det andet trin skal vi maksimere $Q(\theta, \theta^{(r)})$ i forhold til $\theta$, hvor vi for $\alpha_k$, for $k = 1, \ldots, K$, får, at

$$\frac{\partial}{\partial \alpha_k} Q(\theta, \theta^{(r)}) = \sum_{i=1}^{d} f(c_i = k \mid y_i; \theta^{(r)}) \Sigma^{-1}(y_i - \Lambda \alpha_k)\Lambda = 0, \tag{4.57}$$

som løses ved

$$\alpha_k^{(r+1)} = \Lambda^{-1} \frac{\sum_{i=1}^{d} y_i f(c_i = k \mid y_i; \theta^{(r)})}{\sum_{i=1}^{d} f(c_i = k \mid y_i; \theta^{(r)})}. \tag{4.58}$$

Yderligere gælder det, at $\log|\Sigma| = -\log|\Sigma^{-1}|$ og ved at maksimere med hensyn til præcisionsmatricen $\Sigma^{-1}$, får vi af [25, s. 9-10], at

$$\frac{\partial}{\partial \Sigma^{-1}} Q(\theta, \theta^{(r)}) = \sum_{i=1}^{d} \sum_{k=1}^{K} f(c_i = k \mid y_i; \theta^{(r)}) \frac{1}{2}(\Sigma - (y_i - \Lambda \alpha_k)(y_i - \Lambda \alpha_k)^T) = 0, \tag{4.59}$$

hvormed

$$\Sigma^{(r+1)} = \frac{\sum_{i=1}^{d} \sum_{k=1}^{K} f(c_i = k \mid y_i; \theta^{(r)})(y_i - \Lambda \alpha_k)(y_i - \Lambda \alpha_k)^T}{\sum_{i=1}^{n} \sum_{k=1}^{K} f(c_i = k \mid y_i; \theta^{(r)})}. \tag{4.60}$$

For slutteligt at bestemme $\pi_{1k}^{(r+1)}$, husker vi, at disse skal overholde begrænsningen $\sum_{k=1}^{K} \pi_{1k} - 1 = 0$. Vi benytter derfor en *Lagrange multiplier*, $\lambda$, i maksimeringsproblemet [21, s. 310], hvormed vi altså ønsker at løse

$$\begin{aligned} \frac{\partial}{\partial \pi_{1k}} \left( Q(\theta, \theta^{(r)}) - \lambda \left( \sum_{k=1}^{K} \pi_{1k} - 1 \right) \right) = \sum_{i=1}^{d} \frac{1}{\pi_{1k}} f(c_i = k \mid y_i; \theta^{(r)}) - \lambda = 0 \Leftrightarrow \\ \sum_{i=1}^{d} f(c_i = k \mid y_i; \theta^{(r)}) = \lambda \pi_{1k}. \end{aligned} \tag{4.61}$$

Vi bemærker heraf, at

$$\sum_{i=1}^{d} \sum_{k=1}^{K} f(c_i = k \mid y_i; \theta^{(r)}) = \lambda \sum_{k=1}^{K} \pi_{1k}, \tag{4.62}$$

hvormed $\lambda = d$, da summerne over de $K$ grupper begge giver 1. Dermed kan vi altså udregne

$$\pi_{1k}^{(r+1)} = \frac{1}{d} \sum_{i=1}^{d} f(c_i = k \mid y_i; \theta^{(r)}), \tag{4.63}$$

hvormed vi har fundet alle elementerne i $\theta^{(r+1)}$ og dermed maksimeret $Q(\theta, \theta^{(r)})$. ◀

### 4.3.4 Sammenligning af algoritmer

I R er der implementeret flere funktioner til tilpasning af en GMM [33]. Vi har tidligere brugt `hlme`-funktionen, og yderligere er EM-algoritmen implementeret i funktionen `FLXMRlmm` fra pakken `flexmix` [12]. I begge disse pakker er der implementeret mulighed for at bruge algoritmerne gentagne gange for forskellige startværdier. Vi vil nu lave et simulationsstudie for at undersøge disse funktioners konvergenshastigheder. For at gøre dette simulerer vi $N = 100$ datasæt for forskellige værdier af $K$. I alle disse datasæt er $T = 5$ og $d = 100$. Yderligere lader vi

$$\Gamma_1 = \cdots = \Gamma_K = 0, \quad \Gamma_c = 0, \quad \Psi = \text{diag}(5, 0, 0, 0, 0), \quad \Sigma = 5I_5. \tag{4.64}$$

For hvert datasæt simulerer vi indgangene i $\alpha_k$, for $k = 1, \ldots, K$, som uafhængige realiseringer af en uniform fordeling på intervallet $[-10, 10]$, samt inddeler individerne i hver gruppe med sandsynlighed $1/K$. Herefter tilpasser vi en GMM med henholdsvis `hlme` og `FLXMRlmm` på hvert datasæt. I Tabel 4.3 ses gennemsnittene af antal iterationer, tid og maksimum af log-likelihoodfunktionerne for begge funktioner over de $N$ datasæt samt standardafvigelsen af log-likelihoodfunktionerne. Koden brugt til denne simulation kan ses i Appendiks B.3.

| | | iterationer | tid i ms | log-likelihood | standardafvigelse |
|---|---|---|---|---|---|
| $K = 2$ | `FLXMRlmm` | 35.98 | 1881 | -1660.07 | 21.19 |
| | `hlme` | 11.66 | 284 | -1660.09 | 21.19 |
| $K = 3$ | `FLXMRlmm` | 48.22 | 3509 | -1693.76 | 26.71 |
| | `hlme` | 16.76 | 795 | -1691.79 | 22.24 |
| $K = 4$ | `FLXMRlmm` | 51.40 | 4976 | -1712.37 | 23.98 |
| | `hlme` | 21.19 | 1829 | -1713.12 | 26.91 |
| $K = 5$ | `FLXMRlmm` | 66.33 | 7857 | -1721.95 | 24.14 |
| | `hlme` | 29.27 | 4065 | -1726.12 | 27.47 |

***Tabel 4.3:*** Tabel over resultater for de to R-funktioner for $K = 2, 3, 4$ og 5.

Vi ser, at `hlme` generelt er hurtigere for alle værdier af $K$ både med hensyn til iterationer og tid brugt. Eksempelvis bruger denne funktion i gennemsnit omkring halvt så lang tid som `FLXMRlmm` for $K = 5$. Yderligere ser vi, at funktionerne finder tæt på samme værdi for log-likelihoodfunktionerne for hvert $K$ samt tæt på ens standardafvigelser. Vi erfarer yderligere, at `hlme` ofte ikke kunne konvergere, specielt for $K = 4$ og $K = 5$, hvilket ikke var et problem for `FLXMRlmm`.

Alternativt kan vi også undersøge de to funktioners konvergenshastigheder, når $d$ bliver større for et fastholdt $K$. Vi lader derfor $K = 2$ og gentager de førnævnte simuleringer $N = 100$ gange for $d \in \{50, 100, 200, 500\}$. Dermed er det samlede antal observationer for hvert datasæt $5d$.

| | | iterationer | tid i ms | log-likelihood | standardafvigelse |
|---|---|---|---|---|---|
| $d = 50$ | `FLXMRlmm` | 30.39 | 940 | -825.18 | 14.11 |
| | `hlme` | 11.42 | 170 | -825.16 | 14.17 |
| $d = 100$ | `FLXMRlmm` | 28.78 | 1554 | -1662.13 | 16.36 |
| | `hlme` | 11.83 | 280 | -1662.00 | 16.51 |
| $d = 200$ | `FLXMRlmm` | 27.28 | 2873 | -3328.62 | 34.97 |
| | `hlme` | 11.81 | 479 | -3328.03 | 35.11 |
| $d = 500$ | `FLXMRlmm` | 30.54 | 8687 | -8315.78 | 34.97 |
| | `hlme` | 13.84 | 1268 | -8315.73 | 35.11 |

***Tabel 4.4:*** Tabel over resultater for de to R-funktioner for $d \in \{50, 100, 200, 500\}$.

Resultatet af dette kan ses i Tabel 4.4, hvor vi ser, at størrelsen af $d$ ikke ændrer meget på gennemsnittet af antallet af iterationer, som algoritmerne bruger. Tilsvarende resultaterne i tidligere undersøgelse ser vi, at tiden for `FLXMRlmm` bliver meget større end `hlme`, samt at log-likelihoodfunktionen og standardafvigelsen er cirka det samme. Bemærk, at første række i Tabel 4.3 og anden række i Tabel 4.4 er ækvivalente undersøgelser. Vi bemærker yderligere, at denne sammenligning ikke nødvendigvis er helt meningsfuld, idet de to algoritmer hverken har samme konvergenskritier eller startværdier.

En fordel ved at bruge Marquardt-algoritmen frem for EM-algoritmen er, at estimation af standardfejl for estimaterne er en del af selve algoritmen. Dette gøres, idet Marquardt-algoritmen approksimerer den inverse Hessematrix, hvilket svarer til minus den inverse af den observerede informationsmatrix, hvormed vi sætter

$$\widehat{\mathrm{Var}}(\hat{\theta}) = -B^{(r)}. \tag{4.65}$$

De tilsvarende metoder til at finde standardfejl for EM-algoritmen har vist sig at give meget varierende resultater [5, s. 34-35].

Baseret på denne fordel og de ovenstående resultater er `hlme` altså at foretrække.

## 4.4 Valg af antal latente grupper

Når vi opstiller en GMM, skal vi indledningsvist vælge antallet af latente grupper, $K$. Dette valg kan somme tider være oplagt, alt efter hvilken undersøgelse man ønsker at lave. Det kan dog stadig være interessant at undersøge, om der findes et andet mere passende antal grupper, der beskriver data bedre. Vi skal derfor i dette afsnit undersøge forskellige metoder, som kan give en indikation på et passende antal grupper, baseret på hvordan elevernes CBM-scorer har udviklet sig over de fire perioder. Overordnet set vil vi beskæftige os med to forskellige tilgange nemlig at benytte udvalgte informationskriterier og likelihood ratio-test, som er de to mest brugte tilgange til modelselektion for en GMM [4, s. 3]. Begge tilgange består i at estimere log-likelihoodfunktionen, $\ell_K(\theta; y)$, for flere forskellige værdier af $K$, hvorefter vi kan sammenligne de tilhørende modeller.

### 4.4.1 Informationskriterier

En simpel måde, hvorpå vi kan sammenligne modeller for forskellige $K$, er blot ved at sammenligne de maksimerede værdier af de tilhørende log-likelihoodfunktioner. Vi kan dog i så fald risikere at favorisere modeller med unødvendig høj kompleksitet, altså med rigtig mange latente grupper, som kan komme til at overfitte data, hvormed vi også bør tage højde for antallet af parametre i modellen. Alternativt kan vi derfor udregne *Akaike informationskriteriet (AIC)* for modellerne, som er givet ved

$$\mathrm{AIC}(K) = -2\ell_K(\hat{\theta}_K; y) + 2m(K), \tag{4.66}$$

hvor $m(K)$ er antallet af uafhængige parametre i modellen med $K$ latente grupper, og hvor $\hat{\theta}_K$ maksimerer $\ell_K$ [32, s. 5-6]. I denne sammenhæng vil kandidaten til den bedste model være modellen med laveste AIC, hvorfor modellerne *straffes* for at have mange parametre. Simulationsstudier har vist, at dette informationskriterie har en tendens til at favorisere modeller med flere latente grupper, end den sande model er specificeret ved [32, s. 5]. Vi kan derfor alternativt udregne *Bayesian informationskriteriet (BIC)* for modellerne givet ved

$$\mathrm{BIC}(K) = -2\ell_K(\hat{\theta}_K; y) + \log(n)m(K), \tag{4.67}$$

som øger straffen for at have mange parametre og dermed også for at have mange grupper. Dette informationskriterie har dog i stedet en tendens til at favorisere modeller med færre latente grupper end den sande model [32, s. 5-6].

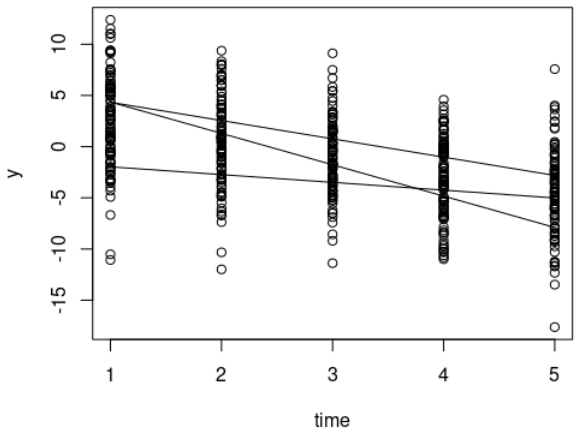
**Eksempel 4.4**
Vi laver nu et simulationsstudie for at undersøge valget af model baseret på ovenstående. Vi simulerer derfor $d = 100$ individer fordelt på tre grupper med sandsynlighed $1/3$, hvor $T = 5$ og

$$\begin{aligned} \alpha_1 &= (-1.97, -0.76, 0, 0, 0)^T, \\ \alpha_2 &= (4.35, -3.06, 0, 0, 0)^T, \\ \alpha_3 &= (4.33, -1.79, 0, 0, 0)^T. \end{aligned} \tag{4.68}$$

Yderligere lader vi

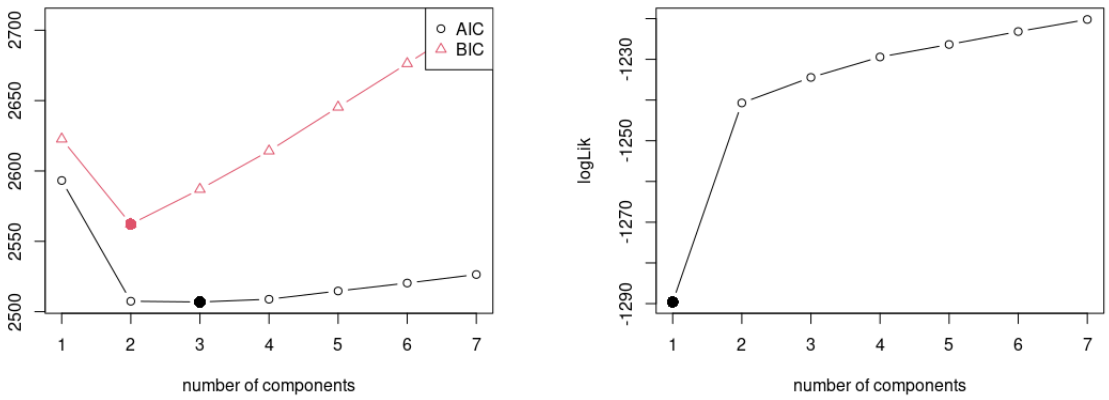
$$\Gamma_1 = \Gamma_2 = \Gamma_3 = 0, \quad \Gamma_c = 0, \quad \Psi = \mathrm{diag}(3, 0, \ldots, 0), \quad \Sigma = 2I_5, \tag{4.69}$$

hvilket resulterer i de lineære vækstkurver, som ses på Figur 4.4.




***Figur 4.4:*** De sande vækstkurver og det simulerede datasæt.

Ved brug af EM-algoritmen tilpasser vi nu en GMM på datasættet for $K = 1, \ldots, 7$, og udregner herefter maksimum af log-likelihoodfunktionen, AIC og BIC. Resultatet af dette kan ses på Figur 4.5, hvor vi ser, at AIC og BIC minimeres ved henholdsvis $K = 3$ og $K = 2$, hvormed AIC altså favoriserer en model med flere grupper end BIC. Det gælder yderligere, at maksimum af log-likelihoodfunktionen vokser, når $K$ bliver større, samt at denne stigning er mest markant mellem $K = 1, 2$. Ved at betragte tre latente grupper, i stedet for to, stiger log-likelihoodfunktionen kun ganske lidt, og vi skal derfor overveje, om den øgede kompleksitet retfærdiggør den marginalt bedre tilpassede model. R-koden til dette eksempel kan ses i Appendiks B.4.




***Figur 4.5:*** AIC, BIC og maksimum af log-likelihoodfunktionen for $K = 1, \ldots, 7$.

### 4.4.2 Likelihood ratio-test

Hvis vi i stedet ønsker at udføre en hypotesetest for det rette antal af grupper, frem for blot at samle evidens ud fra AIC og BIC, kan vi benytte os af en likelihood ratio-test foreslået af Lo, Mendell og Rubin. Denne test kan give evidens for, om data følger en GMM med enten $K_0$ eller $K_1$ grupper, hvor $K_0 < K_1$ [15, s. 767]. Typisk vil vi teste for, om vi skal benytte enten $K-1$ eller $K$ grupper for forskellige værdier af $K$.

Vi lader indledningsvist $L_{K_0}(\theta_{K_0}; y)$ være likelihoodfunktionen for en GMM med $K_0$ grupper, med tilhørende parametervektor $\theta_{K_0}$, og ligeledes for $L_{K_1}(\theta_{K_1}; y)$ og $\theta_{K_1}$. Betragt nu hypoteserne

$$\begin{aligned} \mathcal{H}_0 : &\quad K = K_0, \\ \mathcal{H}_1 : &\quad K = K_1. \end{aligned} \tag{4.70}$$

Da får vi, at teststørrelsen for en likelihood ratio-test for ovenstående nulhypotese er givet ved

$$\mathrm{LR} = -2 \log \left( \frac{L_{K_0}(\hat{\theta}_{K_0}; y)}{L_{K_1}(\hat{\theta}_{K_1}; y)} \right), \tag{4.71}$$

hvor $\hat{\theta}_{K_0}$ og $\hat{\theta}_{K_1}$ er estimerede værdier af henholdsvis $\theta_{K_0}$ og $\theta_{K_1}$ for de to modeller [15, s. 771]. Det gælder dog ikke, at denne teststørrelse asymptotisk følger en $\chi^2$-fordeling med $m(K_1) - m(K_0)$ frihedsgrader, som en sædvanlig likelihood ratio-teststørrelse gør. Dette skyldes, at under $\mathcal{H}_0$ vil nogle af mikstursandsynlighederne, $\pi_{ik}$, være 0, hvorfor $\alpha_c$ vil ligge på randen af parameterrummet, hvormed visse regularitetsbetingelser ikke er opfyldt [17, s. 343]. I stedet er denne teststørrelse asymptotisk fordelt som en vægtet sum af $\widetilde{m} = m(K_1) + m(K_0)$ uafhængige $\chi^2(1)$-fordelte variable, sådan at

$$\mathrm{LR} \xrightarrow{\mathcal{D}} \lambda_1 Z_1 + \cdots + \lambda_{\widetilde{m}} Z_{\widetilde{m}}, \tag{4.72}$$

når $d \to \infty$, hvor $Z_i \sim \chi^2(1)$ for $i = 1, \ldots, \widetilde{m}$ [15, s. 772]. Her gælder det, at $\lambda_i$ er den $i$'te egenværdi til matricen

$$\begin{bmatrix} -B_1(\hat{\theta}_{K_1}) A_1(\hat{\theta}_{K_1})^{-1} & -B_{10}(\hat{\theta}_{K_1}, \hat{\theta}_{K_0}) A_0(\hat{\theta}_{K_0})^{-1} \\ B_{10}(\hat{\theta}_{K_1}, \hat{\theta}_{K_0})^T A_1(\hat{\theta}_{K_1})^{-1} & B_0(\hat{\theta}_{K_0}) A_0(\hat{\theta}_{K_0})^{-1} \end{bmatrix}, \tag{4.73}$$

hvor

$$\begin{aligned} A_1(\theta_{K_1}) &= \mathbb{E}_Y \left[ \frac{\partial^2}{\partial \theta_{K_1} \partial \theta_{K_1}^T} \log L_{K_1}(\theta_{K_1}; Y) \right], \\ A_0(\theta_{K_0}) &= \mathbb{E}_Y \left[ \frac{\partial^2}{\partial \theta_{K_0} \partial \theta_{K_0}^T} \log L_{K_0}(\theta_{K_0}; Y) \right], \\ B_1(\theta_{K_1}) &= \mathbb{E}_Y \left[ \frac{\partial}{\partial \theta_{K_1}} \log L_{K_1}(\theta_{K_1}; Y) \frac{\partial}{\partial \theta_{K_1}^T} \log L_{K_1}(\theta_{K_1}; Y) \right], \\ B_0(\theta_{K_0}) &= \mathbb{E}_Y \left[ \frac{\partial}{\partial \theta_{K_0}} \log L_{K_0}(\theta_{K_0}; Y) \frac{\partial}{\partial \theta_{K_0}^T} \log L_{K_0}(\theta_{K_0}; Y) \right], \\ B_{10}(\theta_{K_1}, \theta_{K_0}) &= \mathbb{E}_Y \left[ \frac{\partial}{\partial \theta_{K_1}} \log L_{K_1}(\theta_{K_1}; Y) \frac{\partial}{\partial \theta_{K_0}^T} \log L_{K_0}(\theta_{K_0}; Y) \right]. \end{aligned} \tag{4.74}$$

Denne teststørrelse konvergerer forholdsvis langsomt mod den asymptotiske fordeling og er derfor kun brugbar for store stikprøver. Vi kan da i stedet betragte den justerede LR-teststørrelse givet ved

$$\mathrm{LR}_{adj} = \frac{\mathrm{LR}}{1 + ((m(K_1) - m(K_0)) \log n)^{-1}}, \tag{4.75}$$

som er hurtigere til at konvergere mod den asymptotiske fordeling, hvormed brugen af denne teststørrelse regnes for at være mere pålidelig end ved brug af teststørrelsen i (4.71) [15, s. 774].

Bemærk, at for at udlede analytiske udtryk for egenværdierne i (4.72) skal vi udregne de afledede og de dobbeltafledede af de to log-likelihoodfunktioner, hvilket kan være både besværligt og tidskrævende. Derudover har det også vist sig, at argumentationen for at LR asymptotisk følger denne fordeling, når $d \to \infty$, indeholder betydelige matematiske fejl [9]. For derfor at undgå denne asymptotiske fordeling kan vi i stedet bruge *parametrisk bootstrap* til at estimere $p$-værdien for hypotesetesten. Denne metode kaldes *Bootstrap likelihood ratio-test (BLRT)* og består af følgende trin [18, s. 229].

(i) Estimer to modeller med henholdsvis $K_0$ og $K_1$ grupper og udregn likelihood ratio-teststørrelsen i (4.71).

(ii) For $b = 1, \ldots, B$, simuler et datasæt, $y^*$, baseret på modellen med $K_0$ grupper, estimer to nye modeller med henholdsvis $K_0$ og $K_1$ grupper og udregn

$$\mathrm{LR}_b^* = -2 \log \left( \frac{L_{K_0}(\hat{\theta}_{K_0}^*; y^*)}{L_{K_1}(\hat{\theta}_{K_1}^*; y^*)} \right). \tag{4.76}$$

(iii) Estimer $p$-værdien for hypotesetesten ved

$$p \approx \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}[\mathrm{LR}_b^* \geq \mathrm{LR}]. \tag{4.77}$$

Bemærk, at $\mathrm{LR}_b^*$ er udregnet under antagelse af, at $\mathcal{H}_0$ er sand, hvorfor det er meningsfuldt at sammenligne disse med LR for at teste hypotesen.

Denne metode er udregningsmæssig tung, idet vi skal vælge en stor værdi af $B$ samt flere forskellige startværdier for hver tilpasning af de to modeller, for hvert $b$, for at få gode approksimationer. Trods dette, er det dog den fortrukne metode til at udføre likelihood ratio-test for en GMM [32, s. 6]. Denne udregningsmæssige udfordring betyder, at den typiske fremgangsmåde til at vælge antallet af grupper er først at undersøge forskellige værdier af $K$ ved hjælp af AIC og BIC, hvorefter vi laver BLRT for udvalgte værdier af $K$.

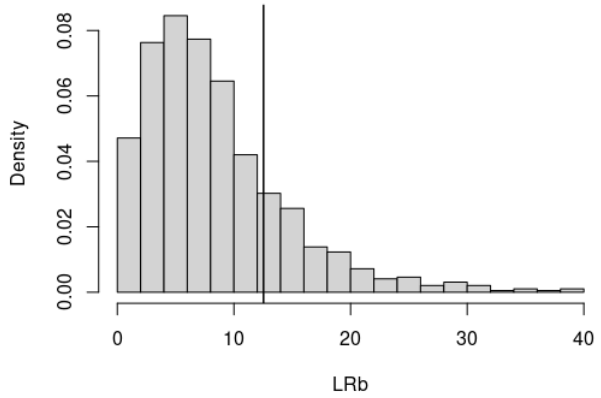
**Eksempel 4.4** *(fortsat)*
Vi husker, at AIC og BIC blev minimeret ved henholdsvis $K = 3$ og $K = 2$. Lad nu $K_0 = 2$ og $K_1 = 3$, hvormed vi får, at

$$\mathrm{LR} = 12.5349. \tag{4.78}$$

Vi simulerer nu $B = 1000$ datasæt baseret på modellen hvor $K = 2$, hvis parameterestimater er givet ved

$$\begin{aligned}\hat{\alpha}_1 &= (0.65, -0.29, -1.06, 0.38, -1.11)^T, \\ \hat{\alpha}_2 &= (5.44, -4.41, 2.39, -1.92, 2.19)^T,\end{aligned} \tag{4.79}$$

$\hat{\Psi} = \text{diag}(11.3524, 0, \ldots, 0)$ og $\hat{\Sigma} = 3.8940 I_5$. Sammenligner vi de simulerede datasæts likelihood-ratios med LR, får vi at $p \approx 0.191$, hvormed vi ikke kan forkaste $\mathcal{H}_0$. Fordelingen af $\text{LR}_b^*$ kan ses i Figur 4.6. Vi ser altså en indikation på, at vi vil få en bedre tilpasset model ved at bruge $K = 2$, selvom det sande antal grupper er tre. R-koden til dette eksempel kan ses i Appendiks B.4.




***Figur 4.6:*** Histogram over $\text{LR}_b^*$, for $b = 1, \ldots, 1000$, med LR markeret ved den lodrette linje.

◀

## 4.5 Dataanalyse ved brug af Growth Mixture modeller

Vi vil i dette afsnit tilpasse en GMM med $K = 3$ på hvert datasæt for herefter at sammenligne med resultaterne i Afsnit 3.5 både i forhold til estimaterne og gruppeinddelingerne. Derudover vil vi undersøge, om der er et mere optimalt valg af $K$ og i så fald tilpasse en model med dette $K$. R-koden til dette afsnit kan findes i Appendiks B.5.

### 4.5.1 Modeller med tre latente grupper med EM-algoritmen

Vi så i Afsnit 3.5 en indikation på, at den mest relevante stokastiske effekt at have med i en mixed model er eleveffekten. Vi ønsker derfor at tilpasse en GMM for $K = 3$ grupper, hvor vi lader variablen `id` indgå som den eneste stokastiske variabel. Derudover ønsker vi ikke at inkludere krydsfaktoren `gender:time` eller den stokastiske krydsfaktor `id:time` i modellen, sådan at resultaterne er mest muligt sammenlignelige med modellen i Afsnit 3.5. Vi får dermed

en model på formen

$$Y_i = \Lambda \left( \sum_{k=1}^{K} (\alpha_k + \Gamma_k x_{\texttt{gender}}) \mathbf{1}[c_i = k] + \zeta_i \right) + \varepsilon_i, \tag{4.80}$$

hvor $\Gamma_k = e_1 \gamma_k$ for standardvektoren $e_1$. Derudover er $\zeta_i \sim N_5(0, \Psi)$ og $\varepsilon_i \sim N_5(0, \sigma^2 I_5)$, hvor $\Psi = \text{diag}(\psi_{\texttt{id}}, 0, 0, 0, 0)$. Vi tillader altså hver elev at have en stokastisk skæring samt væksten og kønseffekten at være gruppespecifikke. Vi benytter da EM-algoritmen, implementeret i R-funktionen `FLXMRlmm`, til at estimere de relevante parametre for datasæt 1, med 1000 forskellige startværdier, hvormed vi får estimaterne i Tabel 4.5 samt de tilsvarende middelvækstkurver på Figur 4.7. På samme måde tilpasser vi også en GMM for datasæt 2, og parameterestimaterne og middelvækstkurverne for denne kan ses henholdsvis i Tabel 4.6 og på Figur 4.8.

`intervention` = 1:

| | gruppe 1 | gruppe 2 | gruppe 3 |
|---|---|---|---|
| $\alpha_{k1}$ | 6.8146 | 6.1655 | 8.5924 |
| $\alpha_{k2}$ | -0.1551 | 0.5216 | 2.1863 |
| $\alpha_{k3}$ | 4.1815 | 1.9301 | 1.7031 |
| $\alpha_{k4}$ | -4.2322 | 5.3162 | -0.1688 |
| $\alpha_{k5}$ | 0.6499 | -18.8117 | -3.3220 |
| $\gamma_k$ | 1.0925 | 6.1828 | 5.2878 |

***Tabel 4.5:*** Parameterestimater for datasæt 1 ved brug af EM-algoritmen.

`intervention` = 2:

| | gruppe 1 | gruppe 2 | gruppe 3 |
|---|---|---|---|
| $\alpha_{k1}$ | 6.2061 | 9.6882 | 11.3850 |
| $\alpha_{k2}$ | 0.9288 | -0.9922 | 4.9546 |
| $\alpha_{k3}$ | -0.1483 | 1.9914 | -3.4152 |
| $\alpha_{k4}$ | 1.9312 | 8.9768 | 0.1804 |
| $\alpha_{k5}$ | -2.3090 | -7.6473 | 0.9846 |
| $\gamma_k$ | 0.5260 | 7.1236 | 1.2751 |

***Tabel 4.6:*** Parameterestimater for datasæt 2 ved brug af EM-algoritmen.

Yderligere får vi for datasæt 1, at

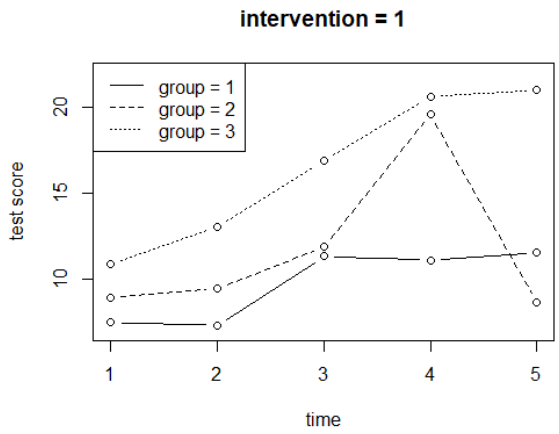
$$\hat{\psi}_{\texttt{id}} = 13.0132 \quad \text{og} \quad \hat{\sigma}^2 = 6.6463, \tag{4.81}$$

og tilsvarende for datasæt 2, at

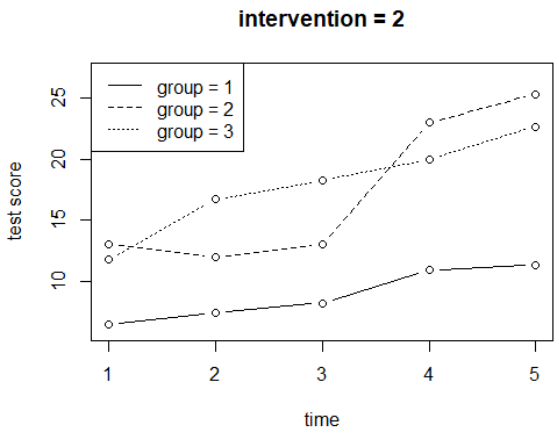
$$\hat{\psi}_{\texttt{id}} = 7.6645 \quad \text{og} \quad \hat{\sigma}^2 = 6.5430. \tag{4.82}$$

Bemærk, at den estimerede residualvarians for datasæt 1 udgør 32% af den sammelede varians, hvilket er tilsvarende modellen, der indeholder den systematiske gruppeeffekt, i Afsnit 3.5, hvor den udgør 33%. Yderligere gælder det for datasæt 2, at residualvariansen udgør 46% sammenlignet med 34% for den tilpassede mixed model. Derudover ser vi også, at elevvariansen er markant mindre end for de tilpassede mixed modeller. Dette kan skyldes den øgede kompleksitet af middelværdistrukturen, samt at modellen prædikterer hvilke elever, der skal tilhøre hvilken gruppe ud fra data.

I Tabel 4.5 ser vi, at gruppe 1 har en ændring i hældningen i anden periode på $\hat{\alpha}_{13} = 4.1815$ i forhold til den forrige periode. Dermed må det gælde, at testscorerne forventes at stige med $\hat{\alpha}_{12} + \hat{\alpha}_{13} = -0.1551 + 4.1815 = 4.0264$ point for denne gruppe i denne periode, idet $\sum_{s=1}^{m} \alpha_{k(s+1)}$ svarer til hældningen i den $m$'te periode. Yderligere er $\hat{\alpha}_{14} = -4.2322$, hvormed hældningen i tredje periode flader ud, idet $4.0264 - 4.2322 = -0.2058$. En oversigt over omregning af estimater kan ses i Appendiks A. Dette svarer til, hvad vi ser på den tilsvarende




*Figur 4.7:* Middelvækstkurverne for datasæt 1 ved brug af EM-algoritmen.




*Figur 4.8:* Middelvækstkurverne for datasæt 2 ved brug af EM-algoritmen.

middelvækstkurve i Figur 4.7. Det gælder altså, at vækstkurven for denne gruppe stiger i anden periode, hvor eleverne er blevet undervist i brøkregning.

Ved at udregne alle hældningerne for middelvækstkurverne for datasæt 1, som i Appendiks A, kan vi sammenligne disse med estimaterne i Tabel 3.2, hvormed vi opnår resultaterne i Tabel 4.7. Vi ser her, at hældningerne for kurverne for gruppe 1 og 3 generelt stemmer overens

| `intervention` = 1: | | | |
|---|---|---|---|
| | mixed model | GMM | afvigelse |
| $\delta^{12}_{\texttt{group1}}$ | -0.0242 | -0.1551 | 0.1793 |
| $\delta^{23}_{\texttt{group1}}$ | 3.5951 | 4.0264 | 0.4313 |
| $\delta^{34}_{\texttt{group1}}$ | 0.6284 | -0.2058 | 0.8342 |
| $\delta^{45}_{\texttt{group1}}$ | -0.3240 | 0.4441 | 0.7681 |
| $\delta^{12}_{\texttt{group2}}$ | 0.4779 | 0.5216 | 0.0437 |
| $\delta^{23}_{\texttt{group2}}$ | 4.3920 | 2.4517 | 1.9403 |
| $\delta^{34}_{\texttt{group2}}$ | 2.0302 | 7.7679 | 5.7377 |
| $\delta^{45}_{\texttt{group2}}$ | -0.0443 | -11.0438 | 10.9995 |
| $\delta^{12}_{\texttt{group3}}$ | 2.4517 | 2.1863 | 0.2654 |
| $\delta^{23}_{\texttt{group3}}$ | 3.1184 | 3.8894 | 0.7710 |
| $\delta^{34}_{\texttt{group3}}$ | 2.4572 | 3.7206 | 1.2634 |
| $\delta^{45}_{\texttt{group3}}$ | 0.1598 | 0.3986 | 0.2388 |

*Tabel 4.7:* Sammenligning af estimater for den tilpassede mixed model og den tilpassede GMM, ved brug af EM-algoritmen.

med hældningerne for kurverne for de lavt og højt præsterende elever. På Figur 4.7 ser vi ligeledes, at middelvækstkurverne for gruppe 1 og 3 ligner kurverne, som vi opnåede i Afsnit 3.5. Specielt gælder det, at kurverne for gruppe 1 og for de lavt præsterende elever ligner forskydninger af hinanden, som også bekræftes af resultaterne i Tabel 4.7. Derimod ser vi, at

kurven for gruppe 2 har en meget stort positiv hældning i tredje periode, efterfulgt af en stor negativ hældning i fjerde periode, som ikke stemmer overens med nogle af kurverne for den tilpassede mixed model.

Afvigelsen fra middelvækstkurverne for den tilpassede mixed model kan skyldes en ny gruppeinddeling, hvorfor vi ønsker at sammenligne disse inddelinger. I Tabel 4.8 ser vi, at modellen for datasæt 1 prædikterer syv elever til at ligge i gruppe 2, i modsætning til gruppen med de middel præsterende elever, baseret på de nationale tests, hvor der er 111 elever. Dette er altså årsagen til, at vækstkurven for gruppe 2 på Figur 4.7 er så forskellig fra kurven for de middel præsterende elever. På Figur 4.9 ser vi de observerede vækstkurver for de syv individer prædikteret til gruppe 2. Her bemærker vi, at alle disse elever har fået en høj testscore i fjerde test efterfulgt af en lav testscore i femte test, hvilket er en tendens, som modellen har opfanget. Yderligere ser vi i datasæt 2, at der er prædikteret 166 elever til gruppe 1, hvoraf 98 af dem tilhører den middel præsterende gruppe. Dette kan have en sammenhæng med, at middelvækstkurverne for de lavt og middel præsterende elever ligger meget tæt, jævnfør Figur 3.3. Bemærk, at modellen for dette datasæt ligeledes prædikterer få elever til at tilhøre gruppe 2, som tilsvarende datasæt 1 har en stor hældning i tredje periode.

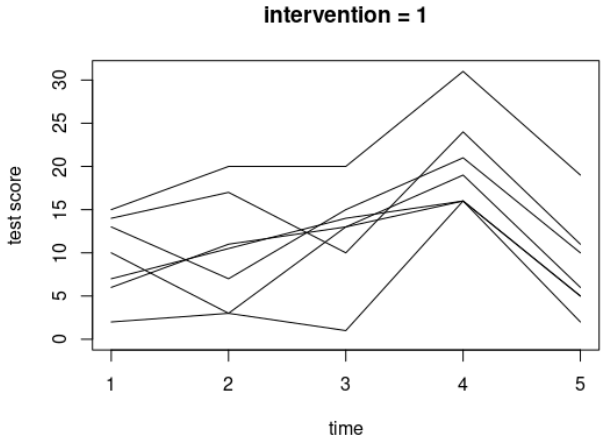
`intervention = 1`:

| NT\GMM | 1 | 2 | 3 | total |
|---|---|---|---|---|
| 1 | 45 | 2 | 5 | 52 |
| 2 | 62 | 4 | 45 | 111 |
| 3 | 15 | 1 | 39 | 55 |
| total | 122 | 7 | 89 | 218 |

`intervention = 2`:

| NT\GMM | 1 | 2 | 3 | total |
|---|---|---|---|---|
| 1 | 46 | 2 | 0 | 48 |
| 2 | 98 | 10 | 13 | 121 |
| 3 | 22 | 2 | 20 | 44 |
| total | 166 | 14 | 33 | 213 |

***Tabel 4.8:*** Sammenligning af gruppeinddelingen baseret på de nationale tests (NT) med inddelingen baseret på den tilpassede GMM, ved brug af EM-algoritmen.




***Figur 4.9:*** Observerede vækstkurver for de syv elever i datasæt 1, som er prædikteret til at tilhøre gruppe 2.

Derudover ser vi, at der kun er få elever i datasæt 1 fra gruppen med de lavt præsterende, som ifølge den tilpassede GMM bør tilhøre gruppe 3 og omvendt. Vi ser også, at næsten alle middel præsterende elever tilhører enten gruppe 1 eller gruppe 3, hvoraf størstedelen bliver

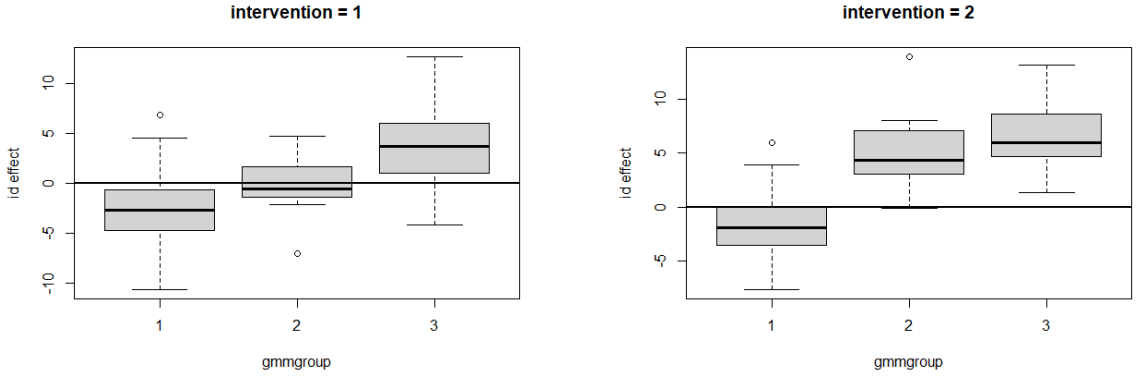
prædikteret til at tilhøre gruppe 1. Dette forklarer forskydningen af den tilhørende vækstkurve i forhold til den vi fik ved brug af en mixed model. Tilsvarende gælder det i datasæt 2, at der ikke er nogle elever, der er i den lavt præsterende gruppe, som bliver prædikteret til gruppe 3. Dog er det omvendte ikke tilfældet. De to gruppeinddelinger stemmer altså overens for 40% af eleverne i datasæt 1 og for 36% af eleverne i datasæt 2.

For at få et udtryk for korrelationen mellem gruppeinddelingerne kan vi bruge *Spearmans korrelationskoefficient*. Lad $r_{x_i}$ være rangen for den $i$'te realisering af den stokastiske variabel $X$, sådan at $r_{x_i} = j$, hvis $x_i$ er den $j$'te mindste værdi iblandt realiseringerne af $X$. For $n$ realiseringer af den stokastiske variabel $(X, Y)$ er Spearmans korrelationskoefficient givet ved

$$\rho = \frac{\sum_{i=1}^{n}(r_{x_i} - \bar{r}_x)(r_{y_i} - \bar{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x_i} - \bar{r}_x)^2 \sum_{j=1}^{n}(r_{y_j} - \bar{r}_y)^2}}, \tag{4.83}$$

hvor $\bar{r}_x$ er gennemsnittet af $r_{x_i}$ for $i = 1, \ldots, n$ [22, s. 430]. Vi kan nu udregne Spearmans korrelationskoefficient mellem de to gruppeindelinger, som er 0.43 og 0.38 for henholdvis datasæt 1 og 2. Der er altså en svag til moderart positiv korrelation mellem inddelingerne.

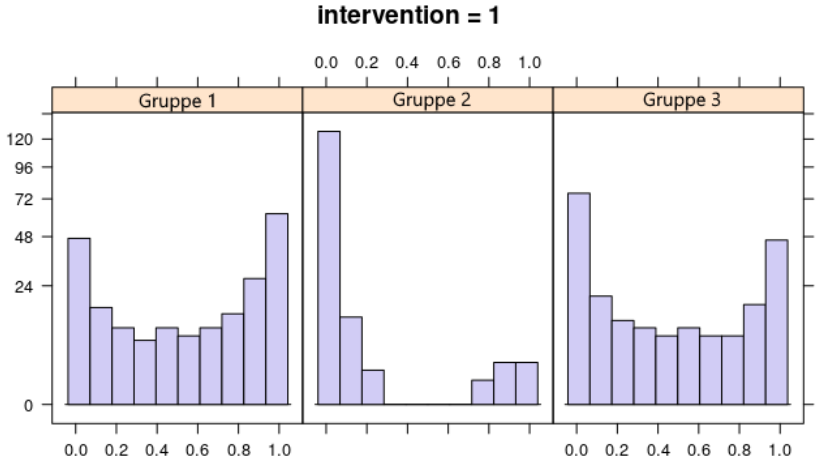
Vi sammenligner nu gruppeinddelingen baseret på den tilpassede GMM med de prædikterede eleveffekter fra Afsnit 3.5, som var prædikteret ud fra modellen uden `group`-faktoren. Denne sammenligning kan ses i Figur 4.10. Vi ser her, at eleverne i gruppe 3 generelt har positive prædiktioner, hvilket især er tilfældet for datasæt 2, samt at eleverne i gruppe 1 generelt har negative prædiktoner. Altså tyder det på, at der er en sammenhæng mellem eleveffekten fra den tilpassede mixed model og prædiktionen af gruppeinddelingen. Bemærk, at idet gruppe 2 for datasæt 1 kun består af syv individer, kan det være misvisende at opfatte denne gruppe som de middel præsterende elever og tilsvarende for datasæt 2.



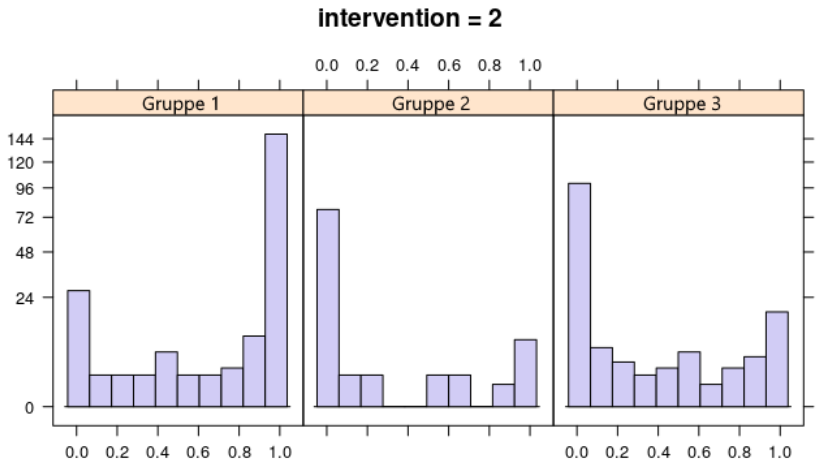

***Figur 4.10:*** Boksplots af de prædikterede eleveffekter fra den tilpassede mixed model fordelt over gruppeinddelingen for den tilpassede GMM, ved brug af EM-algoritmen.

Ved nu at udregne de gennemsnitlige posterior sandsynligheder for hver af de tre grupper, for modellen tilpasset på datasæt 1, får vi 0.8787 for gruppe 1, 0.9043 for gruppe 2 og 0.8724 for gruppe 3. Disse gennemsnit indikerer, at grupperne generelt er godt separerede, idet de alle er tæt på 1, og vi ser den samme indikation i rootogrammet på Figur 4.11. Tilsvarende får vi henholdsvis 0.9751, 0.8623 og 0.8701 for modellen, der er tilpasset på datasæt 2. Idet gruppe 1 har en markant højere ratio end de resterende grupper, gælder det, at denne gruppe

er bedre separeret fra de to andre grupper, end de andre grupper er. Dette stemmer overens med Figur 4.12, hvor rootogrammet for denne gruppe viser en høj frekvens omkring 1.

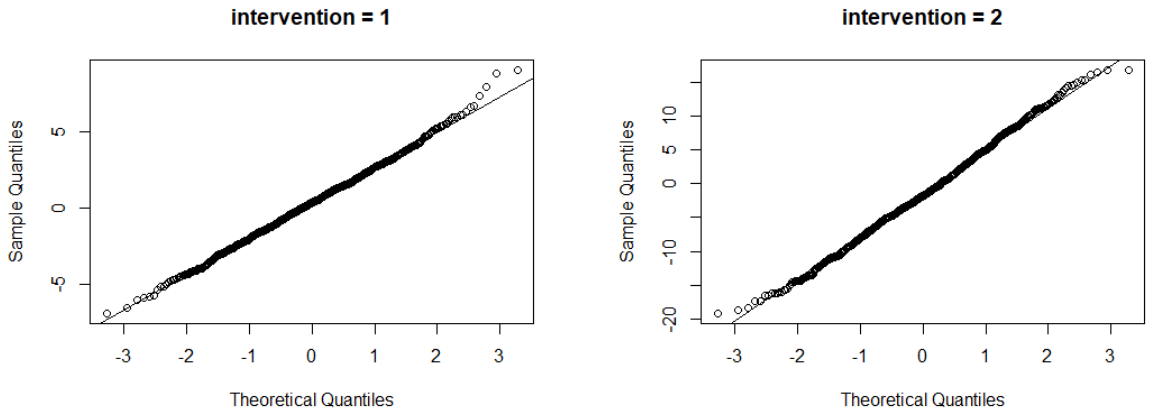



***Figur 4.11:*** Rootogrammer for datasæt 1 ved brug af EM-algoritmen.




***Figur 4.12:*** Rootogrammer for datasæt 2 ved brug af EM-algoritmen.

Ved at udregne $\hat{\zeta}_i$ som i (4.19) for $i = 1, \ldots, d$, kan vi opnå residualvektoren $y - \hat{y}$, hvormed vi får Q-Q-plottene på Figur 4.13. Her ser vi, at antagelsen om normalitet af $Y$ er acceptabel.




***Figur 4.13:*** Q-Q-plots af residualerne for modellerne tilpasset ved EM-algoritmen på henholdsvis datasæt 1 og 2.

### 4.5.2 Modeller med tre latente grupper med Marquardt-algoritmen

Tilpasser vi i stedet modellen i (4.80) ved brug af Marquardt-algoritmen, implementeret i R-funktionen `hlme`, med 1000 forskellige startværdier, får vi estimaterne i Tabel 4.9 og 4.10. Her er grupperene nummereret som grupperne baseret på modellen tilpasset ved EM-algoritmen, sådan at deres vækstkurver stemmer overens. Ved brug af denne metode, får vi estimater, som er nogenlunde tilsvarende estimaterne opnået ved EM-algoritmen, hvilket bekræftes af Figur 4.14 og 4.15.

`intervention` = 1:

| | gruppe 1 | gruppe 2 | gruppe 3 |
|---|---|---|---|
| $\alpha_{k1}$ | 6.5802 | 5.8821 | 8.5951 |
| $\alpha_{k2}$ | -0.2637 | 0.3972 | 1.9231 |
| $\alpha_{k3}$ | 4.2172 | 2.0086 | 2.0315 |
| $\alpha_{k4}$ | -4.3438 | 5.7648 | -0.6498 |
| $\alpha_{k5}$ | 0.6007 | -19.5866 | -2.8222 |
| $\gamma_k$ | 0.6922 | 7.9935 | 4.7236 |

***Tabel 4.9:*** Parameterestimater for datasæt 1 ved brug af Marquardt-algoritmen.

`intervention` = 2:

| | gruppe 1 | gruppe 2 | gruppe 3 |
|---|---|---|---|
| $\alpha_{k1}$ | 6.0316 | 9.7791 | 10.0179 |
| $\alpha_{k2}$ | 0.7673 | -1.1797 | 3.7433 |
| $\alpha_{k3}$ | -0.0630 | 2.1884 | -2.2593 |
| $\alpha_{k4}$ | 1.8820 | 9.2380 | 0.9195 |
| $\alpha_{k5}$ | -2.2882 | -7.9175 | -0.3277 |
| $\gamma_k$ | 0.2576 | 6.9127 | 1.1900 |

***Tabel 4.10:*** Parameterestimater for datasæt 2 ved brug af Marquardt-algoritmen.

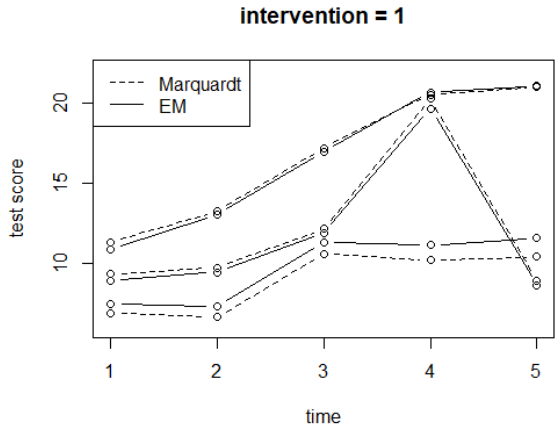
Yderligere får vi for datasæt 1, at

$$\hat{\psi}_{\texttt{id}} = 7.7813 \quad \text{og} \quad \hat{\sigma}^2 = 6.7055, \tag{4.84}$$

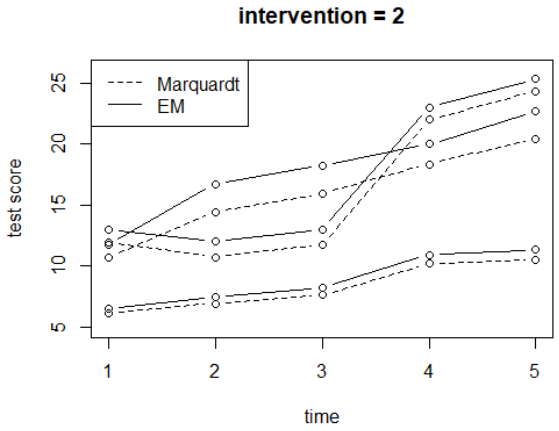
og tilsvarende for datasæt 2, at

$$\hat{\psi}_{\texttt{id}} = 16.9001 \quad \text{og} \quad \hat{\sigma}^2 = 6.5184. \tag{4.85}$$

Altså udgør elevvariansen for datasæt 1 og 2 henholdsvis 54% og 72% af den samlede varians.




***Figur 4.14:*** Middelvækstkurver for datasæt 1 ved brug af Marquardt- og EM-algoritmen.




***Figur 4.15:*** Middelvækstkurver for datasæt 2 ved brug af Marquardt- og EM-algoritmen.

Vi ser nu på Figur 4.14, at middelvækstkurven for gruppe 1 baseret på Marquardt-algoritmen ligger lavere end den tilsvarende kurve baseret på EM-algoritmen. Tilsvarende ser vi på Figur 4.15, at alle kurverne for modellen tilpasset ved Marquardt-algoritmen er lavere end de tilsvarende kurver baseret på EM-algoritmen. Dette kan skyldes, at gruppeinddelingen ved brug af de to metoder ikke stemmer overens for alle individer, og en sammenligning af disse inddelinger kan ses i Tabel 4.11. Det gælder heraf, at de to prædikterede gruppeinddelinger for både datasæt 1 og 2 stemmer overens for henholdsvis 93% og 92% af eleverne. Eksempelvis ser vi for datasæt 1, at den eneste forskel på de to inddelinger er de 16 elever, som modellen tilpasset ved Marquardt-algoritmen prædikterer til at tilhøre gruppe 3, men som modellen tilpasset ved EM-algoritmen prædikterer til at tilhøre gruppe 1. Forskellen i disse inddelinger kan skyldes, at de to implementeringer af algoritmerne ikke har de samme konvergenskriterier.

`intervention = 1`:

| EM\Marq. | 1 | 2 | 3 | total |
|---|---|---|---|---|
| 1 | 106 | 0 | 16 | 122 |
| 2 | 0 | 7 | 0 | 7 |
| 3 | 0 | 0 | 89 | 89 |
| total | 106 | 7 | 105 | 218 |

`intervention = 2`:

| EM\Marq. | 1 | 2 | 3 | total |
|---|---|---|---|---|
| 1 | 149 | 0 | 17 | 166 |
| 2 | 0 | 13 | 1 | 14 |
| 3 | 0 | 0 | 33 | 33 |
| total | 149 | 13 | 51 | 213 |

***Tabel 4.11:*** Sammenligning af gruppeinddelingen baseret på modellen tilpasset ved Marquardt-algoritmen med inddelingen baseret på modellen tilpasset ved EM-algoritmen.

AIC og BIC for modellerne ved EM- og Marquardt-algoritmen kan ses i Tabel 4.12. Bemærk, at vi opnår de laveste værdier af AIC og BIC ved brug af Marquardt-algoritmen, hvilket indikerer, at denne algoritmen giver de bedste tilpassede modeller.

En fordel ved at bruge Marquardt-algoritmen er, at variansen for alle parameterestimaterne udregnes som en del af metoden. Vi kan derfor lave Wald-test for at undersøge nulhypotesen om, at vækstkurvernes hældninger er 0, hvormed elevernes udvikling ikke er signifikant. Bemærk, at disse hældninger er udregnet som i Appendiks A. Teststørrelsen for $\mathcal{H}_0 : \delta^{t(t+1)}_{\texttt{group}k} = 0$

| | | AIC | BIC |
|---|---|---|---|
| datasæt 1 | EM | 5234.71 | 5341.44 |
| | Marq. | 5231.52 | 5312.75 |
| datasæt 2 | EM | 5223.58 | 5330.81 |
| | Marq. | 5219.75 | 5300.43 |

***Tabel 4.12:*** AIC og BIC ved brug af EM- og Marquardt-algoritmen for begge datasæt.

er da givet ved

$$W = \frac{(\hat{\delta}^{t(t+1)}_{\texttt{group}k})^2}{\widehat{\text{Var}}(\hat{\delta}^{t(t+1)}_{\texttt{group}k})}, \quad k = 1, 2, 3, \quad t = 1, \ldots, 4, \tag{4.86}$$

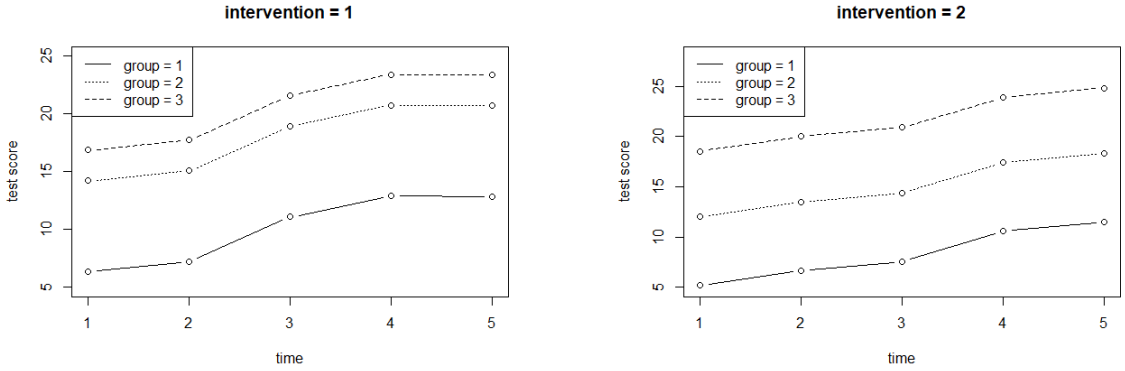
som er asymptotisk $\chi^2(1)$-fordelt under $\mathcal{H}_0$ [16, s. 23]. Denne test er i dette tilfælde meningsfuld, så længe at $\hat{\alpha}_k$ er en god approksimation af MLE for $\alpha_k$, som er asymptotisk normalfordelt, når $d \to \infty$, hvormed $\delta^{t(t+1)}_{\texttt{group}k}$ også er asymptotisk normalfordelte for $t = 1, \ldots, 4$. Resultaterne af disse tests for alle hældningerne for modellen tilpasset på datasæt 1 kan ses i Tabel 4.13. Bemærk, at idet modellen tilpasset ved Marquardt-algoritmen, ligesom ved EM-algoritmen, kun har tildelt syv elever til gruppe 2, er det ikke meningsfuldt at sammenligne middelvækstkurven for denne gruppe med den tilsvarende kurve for den tilpassede mixed model fra Afsnit 3.5. Sammenligner vi dog gruppe 1 og 3 for denne model med de lavt og højt præsterende elever, ser vi, at det er de samme perioder, hvori væksten i elevernes udvikling er signifikant. Vores konklusioner stemmer derfor overens med konklusionerne i Afsnit 3.5. Vi bemærker yderligere, at disse estimater også stemmer godt overens med hældningerne i Tabel 4.7 for modellen tilpasset ved brug af EM-algoritmen.

`intervention` = 1:

| | mixed model | $p$-værdi | | GMM | $p$-værdi | |
|---|---|---|---|---|---|---|
| $\delta^{12}_{\texttt{group1}}$ | -0.0242 | 0.96940 | | -0.2637 | 0.54495 | |
| $\delta^{23}_{\texttt{group1}}$ | 3.5951 | 4.48e-08 | *** | 3.9535 | 0.00000 | *** |
| $\delta^{34}_{\texttt{group1}}$ | 0.6284 | 0.36650 | | -0.3903 | 0.40660 | |
| $\delta^{45}_{\texttt{group1}}$ | -0.3240 | 0.63998 | | 0.2104 | 0.68221 | |
| $\delta^{12}_{\texttt{group2}}$ | 0.4779 | 0.25160 | | 0.3972 | 0.83128 | |
| $\delta^{23}_{\texttt{group2}}$ | 4.3920 | 2.48e-23 | *** | 2.4058 | 0.14938 | |
| $\delta^{34}_{\texttt{group2}}$ | 2.0302 | 8.30e-06 | *** | 8.1706 | 1.67e-06 | *** |
| $\delta^{45}_{\texttt{group2}}$ | -0.0443 | 0.92110 | | -11.4160 | 1.57e-11 | *** |
| $\delta^{12}_{\texttt{group3}}$ | 2.4517 | 4.00e-05 | *** | 1.9231 | 1.18e-05 | *** |
| $\delta^{23}_{\texttt{group3}}$ | 3.1184 | 4.52e-07 | *** | 3.9546 | 0.00000 | *** |
| $\delta^{34}_{\texttt{group3}}$ | 2.4572 | 0.00010 | *** | 3.3049 | 2.09e-10 | *** |
| $\delta^{45}_{\texttt{group3}}$ | 0.1598 | 0.79880 | | 0.4826 | 0.28604 | |

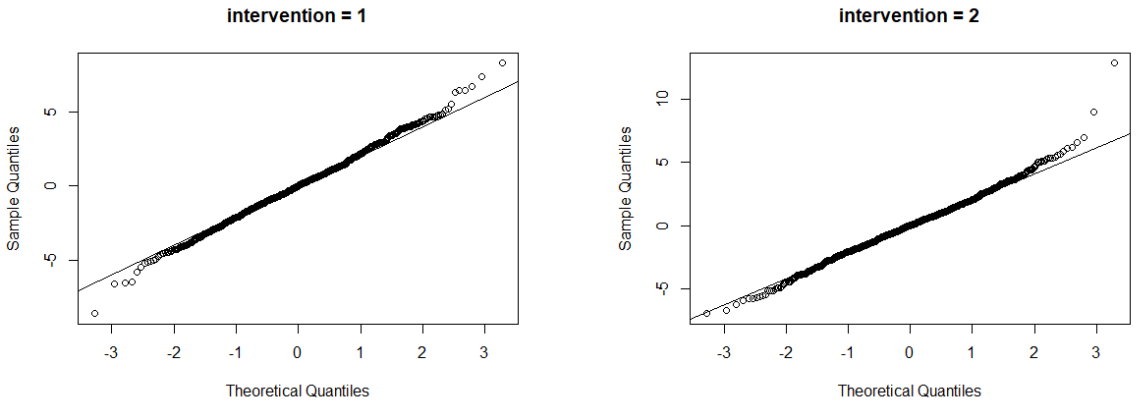
***Tabel 4.13:*** Sammenligning af estimater for modellerne tilpasset på datasæt 1 samt $p$-værdierne for hver test. Signifikanskoder: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

Vi tilpasser nu igen modellen i (4.80) ved brug af Marquardt-algoritmen, men hvor vi lader $\alpha_k$ være ens med undtagelse af $\alpha_{k1}$, for $k = 1, 2, 3$, hvormed vi får ens hældninger på vækstkurverne inden for alle tre grupper men forskellige skæringer. Middelvækstkurverne for denne model kan ses på Figur 4.16, og vi får en AIC og BIC på henholdsvis 5324.67 og 5372.05 for datasæt 1 og 5312.56 og 5359.62 for datasæt 2. Da disse værdier er større end værdierne for modellerne med forskellige hældninger for de tre grupper, indikerer dette, at en GMM forbedres for dette datsæt, hvis vi tillader forskellige hældninger for grupperne. Bemærk, at dette også stemmer overens med F-testen i Afsnit 3.5 om, at krydsfaktoren `group:time` er signifikant for den tilpassede mixed model.




***Figur 4.16:*** Middelvækstkurver for modellerne hvor alle gruppernes vækstkurver har samme hældninger.

Ved igen at udregne $\hat{\zeta}_i$ som i (4.19) for $i = 1, \ldots, d$, opnår vi Q-Q-plottene for residualvektoren, som ses på Figur 4.17. Heraf ser vi også, at normalitet af $Y$ er en acceptabel antagelse.

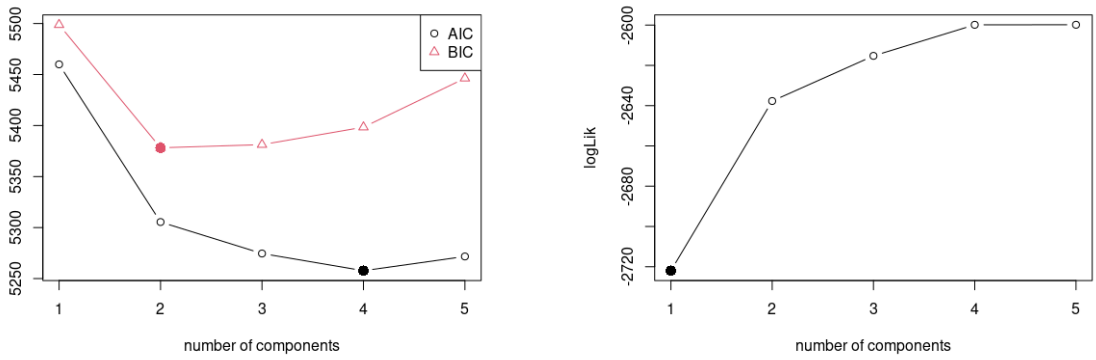



***Figur 4.17:*** Q-Q-plots af residualerne for modellerne tilpasset ved Marquardt-algoritmen på henholdsvis datasæt 1 og 2.

### 4.5.3 Undersøgelse af det optimale antal grupper

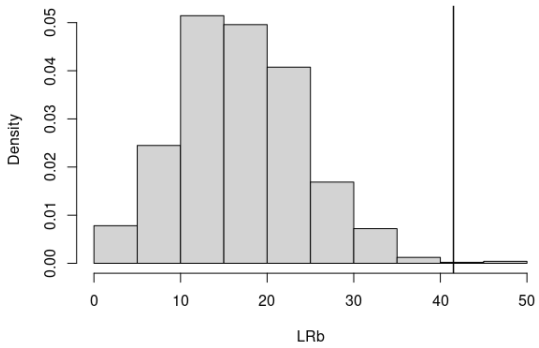
Vi er yderligere interesserede i, om der er et andet mere oplagt antal grupper, som beskriver data bedre. Vi betragter først datasæt 1 og tilpasser en GMM for $K = 1, \ldots, 5$, ved brug

af EM-algoritmen, hvor vi benytter 100 forskellige startværdier for hvert $K$. Vi opnår da værdierne for AIC, BIC og estimater for den maksimerede log-likelihoodfunktion, som kan ses på Figur 4.18. Ifølge dette vil det altså være mest oplagt at vælge $K = 2$ eller $K = 4$. Vi



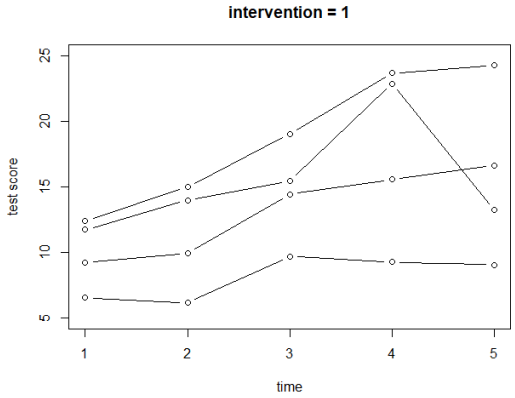

***Figur 4.18:*** AIC, BIC og maksimum af log-likelihoodfunktionen for $K = 1, \ldots, 5$.

laver derfor en BLRT med $K_0 = 2$, $K_1 = 4$ og $B = 1000$, hvor vi yderligere får, at likelihood ratio-teststørrelsen er LR $= 41.5119$. Resultatet af dette kan ses på Figur 4.19, hvor vi får en tilhørende $p$-værdi på 0.002, hvorfor vi forkaster nulhypotesen om, at der er to grupper.

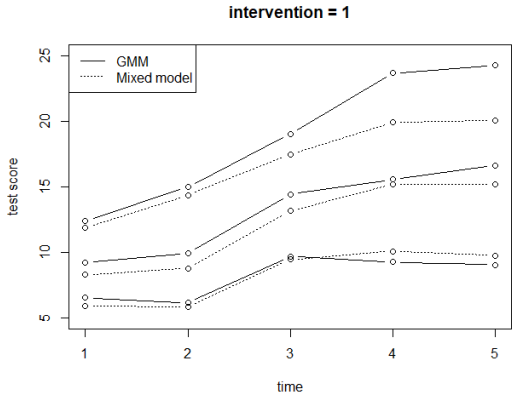



***Figur 4.19:*** BLRT for datasæt 1, hvor den lodrette linje repræsenterer LR $= 41.5119$.

Modellen med fire grupper for datasæt 1 inddeler individerne i gruppestørrelser på henholdvis 86, 75, 12 og 45, og for denne model får vi middelvækstkurverne på Figur 4.20. Vi ser her, at vi stadig har en gruppe af elever, som generelt har høje testscorer til den fjerde test og lave testscorer til den efterfølgende test, som nu består af 12 individer. Yderligere har vi tre kurver, der ligner dem fra den tilpassede mixed model fra Afsnit 3.5, hvilket ses i sammenligningen på Figur 4.21, hvor kurven for gruppen med de 12 indvider er udeladt. Hvis vi nummererer grupperne således, at gruppen med 12 indvider er gruppe 4, samt at de resterende grupper rangeres efter, hvor højt deres middelvækstkurver ligger, kan vi opstille Tabel 4.14, hvormed vi kan sammenligne gruppeinddelingerne.




***Figur 4.20:*** Middelvækstkurver for den tilpassede GMM med $K = 4$ for datasæt 1.



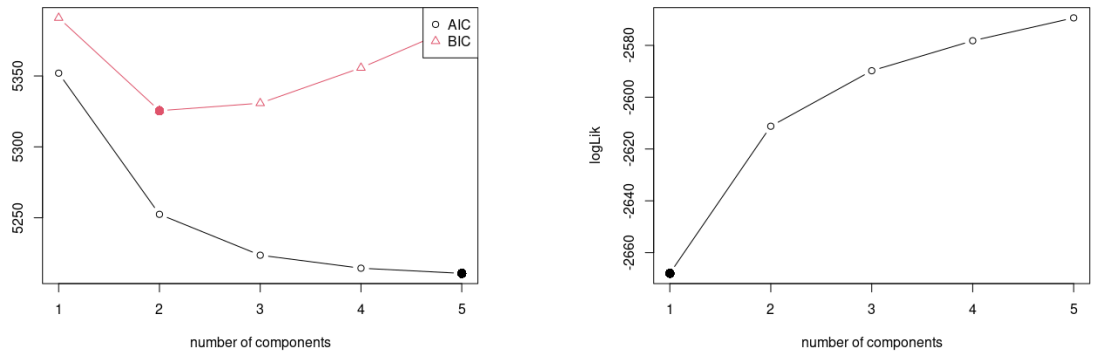

***Figur 4.21:*** Udvalgte middelvækstkurver for GMM med $K = 4$, sammenlignet med middelvækskurverne for mixed model.

`intervention` $= 1$:

| NT\GMM | 1 | 2 | 3 | 4 | total |
|---|---|---|---|---|---|
| 1 | 36 | 12 | 2 | 2 | 52 |
| 2 | 31 | 56 | 19 | 5 | 111 |
| 3 | 8 | 18 | 24 | 5 | 55 |
| total | 75 | 86 | 45 | 12 | 218 |

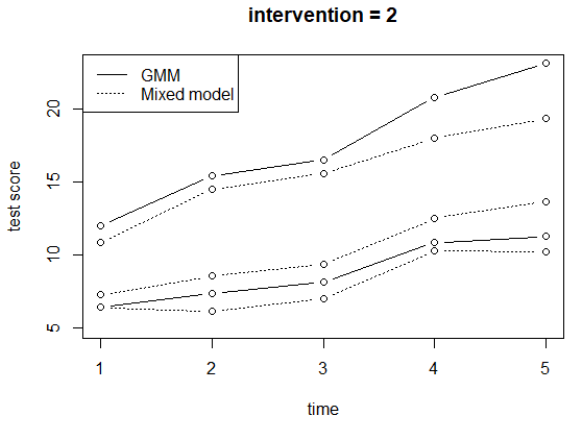
***Tabel 4.14:*** Sammenligning af gruppeinddeling baseret på den tilpassede GMM, med fire grupper, og de nationale tests.

Vi ser her, at eleverne i gruppe 4 kommer fra alle tre præstationsgrupper, samt at det igen gælder, at kun få elever fra den lavt præsterende gruppe prædikteres til at ligge i gruppe 3 og omvendt. Prædiktionerne for den tilpassede GMM stemmer overens med gruppeinddelingen for de nationale tests for 53% af eleverne, og vi får en svag til moderat positiv Spearman korrelation på 0.44 mellem de to gruppeindelinger.




***Figur 4.22:*** AIC, BIC og maksimum af log-likelihoodfunktionen for $K = 1, \ldots, 5$.

Vi benytter nu samme fremgangsmåde for datasæt 2, hvor vi får resultaterne på Figur 4.22. Her ser vi, at AIC falder som antallet af grupper vokser, men at BIC er minimeret ved $K = 2$. Det gælder yderligere, at log-likelihoodfunktionen stiger, når $K$ bliver større. Baseret på BIC er det oplagt at vælge modellen med $K = 2$. Denne model prædikterer gruppestørrelser på henholdsvis 45 og 168 elever, og de tilsvarende middelvækstkurver er på Figur 4.23 sammenlignet med middelvækskurverne for den tilpassede mixed model fra Afsnit 3.5. Vi ser her, at den øverste kurve for den tilpassede GMM stemmer overens med kurven for de højt præsterende elever, samt at den nederste kurve stemmer overens med kurverne for de lavt og middel præsterende elever.




***Figur 4.23:*** Middelvækstkurver for den tilpassede GMM med $K = 2$ for datasæt 2, sammenlignet med vækstkurverne for den tilsvarende mixed model.

Sammenligner vi modellens prædiktioner med gruppeinddelingen baseret på de nationale test, får vi resultaterne i Tabel 4.15. Her ser vi, at størstedelen af de lavt og middel præsterende elever er i gruppe 1, hvorimod halvdelen af de højt præsterende elever er i gruppe 2. Dette kan være årsagen til, at middelvækstkurven for gruppe 2 er højere end kurven for de højt præsterende elever i den tilpassede mixed model.

`intervention` $= 2$:

| NT\GMM | 1 | 2 | total |
|---|---|---|---|
| 1 | 47 | 1 | 48 |
| 2 | 101 | 20 | 121 |
| 3 | 20 | 22 | 44 |
| total | 168 | 45 | 213 |

***Tabel 4.15:*** Sammenligning af gruppeinddelingerne baseret på den tilpassede GMM, med to grupper, og de nationale tests.

# 5 | Fortolkning

Analysen af datasæt 1, ved brug af den tilpassede mixed model i Afsnit 3.5, viste, at de lavt præsterende elever kun udviklede deres matematiske kompetencer inden for brøkregning i anden periode, hvor de blev introduceret for brøker i undervisningen. De middel præsterende elever udviklede sig også i denne periode, men yderligere også i tredje periode, hvor de havde et mere komplekst undervisningsforløb om brøkregning. Derudover udviklede de højt præsterende elever sig også i første periode. Ud fra variansestimaterne så vi, at størstedelen af den totale varians blev udgjort af elevvariansen med 59%, samt at hverken klasse- eller skolevariansen havde megen betydning for modellen. For datasæt 2 kom vi lignende frem til, at de lavt præsterende elevers udvikling kun var signifikant i perioden, hvor de blev introduceret til brøkregning, samt at de højt præsterende elever også udviklede sig i andre perioder. Dette forstærker altså vores resultater, idet de to datasæt er uafhængige.

I Afsnit 4.5 så vi, at vi ved brug af en GMM med tre grupper, tilpasset ved EM-algoritmen, fik middelvækstkurver, som overvejende lignede dem vi fik for lavt og højt præsterende elever i Afsnit 3.5. For begge datasæt var vækstkurven for gruppe 2 dog betydelig anderledes fra alle kurverne baseret på de tilpassede mixed modeller. Dette skyldes angiveligt, at de tilpassede GMM'er prædikterer meget få elever til at tilhøre denne gruppe. Disse elever havde en særlig tendens, som modellerne opfangede, hvilket vi så på Figur 4.9 for datasæt 1. Yderligere så vi, at der er en svag til moderat positiv korrelation mellem disse to gruppeinddelinger.

Ved at sammenligne gruppeinddelingen, baseret på den tilpassede GMM, med de prædikterede eleveffekter fra Afsnit 3.5, så vi, at eleverne i gruppe 1 generelt har negative effekter, samt at eleverne i gruppe 3 generelt har positive effekter for begge datasæt. Denne gruppeinddeling har altså den ønskede sammenhæng med elevernes individuelle effekter på deres udvikling.

Ved i stedet at benytte Marquardt-algoritmen fik vi vækstkurver, som ligner vækstkurverne for modellen tilpasset ved EM-algoritmen, hvilket også bekræftes af, at gruppeinddelingerne stemmer overens for 93% og 92% af eleverne i henholdsvis datasæt 1 og 2. For datasæt 1 er vækstkurven for gruppe 1 en smule lavere end kurven opnået ved brug af EM-algoritmen, hvilket skyldes, at modellen tilpasset ved Marquardt-algoritmen prædikterer nogle elever fra gruppe 1 i den forrige model til at tilhøre gruppe 3. Denne forskel kan skyldes, at de to implementeringer af algoritmerne ikke har de samme konvergenskriterier. Altså minder middelvækstkurverne opnået ved brug af den tilpassede GMM og den tilpassede mixed model overvejende om hinanden med undtagelse af kurverne for gruppe 2 og for de middel præsterende elever. Sammenligningen i Tabel 4.13 for datasæt 1 bekræfter også ovenstående konklusioner. Heraf bemærker vi også, at ved brug af en GMM, er det kun hældningerne i tredje og fjerde periode, som er signifikante for gruppe 2.

Derudover udgør elevvariansen for datasæt 1 og 2 henholdsvis 66% og 54% af den totale varians for modellerne tilpasset ved brug af EM-algoritmen og henholdsvis 54% og 72% for modellerne tilpasset ved brug af Marquardt-algoritmen. Størstedelen af den totale varians er altså udgjort af elevvariansen, hvilket stemmer overens med, hvad vi så i Afsnit 3.5.

Ved brug af F-tests i Afsnit 3.5 kunne vi for begge datasæt konkludere, at gruppernes middelvækstkurver over de fire perioder ikke blot var forskydninger af hinanden. Dermed må der være en betydelig forskel på gruppernes udvikling over tid. For at undersøge dette for en GMM tilpassede vi en model ved brug af Marquardt-algoritmen, hvor hældningerne for de tre gruppers kurver var specificeret ved den samme parametervektor. Vi så da, at informationskriterierne for disse modeller, hvor hældningerne ikke kan variere på tværs af grupperne, havde højere værdier end for de allerede tilpassede modeller. Vi slutter derfor samme konklusion om, at der er forskel på gruppernes udvikling.

For modellerne tilpasset ved EM-algoritmen så vi en generel indikation på, at grupperne baseret på modellerne er godt separerede jævnfør rootogrammerne. Specielt så vi, at gruppe 1 for datasæt 2 er gruppen, der er bedst separeret fra de resterende, hvilket stemmer overens med, at denne gruppe også har den højeste gennemsnitlige posterior sandsynlighed for at tilhøre denne gruppe.

For at undersøge om et andet antal grupper vil være mere passende, tilpassede vi en GMM for $K = 1, \ldots, 5$ ved brug af EM-algoritmen. For datasæt 1 viste det sig, at vi vil opnå den mindste værdi for AIC og BIC ved at vælge en model med henholdsvis $K = 4$ og $K = 2$. Ved da at benytte bootstrap likelihood ratio-test, forkastede vi hypotesen om, at data følger en model med kun to grupper, idet vi fik en $p$-værdi på 0.002. Vi fik derfor en indikation på, at en model med fire grupper vil være det mest optimale valg, hvilket også vil give det største maksimum af log-likelihoodfunktionen. Ved at benytte denne GMM så vi, at gruppeinddelingerne stemmer overens med inddelingen baseret på de nationale tests for 53% af eleverne. Derudover så vi, at eleverne, der blev prædikteret til at tilhøre gruppe 4, kommer fra alle tre præstationsgrupper, samt at der kun er tildelt 12 elever i alt til denne gruppe. Vi så yderligere, at middelvækstkurverne for de øvrige tre grupper ligner meget de tilsvarende kurver opnået ved den tilpassede mixed model i Afsnit 3.5.

For datasæt 2 vurderede vi, at det mest oplagte valg af grupper var $K = 2$. Ved at benytte denne model så vi, at 101 ud af 121 elever fra den middel præsterende gruppe blev prædikteret til at tilhøre gruppe 1, samt at alle fra den lavt præsterende gruppe, på nær en enkelt elev, også blev prædikteret til at tilhøre denne gruppe. Dette kan forklare, at middelvækstkurven for gruppe 1 ligger mellem de to middelvækstkurver for de lavt og middel præsterende grupper. Derudover er de højt præsterende elever omtrent ligeligt fordelt over de to grupper.

## 5.1 Diskussion

Vi så, at de tilpassede GMM'er prædikterede få elever til at tilhøre gruppe 2 for begge datasæt, og grundet disse gruppers størrelse rejser det spørgsmålet, om disse gruppers abnormitet skyldes støj. Vigtigheden af dette spørgsmål bliver yderligere forstærket af, at strukturen af gruppernes middelvækstkurver, og deres undervisningsforløb, er forskellige. I så fald vil det være interessant at tilpasse en model med en gruppe mindre. Dette viste sig at være en god idé for datasæt 2, da ingen af kurverne for denne model ligner støj. Ved at tilpasse en model

med fire grupper for datasæt 1 så vi, at denne model også dannede en mulig støjgruppe. Derfor ville det igen være interessant, at tilpasse en model med en gruppe mindre, hvilket dog svarer til modellen, vi startede med, som også havde en støjgruppe.

Vi skal også overveje, om data bryder antagelsen om normalitet givet ved (4.16). Dette er strengt taget tilfældet, idet testscorerne kun tager værdier i $\mathbb{Z}$, hvormed fordelingen af disse er diskret. Yderligere er testenscoren begrænset til at ligge i intervallet [0,36], hvilket kan resultere i skæve fordelinger i enderne af intervallet. Antagelsen bliver dog bekræftet af Q-Q-plottene for residualvektorerne i analysen.

I analyserne af datasættet har vi kun betragtet tilfældet, hvor elevernes vækstkurver har stokastiske skæringer, men ikke stokastiske hældninger, hvorfor $\Psi = \text{diag}(\psi_{\texttt{id}}, 0, 0, 0, 0)$. Det ville i stedet have været en mere realistisk antagelse at vælge $\Psi$, sådan at alle diagonalindgange er forskellige fra nul, idet elevernes udvikling i så fald kan variere. Samtidig kunne vi også lade $\Gamma_k = (\gamma_{k1}, \ldots, \gamma_{k5})^T$, som ville tillade forskellige vækstkurver for hvert køn inden for hver gruppe, svarende til at inkludere krydsfaktorerne `gender:time` og `gender:group:time`. Disse tilføjelser ville dog øge antallet af parametre, som skal estimeres med $4(K+1)$, hvis $\Psi$ er en diagonalmatrix.

En fordel ved at benytte en GMM er, at modellen er fleksibel og tillader individuel variation inden for latente grupper, samt at den kan modellere data ikke-lineært for et passende valg af $\Lambda$. Der opstår dog en problemstilling i, at vi selv skal vurdere hvilket valg af $K$, der er mest optimalt. Antallet af grupper afhænger dog også af, hvilket spørgsmål man ønsker at besvare.

Derudover er en GMM muligvis ikke det rigtige værktøj, hvis målet med en undersøgelse er at lave en gruppeinddeling af lavt, middel og højt præsterende elever. Dette skyldes, at denne model ikke er garanteret at lave sådan en type inddeling, men også kan basere denne på andre tendenser i data som eksempelvis tendensen for gruppe 2 på Figur 4.7. Generelt skal vi altså være opmærksomme på om den struktur, modellen opfanger, også er den, vi ønsker at undersøge. Hvis vi på forhånd kender en troværdig og relevant inddeling af eleverne, er det derfor fordelagtigt blot at tilpasse en mixed model på data.

Yderligere er tilpasning af en GMM også en beregningsmæssig tung opgave, idet de numeriske metoder skal benytte mange forskellige startværdier.

# 6 | Bibliografi


[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi:10.18637/jss.v067.i01.

[2] K. A. Bollen and P. J. Curran. *Latent Curve Models a Structural Equation Perspective*. Wiley-Interscience, 2006.

[3] O. C. Carrasco. Gaussian mixture models explained [online]. 2019. URL: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95 [cited 2023-29-05].

[4] Q. Chen, W. Luo, G. J. Palardy, R. Glaman, and A. McEnturff. The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: A Monte Carlo study. *SAGE Open*, 7(1), 2017. URL: https://doi.org/10.1177/2158244017700459.

[5] S. W. Enck. *Latent Class Linear Mixed Models: a General Approach Implemented Via Sas Macro with a Tutorial for Clinical Researchers*. PhD thesis, Chapel Hill, NC: University of North Carolina at Chapel Hill, 2009. doi:10.17615/gqp1-yc24.

[6] R. Fletcher. *Practical Methods of Optimization*. A Wiley-Interscience publication. Wiley, 2. edition, 1987.

[7] A. Guessab and G. Schmeisser. Necessary and sufficient conditions for the validity of Jensen's inequality. *Archiv der Mathematik*, 100:561–570, 2013. URL: https://doi.org/10.1007/s00013-013-0522-3.

[8] U. Halekoh and S. Højsgaard. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software*, 59(9):1–30, 2014. URL: https://www.jstatsoft.org/v59/i09/.

[9] N. O. Jeffries. A note on 'Testing the number of components in a normal mixture'. *Biometrika*, 90(4):991–994, 2003. URL: http://www.jstor.org/stable/30042105.

[10] J. Jiang and T. Nguyen. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. Springer New York, NY, 2021.

[11] M. G. Kenward and J. H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997, 1997. URL: http://www.jstor.org/stable/2533558.

[12] F. Leisch. Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004. URL: https://www.jstatsoft.org/index.php/jss/article/view/v011i08.

[13] C. Li. Little's test of missing data. *The Stata Journal*, 13(4):795–809, 2013. URL: https://journals.sagepub.com/doi/10.1177/1536867X1301300407.

[14] R. J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988. URL: http://www.jstor.org/stable/2290157.

[15] Y. Lo, N. R. Mendell, and D. B. Rubin. Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778, 2001. URL: http://www.jstor.org/stable/2673445.

[16] H. Madsen and P. Thyregod. *Introduction to general and generalized linear models*. Texts in statistical science. CRC Press, 2011.

[17] G. McLachlan and S. Rathnayake. On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4, 09 2014. doi:10.1002/widm.1135.

[18] D. McNeish and J. R. Harring. The effect of model misspecification on growth mixture model class enumeration. *Journal of of Classification*, 34:223–248, 2017. doi:10.1007/s00357-017-9233-y.

[19] P. Menon. Data science simplified part 9: Interactions and limitations of regression models [online]. 2017. URL: https://tinyurl.com/4dvfkrkr [cited 2023-28-04].

[20] B. Muthén and K. Shedden. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2):463–469, 1999. URL: https://pubmed.ncbi.nlm.nih.gov/11318201/.

[21] J. Nocedal and S. Wright. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, 2. ed. edition, 2006.

[22] P. Olofsson and M. Andersson. *Probability, Statistics, and Stochastic Processes*. Wiley, 2. edition, 2012.

[23] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford science publications. OUP Oxford, 2001.

[24] P. L. Pedersen, P. Aunio, P. B. Sunde, M. Bjerre, and R. Waagepetersen. Differences in high- and low-performing students' fraction learning in the fourth grade. *The Journal of Experimental Education*, 2022. URL: https://doi.org/10.1080/00220973.2022.2107603.

[25] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2012. Version 20121115. URL: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html.

[26] C. Proust-Lima and H. Jacqmin-Gadda. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer Methods and Programs in Biomedicine*, 78(2):165–173, 2005. URL: https://www.sciencedirect.com/science/article/pii/S0169260705000209.

[27] C. Proust-Lima, V. Philipps, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2), 2017. URL: https://arxiv.org/abs/1503.00890.

[28] R Core Team. *optim: General-purpose Optimization*. R Foundation for Statistical Computing. URL: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/optim.

[29] M. Taboga. EM algorithm, lectures on probability theory and mathematical statistics. [online]. Kindle Direct Publishing, 2021. URL: https://www.statlect.com/fundamentals-of-statistics/EM-algorithm [cited 2023-03-09].

[30] N. Tierney and D. Cook. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, 105(7):1–31, 2023. doi:10.18637/jss.v105.i07.

[31] T. Tjur. Analysis of variance models in orthogonal designs. *International Statistical Review*, 52(1):33–65, 1984. URL: http://www.jstor.org/stable/1403242.

[32] G. van der Nest, V. L. Passos, M. J. Candel, and G. J. van Breukelen. An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, 43:100323, 2020. URL: https://www.sciencedirect.com/science/article/pii/S1040260819301881.

[33] K. J. Wardenaar. Latent class growth analysis and growth mixture modeling using R: A tutorial for two R-packages and a comparison with Mplus, 2022. URL: https://doi.org/10.31234/osf.io/m58wx.

# A | Appendiks - Omregning af estimater

Ud fra estimaterne opnået for de tilpassede mixed modeller i Afsnit 3.5, kan vi udregne den forventede gennemsnitlige CMB-score for piger til de fem tests i hver af de tre grupper. For gruppe 1 er disse givet ved

$$\begin{aligned} &\beta_0 \\ &\beta_0 + \beta_{\texttt{time2}} \\ &\beta_0 + \beta_{\texttt{time3}} \\ &\beta_0 + \beta_{\texttt{time4}} \\ &\beta_0 + \beta_{\texttt{time5}} \end{aligned} \tag{A.1}$$

og

$$\begin{aligned} &\beta_0 + \beta_{\texttt{group}k} \\ &\beta_0 + \beta_{\texttt{time2}} + \beta_{\texttt{group}k} + \beta_{\texttt{group}k\texttt{:time2}} \\ &\beta_0 + \beta_{\texttt{time3}} + \beta_{\texttt{group}k} + \beta_{\texttt{group}k\texttt{:time3}} \\ &\beta_0 + \beta_{\texttt{time4}} + \beta_{\texttt{group}k} + \beta_{\texttt{group}k\texttt{:time4}} \\ &\beta_0 + \beta_{\texttt{time5}} + \beta_{\texttt{group}k} + \beta_{\texttt{group}k\texttt{:time5}}, \end{aligned} \tag{A.2}$$

for $k = 2, 3$. For estimaterne opnået ved brug af en GMM i Afsnit 4.5, svarer ovenstående til

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 1 & 0 \\ 1 & 4 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{k1} \\ \alpha_{k2} \\ \alpha_{k3} \\ \alpha_{k4} \\ \alpha_{k5} \end{bmatrix} = \begin{bmatrix} \alpha_{k1} \\ \alpha_{k1} + \alpha_{k2} \\ \alpha_{k1} + 2\alpha_{k2} + \alpha_{k3} \\ \alpha_{k1} + 3\alpha_{k2} + 2\alpha_{k3} + \alpha_{k4} \\ \alpha_{k1} + 4\alpha_{k2} + 3\alpha_{k3} + 2\alpha_{k4} + \alpha_{k5} \end{bmatrix} \tag{A.3}$$

for $k = 1, 2, 3$. For tilsvarende at få den gennemsnitlige testscore for drenge, skal vi tilføje $\beta_{\texttt{gender}}$ i (A.1) og (A.2), samt $\gamma_k$ på alle indgange i (A.3). Det gælder derfor også, at hældningerne i de fire perioder for de tre grupper er givet ved

$$\begin{aligned} \delta^{12}_{\texttt{group1}} &= \beta_{\texttt{time2}} = \alpha_{12} \\ \delta^{23}_{\texttt{group1}} &= \beta_{\texttt{time3}} - \beta_{\texttt{time2}} = \alpha_{12} + \alpha_{13} \\ \delta^{34}_{\texttt{group1}} &= \beta_{\texttt{time4}} - \beta_{\texttt{time3}} = \alpha_{12} + \alpha_{13} + \alpha_{14} \\ \delta^{45}_{\texttt{group1}} &= \beta_{\texttt{time5}} - \beta_{\texttt{time4}} = \alpha_{12} + \alpha_{13} + \alpha_{14} + \alpha_{15}, \end{aligned} \tag{A.4}$$

og

$$\begin{aligned}
\delta^{12}_{\texttt{group}k} &= \beta_{\texttt{time2}} + \beta_{\texttt{group}k\texttt{:time2}} = \alpha_{k2} \\
\delta^{23}_{\texttt{group}k} &= \beta_{\texttt{time3}} + \beta_{\texttt{group}k\texttt{:time3}} - (\beta_{\texttt{time2}} + \beta_{\texttt{group}k\texttt{:time2}}) = \alpha_{k2} + \alpha_{k3} \\
\delta^{34}_{\texttt{group}k} &= \beta_{\texttt{time4}} + \beta_{\texttt{group}k\texttt{:time4}} - (\beta_{\texttt{time3}} + \beta_{\texttt{group}k\texttt{:time3}}) = \alpha_{k2} + \alpha_{k3} + \alpha_{k4} \\
\delta^{45}_{\texttt{group}k} &= \beta_{\texttt{time5}} + \beta_{\texttt{group}k\texttt{:time5}} - (\beta_{\texttt{time4}} + \beta_{\texttt{group}k\texttt{:time4}}) = \alpha_{k2} + \alpha_{k3} + \alpha_{k4} + \alpha_{k5},
\end{aligned} \tag{A.5}$$

for $k = 2, 3$.

# B | Appendiks - R-kode

Dette appendiks indeholder den relevante R-kode for eksempler og analyser igennem projektet.

## B.1 Kode til Afsnit 3.5

For at tilpasse modellerne i afsnittet og opnå de tilhørende parameterestimater, bruger vi følgende kode, hvor datasættet er gemt som `NT`.

```
model1 <- lmer(test ~ time * group + gender +
               (1|id) + (1|class_id) + (1|school_id),
               data = NT, subset = intervention == 1)
summary(model1)
model2 <- lmer(test ~ time * group + gender +
               (1|id) + (1|class_id) + (1|school_id),
               data = NT, subset = intervention == 2)
summary(model2)
```

For at udregne estimater for hældningerne af vækstkurverne samt lave t-tests for disse, bruger vi følgende for datasæt 1.

```
df <- table(NT$intervention)[1] - 1

txt <- c("konstrast_low12","konstrast_low23","konstrast_low34",
       "konstrast_low45","konstrast_highlow12","konstrast_highlow23",
       "konstrast_highlow34","konstrast_highlow45","konstrast_midlow12",
       "konstrast_midlow23","konstrast_midlow34","konstrast_midlow45",
       "konstrast_highmid12","konstrast_highmid23","konstrast_highmid34",
       "konstrast_highmid45","kontrast_high12","kontrast_high23",
       "kontrast_high34","kontrast_high45","kontrast_mid12","kontrast_mid23",
       "kontrast_mid34","kontrast_mid45")
est <- rep(0,length(txt))
SE <- rep(0,length(txt))
t <- rep(0,length(txt))
pval <- rep(0,length(txt))

#low
est[1] <- summary(model1)$coef[2,1]
SE[1] <- summary(model1)$coef[2,2]
t[1] <- summary(model1)$coef[2,4]
pval[1] <- summary(model1)$coef[2,5]

for (i in 3:5){
  est[i-1] <- summary(model1)$coef[i,1]-summary(model1)$coef[i-1,1]
  SE[i-1] <- sqrt(summary(model1)$coef[i,2]^2
```

```
              +summary(model1)$coef[i-1,2]^2-2*summary(model1)$vcov[i,i-1])
  t[i-1] <- est[i-1]/SE[i-1]
  pval[i-1] <- 2*pt(-abs(t[i-1]), df)
}
#highlow
est[5] <- summary(model1)$coef[9,1]
SE[5] <- summary(model1)$coef[9,2]
t[5] <- summary(model1)$coef[9,4]
pval[5] <- summary(model1)$coef[9,5]
for (i in 10:12){
  est[i-4] <- summary(model1)$coef[i,1]-summary(model1)$coef[i-1,1]
  SE[i-4] <- sqrt(summary(model1)$coef[i,2]^2
              +summary(model1)$coef[i-1,2]^2-2*summary(model1)$vcov[i,i-1])
  t[i-4] <- est[i-4]/SE[i-4]
  pval[i-4] <- 2*pt(-abs(t[i-4]), df)
}

#midlow
est[9] <- summary(model1)$coef[13,1]
SE[9] <- summary(model1)$coef[13,2]
t[9] <- summary(model1)$coef[13,4]
pval[9] <- summary(model1)$coef[13,5]
for (i in 14:16){
  est[i-4] <- summary(model1)$coef[i,1]-summary(model1)$coef[i-1,1]
  SE[i-4] <- sqrt(summary(model1)$coef[i,2]^2
              +summary(model1)$coef[i-1,2]^2-2*summary(model1)$vcov[i,i-1])
  t[i-4] <- est[i-4]/SE[i-4]
  pval[i-4] <- 2*pt(-abs(t[i-4]), df)

}

#highmid
est[13] <- summary(model1)$coef[9,1]-summary(model1)$coef[13,1]
SE[13] <- sqrt(summary(model1)$coef[9,2]^2
          +summary(model1)$coef[13,2]^2-2*summary(model1)$vcov[13,9])
t[13] <- est[13]/SE[13]
pval[13] <- 2*pt(-abs(t[13]), df)
for (i in 10:12){
  est[i+4] <- summary(model1)$coef[i,1]-summary(model1)$coef[i+4,1]
               -summary(model1)$coef[i-1,1]+summary(model1)$coef[i+3,1]
  SE[i+4] <- sqrt(summary(model1)$coef[i,2]^2+summary(model1)$coef[i+4,2]^2
              +summary(model1)$coef[i-1,2]^2+summary(model1)$coef[i+3,2]^2
              -2*summary(model1)$vcov[i,i+4]-2*summary(model1)$vcov[i,i-1]
              +2*summary(model1)$vcov[i,i+3]-2*summary(model1)$vcov[i+4,i-1]
              +2*summary(model1)$vcov[i+4,i+3]+2*summary(model1)$vcov[i-1,i+3])
  t[i+4] <- est[i+4]/SE[i+4]
  pval[i+4] <- 2*pt(-abs(t[i+4]), df)
}

#high
est[17] <- summary(model1)$coef[9,1]+summary(model1)$coef[2,1]
SE[17] <- sqrt(summary(model1)$coef[9,2]^2+summary(model1)$coef[2,2]^2
          +2*summary(model1)$vcov[2,9])
t[17] <- est[17]/SE[17]
pval[17] <- 2*pt(-abs(t[17]), df)
for (i in 10:12){
  est[i+8] <- summary(model1)$coef[i,1]+summary(model1)$coef[i-7,1]
```

```
               -summary(model1)$coef[i-1,1]-summary(model1)$coef[i-8,1]
  SE[i+8] <- sqrt(summary(model1)$coef[i,2]^2+summary(model1)$coef[i-7,2]^2
             +summary(model1)$coef[i-1,2]^2+summary(model1)$coef[i-8,2]^2
             +2*summary(model1)$vcov[i,i-7]-2*summary(model1)$vcov[i,i-1]
             -2*summary(model1)$vcov[i,i-8]-2*summary(model1)$vcov[i-7,i-1]
             -2*summary(model1)$vcov[i-7,i-8]+2*summary(model1)$vcov[i-1,i-8])
  t[i+8] <- est[i+8]/SE[i+8]
  pval[i+8] <- 2*pt(-abs(t[i+8]), df)
}

#mid
est[21] <- summary(model1)$coef[13,1]+summary(model1)$coef[2,1]
SE[21] <- sqrt(summary(model1)$coef[13,2]^2+summary(model1)$coef[2,2]^2
          +2*summary(model1)$vcov[2,13])
t[21] <- est[21]/SE[21]
pval[21] <- 2*pt(-abs(t[21]),df)
for (i in 14:16){
  est[i+8] <- summary(model1)$coef[i,1]+summary(model1)$coef[i-11,1]
               -summary(model1)$coef[i-1,1]-summary(model1)$coef[i-12,1]
  SE[i+8] <- sqrt(summary(model1)$coef[i,2]^2+summary(model1)$coef[i-11,2]^2
             +summary(model1)$coef[i-1,2]^2+summary(model1)$coef[i-12,2]^2
             +2*summary(model1)$vcov[i,i-11]-2*summary(model1)$vcov[i,i-1]
             -2*summary(model1)$vcov[i,i-12]-2*summary(model1)$vcov[i-11,i-1]
             -2*summary(model1)$vcov[i-11,i-12]
             +2*summary(model1)$vcov[i-1,i-12])
  t[i+8] <- est[i+8]/SE[i+8]
  pval[i+8] <- 2*pt(-abs(t[i+8]), df)
}

tab <- t(rbind(txt, est, SE, t, pval))
```

Vi bruger følgende kode til at lave F-test for datasæt 1.

```
model1.test <- lmer(test ~ time + group + gender
                    + (1|id) + (1|class_id) + (1|school_id),
                    data = NT, subset = intervention == 1)
KRmodcomp(model1, model1.test)
```

Slutteligt tilpasser vi modellerne uden `group`-faktoren samt udregner prædiktionerne af de stokastiske effekter.

```
model1ug <- lmer(test ~ time + gender +
                (1|id) + (1|class_id) + (1|school_id),
                data = NT, subset = intervention == 1, )
model2ug <- lmer(test ~ time + gender +
                (1|id) + (1|class_id) + (1|school_id),
                data = NT, subset = intervention == 2)

u1 <- ranef(model1ug)
u2 <- ranef(model2ug)
```

## B.2 Kode til Eksempel 4.2

Vi simulerer data til første del af eksemplet ved følgende.

```
Tn <- 5
k <- 2
n <- 100
psi1 <- 2
sigmai <- 1
Sigma <- sigmai*diag(Tn)
Psi <- diag(c(psi1, rep(0,Tn-1)))
x <- seq(1, Tn, length = Tn)
Lambda <- matrix(c(rep(1,Tn), 0, 1:(Tn-1),
                   rep(0,2), 1:(Tn-2),
                   rep(0,3), 1:(Tn-3),
                   rep(0,4), 1:(Tn-4)), nrow = Tn, byrow = FALSE)

set.seed(123)
d <- table(sample(1:k, n, replace=TRUE))
group <- 0
for (j in 1:k){
  group <- c(group, rep(j,d[j]))
}
group <- group[-1]

slope <- matrix(c(8,2,6,-4,2,
                  10,-2,2,2,3), nrow = k, byrow = TRUE)
alpha <- slope
for(j in 3:Tn){
  alpha[,j] <- alpha[,j] - alpha[,j-1]
}

set.seed(1234)
eta <- 0
y <- 0
for(j in 1:k){
  eta <- matrix(rep(alpha[j,], d[j]), nrow = d[j], byrow = TRUE)
  eta[,1] <- eta[,1] + rnorm(d[j], 0, psi1)
  y <- c(y,as.vector(Lambda%*%t(eta) + rnorm(Tn*d[j], 0, sigmai)))
}
y <- y[-1]

simdata <- data.frame(time = rep(x,n), y = y,
                      group =  as.factor(rep(group,each= Tn)),
                      id = rep(1:n, each = Tn))
```

Dernæst maksimeres log-likelihoodfunktionen ved brug af `hlme`.

```
start <- hlme(y~factor(time), random = ~1, subject = "id", ng = 1,
        data = simdata)
start.time <- Sys.time()
hlmemod <- hlme(y~factor(time), random = ~1, subject = "id",
                mixture = ~factor(time),
                ng = 2, data = simdata, B = start)
end.time <- Sys.time()
hlmetime <- end.time-start.time
```

For at gøre det samme ved brug af `optim`, definerer vi først log-likelihoodfunktionen, samt udregner startværdierne baseret på `start`.

```
likelihood <- function(theta){
  pii1 <- 1/(1+exp(-theta[1]))
  pii2 <- 1-pii1
  prod <- 0
  Sigma <- exp(theta[13])*diag(Tn)
  Psi <- diag(c(exp(theta[12]), rep(0,Tn-1)))
  for(i in 1:n){
    fi1 <- dmvnorm(as.vector(y[,i]), as.vector(Lambda%*%theta[2:6]),
                   Lambda%*%Psi%*%t(Lambda)+Sigma)
    fi2 <- dmvnorm(as.vector(y[,i]), as.vector(Lambda%*%theta[7:11]),
                   Lambda%*%Psi%*%t(Lambda)+Sigma)
    prod[i] <- sum(c(pii1,pii2)*c(fi1,fi2))
  }
  return(-sum(log(prod)))
}

alpha0 <- solve(Lambda)%*%(start$best[1]+c(0,start$best[2:5]))
se0 <- sqrt(vcov(start)[1,1]) + c(0,sqrt(diag(vcov(start)[2:5,2:5])))

alpha01 <- alpha0+(1-3/2)*se0
alpha02 <- alpha0+(2-3/2)*se0

start.time <- Sys.time()
optimmod <- optim(par = c(0, alpha01, alpha02, log(start$best[6]),
                  log(start$best[7]^2)), fn =likelihood, gr=NULL,
                  method = "BFGS")
end.time <- Sys.time()
BFGStime <- end.time-start.time

preds <- function(y){
  pi <- 1/exp(1+exp(-optimmod$par[1]))
  Sigma <- exp(optimmod$par[13])*diag(Tn)
  Psi <- diag(c(exp(optimmod$par[12]),rep(0,Tn-1)))
  fi1 <- dmvnorm(as.vector(y), as.vector(Lambda%*%optimmod$par[2:6]),
                 Lambda%*%Psi%*%t(Lambda)+Sigma)
  fi2 <- dmvnorm(as.vector(y), as.vector(Lambda%*%optimmod$par[7:11]),
                 Lambda%*%Psi%*%t(Lambda)+Sigma)
  pi*fi1/sum(c(pi,1-pi)*c(fi1,fi2))
}

class <- 0
for(i in 1:n){
  pred <- preds(y[,i])
  class[i] <- ifelse(pred >= 0.5, 1,2)
}
```

Vi gentager nu ovenstående for $N = 100$ datasæt, hvormed vi opnår resultaterne i anden del af eksemplet.

```
optime <- 0
matime <- 0
opite <- 0
maite <- 0
oplog <- 0
malog <- 0
```

```
N <- 100
k <- 2
n <- 100
for (i in 1:N){
  d <- table(sample(1:k,n,replace=TRUE))
  group <- 0
  for (j in 1:k){
    group <- c(group,rep(j,d[j]))
  }
  group <- group[-1]

  alpha <- matrix(runif(Tn*k,-10,10),nrow = k)

  eta <- 0
  y <- 0
  for(j in 1:k){
    eta <- matrix(rep(alpha[j,],d[j]),nrow = d[j], byrow = TRUE)
    eta[,1] <- eta[,1] + rnorm(d[j],0,psi1)
    y <- c(y,as.vector(Lambda %*% t(eta) + rnorm(Tn*d[j],0,sigmai)))
  }
  y <- y[-1]

  simdata <- data.frame(time = rep(x,n), y = y,
                        group =  as.factor(rep(group,each= Tn)),
                        id = rep(1:n, each = Tn))
  y <- matrix(simdata$y, nrow=Tn, byrow = FALSE)

  start <- hlme(y~factor(time), random = ~1, subject = "id", ng = 1,
                data = simdata)
  start.time <- Sys.time()
  hlmemod <- hlme(y~factor(time), random = ~1, subject = "id",
                  mixture = ~factor(time),
                  ng = 2, data = simdata, B= start)
  end.time <- Sys.time()
  matime[i] <- end.time-start.time
  malog[i] <- hlmemod$loglik
  maite[i] <- hlmemod$niter

  alpha0 <- start$best[1]+c(0,start$best[2],start$best[3],start$best[4],
                            start$best[5])
  alpha0 <- solve(Lambda)%*%alpha0

  se0 <- sqrt(vcov(start)[1,1]) + c(0,sqrt(diag(vcov(start)[2:5,2:5])))
  alpha01 <- alpha0+(1-3/2)*se0
  alpha02 <- alpha0+(2-3/2)*se0

  start.time <- Sys.time()
  optimmod <- optim(par = c(0,alpha01,alpha02,
                    log(start$best[6]),log(start$best[7]^2)), fn=likelihood,
                    gr=NULL, method = "BFGS")
  end.time <- Sys.time()
  optime[i] <- end.time-start.time
  oplog[i] <- - optimmod$value
  opite[i] <- optimmod$counts[2]
}
res <- data.frame(maite, opite, matime, optime, malog, oplog)
```

## B.3 Kode til Afsnit 4.3.4

For at lave simulationsstudiet i dette afsnit, bruger vi følgende R-funktion, hvor `K` er antal grupper, `N` er antal simulationer og `n` er antal individer.

```
consim <- function(k,N,n){
  Tn <- 5
  x <- 1:Tn
  psi1 <- 5
  sigmai <- 5
  Lambda <- matrix(c(rep(1,Tn),0,1:(Tn-1),
                     rep(0,2),1:(Tn-2),
                     rep(0,3),1:(Tn-3),
                     rep(0,4),1:(Tn-4)), nrow = Tn, byrow = FALSE)

  emc <- 0
  mac <- 0
  emlog <- 0
  malog <- 0
  emtime <- 0
  matime <- 0

  for (i in 1:N){
    d <- table(sample(1:k,n,replace=TRUE))
    group <- 0
    for (j in 1:k){
      group <- c(group,rep(j,d[j]))
    }
    group <- group[-1]

    alpha <- matrix(runif(Tn*k,-10,10),nrow = k)

    eta <- 0
    y <- 0
    for(j in 1:k){
      eta <- matrix(rep(alpha[j,],d[j]),nrow = d[j], byrow = TRUE)
      eta[,1] <- eta[,1] + rnorm(d[j],0,psi1)
      y <- c(y,as.vector(Lambda %*% t(eta) + rnorm(Tn*d[j],0,sigmai)))
    }
    y <- y[-1]

    simdata <- data.frame(time = rep(x,n), y = y,
                          group =  as.factor(rep(group,each= Tn)),
                          id = rep(1:n, each = Tn))

    start.time <- Sys.time()
    gmm.con.em <- flexmix(.~ .|id, k = k,
                          model = FLXMRlmm(y ~ factor(time), random = ~ 1,
                                           varFix = c(Random = TRUE,
                                                      Residual = TRUE)),
                          data = simdata,
                          control = list(iter.max = 2000, minprior = 0))
    end.time <- Sys.time()
    emtime[i] <- end.time - start.time

    start <- hlme(y ~ factor(time), subject = "id", random = ~ 1,
                  ng = 1, data = simdata, idiag = TRUE)
```

```
    start.time <- Sys.time()
    gmm.con.ma <- hlme(y ~ factor(time), subject = "id", random = ~ 1,
                       mixture = ~ factor(time),
                       ng = k, data = simdata, idiag = TRUE, B = start)
    end.time <- Sys.time()
    matime[i] <- end.time - start.time

    emc[i] <- gmm.con.em@iter
    mac[i] <- gmm.con.ma$niter
    emlog[i] <- gmm.con.em@logLik
    malog[i] <- gmm.con.ma$loglik
  }
  return(data.frame(emiter = emc, maiter=mac,
                    emtime = emtime, matime = matime,
                    emlog = emlog, malog = malog))
}
```

## B.4 Kode til Eksempel 4.4

Vi simulerer data til eksemplet og udregner AIC og BIC for $K = 1, \ldots, 5$ på følgende måde.

```
Tn <- 5
n <- 100
k <- 3
x <- 1:Tn
psi1 <- 3
sigmai <- 2
Lambda <- matrix(c(rep(1,Tn),
                   0,1:(Tn-1),
                   rep(0,2),1:(Tn-2),
                   rep(0,3),1:(Tn-3),
                   rep(0,4),1:(Tn-4)), nrow = Tn, byrow = FALSE)

alpha <- matrix(c(-1.97, -0.76, 0, 0, 0,
                   4.35, -3.06, 0, 0, 0,
                   4.33, -1.79, 0, 0, 0), nrow = k, byrow =TRUE)

set.seed(123)
d <- table(sample(1:k,n,replace=TRUE))
group <- 0
for (j in 1:k){
  group <- c(group,rep(j,d[j]))
}
group <- group[-1]

set.seed(1)
eta <- 0
y <- 0
for(j in 1:k){
  eta <- matrix(rep(alpha[j,],d[j]),nrow = d[j], byrow = TRUE)
  eta[,1] <- eta[,1] + rnorm(d[j],0,psi1)
  y <- c(y,as.vector(Lambda %*% t(eta) + rnorm(Tn*d[j],0,sigmai)))
}
y <- y[-1]
```

```
simdata <- data.frame(time = rep(x,n), y = y,
                      group =  as.factor(rep(group,each= Tn)),
                      id = rep(1:n, each = Tn))

gmmMixsim <- stepFlexmix(.~ .|id, k = 1:7, nrep = 100,
                         model = FLXMRlmm(y ~ factor(time), random = ~ 1,
                                          varFix = c(Random = TRUE,
                                          Residual = TRUE)),
                         data = simdata,
                         control = list(iter.max = 2000, minprior = 0))

plot(gmmMixsim, what = c("AIC", "BIC"))
plot(gmmMixsim, what = c("logLik"))
```

Vi laver da bootstrap likelihood ratio-test.

```
gmm <- getModel(gmmMixsim, which = 2)
gmmalt <- getModel(gmmMixsim, which = 3)

LR <- 2*gmmalt@logLik-2*gmm@logLik

beta21 <- parameters(gmm)[1,1]+c(0, parameters(gmm)[2:5,1])
beta22 <- parameters(gmm)[1,2]+c(0, parameters(gmm)[2:5,2])

alpha01 <- solve(Lambda)%*%beta21
alpha02 <- solve(Lambda)%*%beta22

k <- 2
alpha0 <- matrix(c(alpha01,alpha02), byrow = TRUE, nrow = k)

LRb <- 0
B <- 1000
for (i in 1:B){
  db <- table(sample(gmm@cluster, n, replace = TRUE))

  groupb <- 0
  for (j in 1:k){
    groupb <- c(groupb, rep(j, db[j]))
  }
  groupb <- groupb[-1]

  etab <- 0
  yb <- 0
  for(j in 1:k){
    etab <- matrix(rep(alpha0[j,], db[j]), nrow = db[j], byrow = TRUE)
    etab[,1] <- etab[,1] + rnorm(db[j], 0, parameters(gmm)[6,j])
    yb <- c(yb, as.vector(Lambda%*%t(etab)
             + rnorm(Tn*db[j], 0, parameters(gmm)[7,j])))
  }
  yb <- yb[-1]

  simdatab <- data.frame(time = rep(x,n), y = yb,
                         group =  as.factor(rep(groupb, each= Tn)),
                         id = rep(1:n, each = Tn))

  gmmboot <- stepFlexmix(.~ .|id, k = 2:3, nrep = 5,
                         model = FLXMRlmm(y ~ factor(time), random = ~ 1,
                                          varFix = c(Random = TRUE,
```

```
                                                    Residual = TRUE)),
                          data = simdatab,
                          control = list(iter.max = 2000, minprior = 0))

  LRb[i]<- 2*getModel(gmmboot,which = 2)@logLik
           -2*getModel(gmmboot,which = 1)@logLik
}

hist(LRb, prob = TRUE, breaks = 15)
abline(v=LR, lwd = 1.5)
mean(LRb >= LR)
```

## B.5 Kode til Afsnit 4.5

For at tilpasse en GMM ved brug af EM- og Marquardt-algoritmen for 1000 forskellige startværdier, for datasæt 1, bruger vi følgende.

```
data1 <- NT[NT$intervention == 1,]

gmmModK3 <- stepFlexmix(.~ .|id, k = 3, nrep = 1000,
                        model = FLXMRlmm(test ~ time + gender, random = ~ 1,
                                         varFix = c(Random = TRUE,
                                                    Residual = TRUE)),
                        data = data1,
                        control = list(iter.max = 2000, minprior = 0))

start <- hlme(test ~ factor(time)+factor(gender),
              subject = "id", random= ~ 1, ng = 1, data = data1)
gmmhlme1 <- gridsearch(rep = 1000, maxiter = 100, minit = start,
                       hlme(test ~ factor(time)+factor(gender),
                            subject = "id", random= ~ 1, ng = 3, data = data1,
                            mixture = ~ factor(gender)+factor(time),
                            nwg = TRUE))
```

Derudover er parameterestimaterne og middelvækstkurverne for `FLXMRlmm` udregnet og lavet ved brug af det følgende kode.

```
flxdata1 <- unique(cbind(id = data1$id, gender = data1$gender,
                         gmmgroup = gmmModK3@cluster))

meanboys_flx11 <- mean(flxdata1[flxdata1$gmmgroup == 1,]$gender)
meanboys_flx12 <- mean(flxdata1[flxdata1$gmmgroup == 2,]$gender)
meanboys_flx13 <- mean(flxdata1[flxdata1$gmmgroup == 3,]$gender)

beta1 <- parameters(gmmModK3)[1,1] + c(0, parameters(gmmModK3)[2:5,1])
beta2 <- parameters(gmmModK3)[1,2] + c(0, parameters(gmmModK3)[2:5,2])
beta3 <- parameters(gmmModK3)[1,3] + c(0, parameters(gmmModK3)[2:5,3])

Lambda <- matrix(c(rep(1,Tn),
                   0,1:(Tn-1),
                   rep(0,2),1:(Tn-2),
                   rep(0,3),1:(Tn-3),
                   rep(0,4),1:(Tn-4)), nrow = Tn, byrow = FALSE)

alpha1 <- solve(Lambda, beta1)
alpha2 <- solve(Lambda, beta2)
```

```
alpha3 <- solve(Lambda, beta3)

plot(NA, NA, xlim= c(1,5), ylim = c(7,22), ylab ="test score", xlab ="time",
     main = "intervention = 1")
points(beta1 + meanboys_flx11*parameters(gmmModK3)[6,1], type = "b",
       lty = "dotted")
points(beta2 + meanboys_flx12*parameters(gmmModK3)[6,2], type = "b",
       lty = "dashed")
points(beta3 + meanboys_flx13*parameters(gmmModK3)[6,3], type = "b",
       lty = "solid")
legend("topleft", legend = c("group = 1", "group = 2", "group = 3"),
       lty = c("solid", "dashed", "dotted" ))
```

Lignende kode er benyttet til at lave middelvækstkurver og udregne parameterestimater for `hlme`.

Vi prædikterer nu $\zeta_i$ og $Y_i$ for modellen tilpasset ved EM-algoritmen, ved det følgende.

```
predictzetai <- function(Psi, Sigma, yi, alphak, Gammak, xi){
  return((Psi%*% t(Lambda)%*%solve(Lambda %*% Psi %*% t(Lambda)+Sigma)
         %*%(yi-Lambda%*%(alphak+Gammak*xi)))[1])
}

Gamma1 <- c(parameters(gmmModK3)[6,1],0,0,0,0)
Gamma2 <- c(parameters(gmmModK3)[6,2],0,0,0,0)
Gamma3 <- c(parameters(gmmModK3)[6,3],0,0,0,0)

alphak <- 0
Gammak <- 0
zeta <- 0
Y <- matrix(rep(0,5), ncol=5)
for(i in 1:nrow(flxdata1)){
  id <- flxdata1$id[i]
  if(flxdata1$cluster[i] == 1){
    alphak <- alpha1
    Gammak <- Gamma1
  }
  if(flxdata1$cluster[i] == 2){
    alphak <- alpha2
    Gammak <- Gamma2
  }
  if(flxdata1$cluster[i] == 3){
    alphak <- alpha3
    Gammak <- Gamma3
  }
  xi <- as.integer(flxdata1[i,2])-1
  zeta[i] <- predictzetai(Psi = diag(c(parameters(gmmModK3)[7,1],0,0,0,0)),
               Sigma = diag(rep(parameters(gmmModK3)[8,1],5)),
               yi = data1[data1$id == id,]$test,
               alphak = alphak,
               Gammak = Gammak,
               xi = xi)
  hatY <- Lambda%*%(alphak + Gammak*xi + c(zeta[i],0,0,0,0))
  Y <- rbind(Y,t(hatY))
}
Y <- Y[-1,]
```

Yderligere udregner vi $p$-værdierne for Wald-testene for estimaterne af hældningerne i Tabel 4.13, for de tre grupper.

```
varcov <- vcov(gmmhlme1)

kontrast1 <- 0
SE1 <- 0
kontrast1[1] <- gmmhlme1$best[6]
SE1[1] <- varcov[6,6]
for (i in 2:4){
  kontrast1[i] <- gmmhlme1$best[3*i+3]-gmmhlme1$best[3*i]
  SE1[i] <- varcov[3*i+3,3*i+3]+varcov[3*i,3*i]-2*varcov[3*i+3,3*i]
}
w1 <- kontrast1^2/SE1
pval1 <- 1-pchisq(w1, 1)
tab1 <- t(rbind(kontrast1, SE1, w1, pval1)); tab1

kontrast2 <- 0
SE2 <- 0
kontrast2[1] <- gmmhlme1$best[7]
SE2[1] <- varcov[7,7]
for (i in 2:4){
  kontrast2[i] <- gmmhlme1$best[3*i+4]-gmmhlme1$best[3*i+1]
  SE2[i] <- varcov[3*i+4,3*i+4]+varcov[3*i+1,3*i+1]-2*varcov[3*i+4,3*i+1]
}
w2 <- kontrast2^2/SE2
pval2 <- 1-pchisq(w2, 1)
tab2 <- t(rbind(kontrast2, SE2, w2, pval2)); tab2

kontrast3 <- 0
SE3 <- 0
kontrast3[1] <- gmmhlme1$best[8]
SE3[1] <- varcov[8,8]
for (i in 2:4){
  kontrast3[i] <- gmmhlme1$best[3*i+5]-gmmhlme1$best[3*i+2]
  SE3[i] <- varcov[3*i+5,3*i+5]+varcov[3*i+2,3*i+2]-2*varcov[3*i+5,3*i+2]
}
w3 <- kontrast3^2/SE3
pval3 <- 1-pchisq(w3, 1)
tab3 <- t(rbind(kontrast3, SE3, w3, pval3)); tab3
```

For at tilpasse modellerne med samme hældninger, bruger vi koden, hvor vi ikke inkluderer `time` som en del af `mixture`.

```
start <- hlme(test ~ factor(time)+factor(gender), subject = "id", random = ~ 1,
              data = NT, subset = intervention == 1,
              idiag = TRUE, ng = 1)
model <- gridsearch(rep=100, maxiter = 100, minit = start,
                    hlme(test ~ factor(time)+factor(gender), subject = "id",
                         random = ~ 1, mixture = ~ factor(gender),
                         ng = 3, data = NT, subset = intervention == 1,
                         idiag = TRUE))
```