Eksamen i Matematisk Statistik, 22. august 2019

Vejledende besvarelse

Opgave 1

1. Da både X_1 og X_2 er standardnormalfordelte (og dermed specielt regulært normalfordelte), så følger det af EH Sætning 9.26, at $X = (X_1, X_2)^T$ er regulært normalfordelt på $\mathbb{R}^2$ med middelværdi $\xi = (0,0)^T$ og varians Σ. Da X_1 of X_2 er uafhængige, så følger specielt (fx. af EH Sætning 9.48), at kovarianserne $\Sigma_{12} = \Sigma_{21} = 0$ og vi har at $\Sigma_{11} = \Sigma_{22} = 1$. Dermed er argumenteret for, at X er standardnormalfordelt på $\mathbb{R}^2$. Det er også muligt at gennemføre dette argument ved at trækkes på EH Korollar 9.39.

   Det følger af EH Korollar 9.46 at $Y = (Y_1, Y_2)^T$ er normalfordelt med middelværdi

$$EY = \begin{pmatrix} 1 & 1 \\ a & b \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

   og varians

$$\Sigma_Y = \begin{pmatrix} 1 & 1 \\ a & b \end{pmatrix} \Sigma \begin{pmatrix} 1 & a \\ 1 & b \end{pmatrix} = \begin{pmatrix} 2 & a+b \\ a+b & a^2+b^2 \end{pmatrix}.$$

   Ifølge EH Korollar 9.43 er Y regulært normalfordelt hvis og kun hvis variansen Σ er invertibel. Det ses at determinanten af Σ er lig med

$$2 \cdot (a^2+b^2) - (a+b)^2 = a^2 + b^2 - 2ab = (a-b)^2.$$

   Vi konkluderer at Y er regulært normalfordelt netop hvis $a \neq b$.

2. Fra delopgave 1. ved vi, at covariansen mellem Y_1 og Y_2 er $a+b$, hvorfor det specielt følger af EH Sætning 9.48, at de er uafhængige for $a = 1$ og $b = -1$.

   For dette valg af konstanter har vi desuden, at Y_1 og Y_2 begge er normalfordelte med middelværdi 0 og varians 2. Dermed er $\frac{1}{\sqrt{2}}Y_1 \sim N(0,1)$ og $(\frac{1}{\sqrt{2}}Y_1)^2 = \frac{1}{2}Y_1^2$ er χ_1^2-fordelt.

   Tilsvarende er $\frac{1}{\sqrt{2}}Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2) \sim N(0,1)$ og uafhængig af $\frac{1}{2}Y_1^2$. Det følger at forholdet $\frac{\frac{1}{\sqrt{2}}Y_2}{\sqrt{\frac{1}{2}Y_1^2}} = \frac{X_1 - X_2}{|X_1 + X_2|}$ er t-fordelt med 1 frihedsgrad. Denne konstruktion har fx. været benyttet i forbindelse med udledning af konfidensintervaller og prædiktionsintervaller i den lineære normale model, eller mere direkte i HS opgave 9. Resultatet er også kendt fra MI eksempel 20.27.

Opgave 2

Lad $X_1,\ldots,X_n,Y_1,\ldots,Y_n$ være uafhængige stokastiske variable, hvor alle X_i er poissonfordelte med middelværdi λ, og alle Y_i er poissonfordelte med middelværdi λ^2. Her er $\lambda > 0$ en ukendt parameter.

1. Opskriv log-likelihoodfunktionen og vis at scorefunktionen er givet som

$$S_n(\lambda;x,y) = Dl_{x,y}(\lambda) = -\frac{S_x+2S_y}{\lambda} + n + 2n\lambda$$

For givne observationer $x = (x_1,\ldots,x_n) \in \mathbb{N}_0^n$ og $y = (y_1,\ldots,y_n) \in \mathbb{N}_0^n$ får vi (på nær en multiplikativ konstant) likelihoodfunktionen

$$L_{x,y}(\lambda) = \prod_{i=1}^n \left(\lambda^{x_i} e^{-\lambda}\right) \prod_{i=1}^n \left(\lambda^{2y_i} e^{-\lambda^2}\right) = \lambda^{S_x+2S_y} e^{-n\lambda} e^{-n\lambda^2}, \quad \lambda > 0$$

hvor $S_x = \sum_{i=1}^n x_i$ og $S_y = \sum_{i=1}^n y_i$ Derefter fås

$$\begin{aligned} l_{x,y}(\lambda) &= -\log L_{x,y}(\lambda) = -(S_x+2S_y)\log\lambda + n\lambda + n\lambda^2 \\ Dl_{x,y}(\lambda) &= -\frac{S_x+2S_y}{\lambda} + n + 2n\lambda. \end{aligned}$$

hvor jeg har brugt notationen $S_Y = \sum_{i=1}^n X_i$ og $S_Y = \sum_{i=1}^n Y_i$.

2. Vis at maksimaliseringsestimatoren (MLE) for λ er asymptotisk veldefineret og entydigt givet ved

$$\hat{\lambda} = \frac{-n + \sqrt{n^2 + 8n(S_X+2S_Y)}}{4n}$$

hvor $S_X = \sum_{i=1} X_i$ og $S_Y = \sum_{i=1}^n Y_i$.

Scoreligningen er

$$Dl_{x,y}(\lambda) = 0 \Leftrightarrow 2n\lambda^2 + n\lambda - S_x - 2S_y = 0,$$

dvs. en andengradsligning i λ. Diskriminanten er $n^2 + 8n(S_x+2S_y) > 0$, så der er to reelle løsninger:

$$\frac{-n \pm \sqrt{n^2 + 8n(S_x+2S_y)}}{4n}$$

Løsningen med "+" er altid ikke-negativ, mens løsningen med "–" altid er ikke-positiv, så scoreligningen har netop en løsning i parameterområdet hvis og kun hvis $S_X + S_Y > 0$ hvilket sker med en sandsynlighed gående mod 1, så estimatoren er asymptotisk veldefineret. Da vi endvidere har at $D^2 l_{x,y}(\lambda) > 0$ for alle $\lambda > 0$, er løsningen et minimum for $l_{x,y}$ på $(0,1)$. Samlet set får vi at

$$\hat{\lambda} = \frac{-n + \sqrt{n^2 + 8n(S_X+2S_Y)}}{4n}$$

er en entydig MLE for λ.

3. Vis at familien $\mathscr{P} = \{P_\lambda, \lambda > 0\}$ angivet ovenfor kan repræsenteres som en regulær eksponentiel familie af dimension 1. Angiv familiens kanoniske parameter, kanoniske stikprøvefunktion, samt grundmål.

   Vi lader $\theta = \log\lambda$ og skriver tætheden som

$$p(x_1,\ldots,x_n,y_1,\ldots,y_n;\theta) = \prod_i \frac{1}{x_i!}\prod_i \frac{1}{y_i!} e^{\theta(S_X+2S_Y)-n(e^\theta+e^{2\theta})}$$

   hvoraf det fremgår at grundmålet er $\prod_i \frac{1}{x_i!}\prod_i \frac{1}{y_i!}\cdot m$ hvor m er tællemål, den kanoniske stikprøvefunktion er (S_X+2S_Y), osv.

4. Vis at Fisherinformationen for λ er givet som

$$i_n(\lambda) = \frac{n}{\lambda} + 4n$$

   og angiv den asymptotiske fordeling af maksimaliseringsestimatoren $\hat{\lambda}_n$.

   Vi får Fisherinformationen

$$i(\lambda) = ED^2 l_{X,Y} = E\left(\frac{S_X+2S_Y}{\lambda^2}+2n\right) = \frac{n\lambda+2n\lambda^2}{\lambda^2}+2n = \frac{n}{\lambda}+4n.$$

   I en regulær eksponentiel familie er MLE asymptotisk normalfordelt med den inverse Fisherinformation som varians:

$$\hat{\lambda}_n \overset{\text{as}}{\sim} N\left(\lambda, \frac{\lambda}{n(1+4\lambda)}\right).$$

5. Betragt nu den alternative estimator $\tilde{\lambda}_n$ hvor

$$\tilde{\lambda}_n = \frac{S_X/n + \sqrt{S_Y/n}}{2},$$

   og vis, at den asymptotiske fordeling er

$$\tilde{\lambda}_n \overset{\text{as}}{\sim} N\left(\lambda, \frac{4\lambda+1}{16n}\right).$$

   Vi har at $S_X/n \overset{\text{as}}{\sim} N(\lambda,\lambda/n)$ og $S_Y/n \overset{\text{as}}{\sim} N(\lambda^2,\lambda^2/n)$. Deltametoden giver at

$$\sqrt{S_Y/n} \overset{\text{as}}{\sim} N\left(\sqrt{\lambda^2}, \frac{1}{(2\sqrt{\lambda^2})^2}\frac{\lambda^2}{n}\right) = N\left(\lambda, \frac{1}{4n}\right).$$

   Alt i alt får vi at

$$\tilde{\lambda}_n \overset{\text{as}}{\sim} N\left(\frac{\lambda+\lambda}{2}, \frac{1}{4}\left(\frac{\lambda}{n}+\frac{1}{4n}\right)\right) = N\left(\lambda, \frac{4\lambda+1}{16n}\right)$$

6. Sammenlign de to estimatorer $\hat{\lambda}_n$ og $\tilde{\lambda}_n$.

   Begge estimatorer er asymptotisk konsistente, men den alternative estimator har større asymptotisk varians end MLE. Forholdet mellem de asymptotiske varianser er

$$\frac{\mathbf{V}_{as}(\tilde{\lambda}_n)}{\mathbf{V}_{as}(\hat{\lambda}_n)} = \frac{4\lambda+1}{16}\frac{1+4\lambda}{\lambda} = \frac{(1+4\lambda)^2}{16\lambda} > 1.$$

   Det er særlig slemt for store værdier af λ, hvor forholdet går mod uendelig.

Opgave 3

1. Ved at lave en krydstabel over antallet af observationer for de forskellige kombinationer af faktorerne fås

G	V = -	V = +
A	20	20
B	20	20
C	20	20

G	K = 0	K = lav	K = hoj
A	0	20	20
B	0	20	20
C	40	0	0

   Vi ser at faktorerne `G` og `V` opfylder EH Sætning 14.8, hvorfor de er geometrisk ortogonale. Tilsvarende ses, at minimum af `G` og `K` er en faktor med to niveauer, som angiver om en jordlod er blevet kunstgødet (med `G = A` eller `G = B`) eller ej (svarende til `G = C`).

2. Det følger af EH (13.1) og Lemma 14.6 at

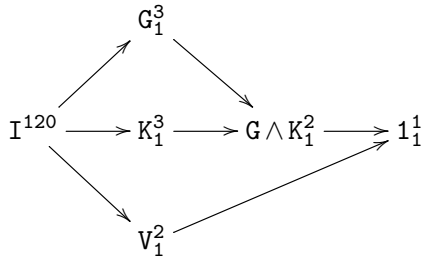
$$dim(L_G + L_V) = dim(L_G) + dim(L_V) - dim(L_{G\wedge V}) = 3 + 2 - 1 = 4,$$

   hvor det er benyttet, at minimum af faktorerne `G` og `V` er den konstante faktor `1`.

   Designet $\mathbb{G} = \{\mathtt{G}, \mathtt{V}, \mathtt{G}\wedge\mathtt{K}, \mathtt{K}, \mathtt{1}\}$ er geometrisk ortonalt og afsluttet over for dannelse af minimum. Dimensionerne af V_G-rummene, $G \in \mathbb{G}$, fra sætningen om den ortogonale dekomposition (EH: Sætning 14.21) kan let beregnes, hvorefter vi finder, at

$$\begin{aligned} dim(L_G + L_V + L_K) &= dim(V_G) + dim(V_V) + dim(V_{G\wedge K}) + dim(V_K) + dim(V_1) \\ &= 1 + 1 + 1 + 1 + 1 = 5. \end{aligned}$$

   Det kan her være nyttigt at støtte sig op ad et faktorstrukturdiagram, for at holde styr på ordningen af faktorerne



3. Der er tale om en lineær normal model, hvor middelværdiunderrrummet er parametriseret ved designmatricen A. Det følger derfor at EH Korollar 10.21 at

$$\hat{\beta} = (A^T A)^{-1} A^T X \sim N(\beta, \sigma^2 (A^T A)^{-1})$$

   og at

$$\hat{\sigma}^2 = \frac{||X - PX||^2}{120} = \frac{||X - A\hat{\beta}||^2}{120} \sim \frac{\sigma^2}{120} \chi^2_{120-k} - \text{fordelt.}$$

Opgave 4

1. Modellen kan udtrykkes ved at vektoren $X = (X_i)_{i \in I}$ bestående af målinger af den maksimale bøjningsvinkel `v` er normalfordelt på $\mathbb{R}^{64}$ med $\xi = EX \in L_{\texttt{ben} \times \texttt{fase}}$ og varians $VX = \sigma^2 + v_1^2 B_1 B_1^T$, hvor B_1 er effektmatricen hørende til parret $(\texttt{subj}, 1)$. Vi er ikke specifikt interesserede i lige præcis de 16 personer, som indgår i eksperimentet, hvorfor vi lader `subj` indgå med en tilfældig effekt.

   Kovariansmatricen for de 4 målinger på samme person kan udtrykkes som

$$\begin{pmatrix} \sigma^2 + v_1^2 & v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & \sigma^2 + v_1^2 & v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & \sigma^2 + v_1^2 & v_1^2 \\ v_1^2 & v_1^2 & v_1^2 & \sigma^2 + v_1^2 \end{pmatrix}.$$

   En mere avanceret løsning består i at inddrage vekselvirkningerne `subj` $\times$ `ben` og `subj` $\times$ `fase` i modellen. Alle faktorer der involverer `subj` bør indgå med tilfældig effekt, således at der bliver op til tre tilfældige effekter svarende til hvert af effektparrene $(\texttt{subj}, 1)$, $(\texttt{subj} \times \texttt{ben}, 1)$ og $(\texttt{subj} \times \texttt{fase}, 1)$. Lader vi B_1, B_2 og B_3 betegne effektmatricerne hørende til de tre effektpar, så kan modellen udtrykkes ved at $X = (X_i)_{i \in I}$ er normalfordelt på $\mathbb{R}^{64}$ med $\xi = EX \in L_{\texttt{ben} \times \texttt{fase}}$ og $VX = \sigma^2 I + v_1^2 B_1 B_1^T + v_2^2 B_2 B_2^T + v_3^2 B_3 B_3^T$.

   Organiseres de 4 målinger for samme person, så vi først har de to målinger for `ben = D` og dernæst de to målinger for `ben = N`, så bliver kovariansmatricen

$$\begin{pmatrix} \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 + v_2^2 & v_1^2 + v_3^2 & v_1^2 \\ v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 & v_1^2 + v_3^2 \\ v_1^2 + v_3^2 & v_1^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 & v_1^2 + v_2^2 \\ v_1^2 & v_1^2 + v_3^2 & v_1^2 + v_2^2 & \sigma^2 + v_1^2 + v_2^2 + v_3^2 \end{pmatrix}.$$

   Det betragtes som en tilstrækkelig besvarelse, selvom man vælger ikke at inddrage vekselvirkninger med `subj` som tilfældige effekter i modellen.

2. Modellen fittes i R med følgende kode (angivelse af kode ikke påkrævet)

```r
library(lme4)
knee <- read.table(file = "MatStatAug2019.txt", header = T)
mod1 <- lmer(v ~ ben * fase + (1|subj), data = knee)
mod1


## Linear mixed model fit by REML ['lmerMod']
## Formula: v ~ ben * fase + (1 | subj)
##    Data: knee
## REML criterion at convergence: 406.8466
## Random effects:
##  Groups   Name        Std.Dev.
##  subj     (Intercept) 3.989
##  Residual             5.720
## Number of obs: 64, groups:  subj, 16
## Fixed Effects:
##    (Intercept)            benN       fasesving  benN:fasesving
##        12.6222         -0.7094         41.8129         -1.6462
```

Estimaterne (REML!) for variansparametrene er $\hat{\sigma} = 5.720$ og $\hat{v}_1 = 3.989$. Den forventede maksimale bøjningsvinkel for det dominante ben (`ben = D`) bliver 12.6222 grader (`fase = kontakt`) og $12.6222 + 41.8129$ grader (`fase = sving`).

3. Modellen fra foregående delspørgsmål genfittes nu med en parametrisering, hvor man direkte kan aflæse forskellen i maksimal bøjningsvinkel (`ben = N` fratrukket `ben = D`) for hver af de faser. På baggrund af estimater og konfidensintervaller for tilvæksterne kan man nu konkludere på de ønskede forskelle.

```
mod1refit <- lmer(v ~ fase + ben : fase - 1 + (1|subj), data = knee)
mod1refit


## Linear mixed model fit by REML ['lmerMod']
## Formula: v ~ fase + ben:fase - 1 + (1 | subj)
##    Data: knee
## REML criterion at convergence: 406.8466
## Random effects:
##  Groups   Name        Std.Dev.
##  subj     (Intercept) 3.989
##  Residual             5.720
## Number of obs: 64, groups:  subj, 16
## Fixed Effects:
##      fasekontakt         fasesving  fasekontakt:benN    fasesving:benN
##          12.6222           54.4351           -0.7094           -2.3555


confint(mod1refit)


## Computing profile confidence intervals ...


##                      2.5 %    97.5 %
## .sig01            1.910588  6.469507
## .sigma            4.591540  6.862495
## fasekontakt       9.245042 15.999395
## fasesving        51.057917 57.812270
## fasekontakt:benN -4.625514  3.206736
## fasesving:benN   -6.271671  1.560580
```

Vi ser at der hverken er signifikante forskelle i bøjningsvinklen for `ben = N` og `ben = D` når man betragter den maksimale vinkel under kontakt-fasen (-0.71 [-4.63-3.21]) eller under spring-fasen (forskel: -2.36 [-6.27-1.56]).

Det er også muligt at lave et (simultant) test for, om maksimal bøjningsvinkel varierer mellem `ben = N` og `ben = D`.

```
mod2 <- lmer(v ~ fase + (1|subj), data = knee)
anova(mod2, mod1)


## refitting model(s) with ML (instead of REML)


## Data: knee
## Models:
```

```
## mod2: v ~ fase + (1 | subj)
## mod1: v ~ ben * fase + (1 | subj)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod2  4 427.56 436.20 -209.78   419.56
## mod1  6 430.01 442.96 -209.00   418.01 1.5528      2     0.4601
```

Her fås en likelihoodratio-teststørrelsen $LRT = 1.5528$ som i en tabel over χ^2- fordelingen med 2 frihedsgrader oversættes til en approksimativ P-værdi på 0.4601. Vi kan altså ikke afvise en hypotese om, at der ikke er forskel på maksimal bøjningsvinkel for `ben = N` og `ben = D`.