Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Reeksamen i Matematisk Statistik
22. august 2019

4 timers skriftlig prøve. Alle hjælpemidler tilladt (inkl. computer uden netforbindelse). Besvarelsen må skrives med blyant. Opgavesættet består af 4 opgaver med i alt 14 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Data til Opgave 4 findes på den udleverede USB-nøgle i filen `MatStatAug2019.txt`*. USB-nøglen skal returneres efter eksamen, men udelukkende for at den kan genbruges. Filer på denne USB-nøgle vil således ikke kunne indgå som en del af besvarelsen.*

Opgave 1

Lad X_1 og X_2 være uafhængige og normalfordelte, begge med middelværdi 0 og varians 1. Sæt $Y_1 = X_1 + X_2$ og $Y_2 = aX_1 + bX_2$ hvor $a, b \in \mathbb{R}$ er konstanter, og lad $Y = (Y_1, Y_2)^T$.

1. Argumenter for, at Y er normalfordelt på $\mathbb{R}^2$ med

$$\Sigma = \mathrm{Var}(Y) = \begin{pmatrix} 2 & a+b \\ a+b & a^2+b^2 \end{pmatrix}$$

   og angiv betingelser på a, b som sikrer, at Y er regulært normalfordelt.

2. Argumenter for at hvis $a = 1$ og $b = -1$, så er Y_1 og Y_2 uafhængige. Find dernæst de marginale fordelinger af $Y_1^2/2$ og $(X_1 - X_2)/|X_1 + X_2|$.

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

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

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

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

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

3. Lad P_λ betegne fordelingen af $(X_i, Y_i), i = 1, \ldots, n$. Vis at familien $\mathcal{P} = \{P_\lambda, \lambda > 0\}$ kan repræsenteres som en regulær eksponentiel familie af dimension 1. Angiv familiens kanoniske parameter, kanoniske stikprøvefunktion, samt grundmål.

4. Vis at Fisherinformationen for λ er givet som

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

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

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 dens asymptotiske fordeling er

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

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

Opgave 3

Det planlægges at udføre et dyrkningsforsøg på 120 jordlodder, hvor der blandt andet skal afprøves to typer af kunstgødning (`A,B`). Desuden udtages et antal jordlodder, hvor der ikke skal anvendes kunstgødning (kontrolforsøg). Vi indfører faktoren `G` med tre niveauer `A, B, C`, hvor niveauet `C` betegner kontrolforsøg helt uden gødning. For de jordlodder som kunstgødes (med enten `G = A` eller `G = B`) udvælges tilfældigt halvdelen af jordlodderne, hvor der blandes ekstra kalk i gødningen. Dette er givet ved faktoren kalk (`K`) med niveauerne `hoj` og `lav`. Jordlodder uden gødning (med `G = C`) tilsættes slet ikke kalk, hvilket angives ved niveauet `K = 0`. Endelig indgår faktoren vanding (`V`) med to niveauer `+` (ekstra vanding) og `-` (ingen ekstra vanding). En oversigt over antallet af jordlodder hvor der benyttes de forskellige kombinationer af faktorerne `G, V` og `K` er angivet i følgende tabel

G	V	K	antal jordlodder
A	+	lav	10
A	+	hoj	10
A	-	lav	10
A	-	hoj	10
B	+	lav	10
B	+	hoj	10
B	-	lav	10
B	-	hoj	10
C	+	0	20
C	-	0	20

1. Argumenter for at faktorerne `G` og `V` er geometrisk ortogonale og find minimum af faktorerne `G` og `K`.
2. Angiv dimensionen af de additive underrum $L_{\mathtt{G}} + L_{\mathtt{V}}$ og $L_{\mathtt{G}} + L_{\mathtt{V}} + L_{\mathtt{K}}$.

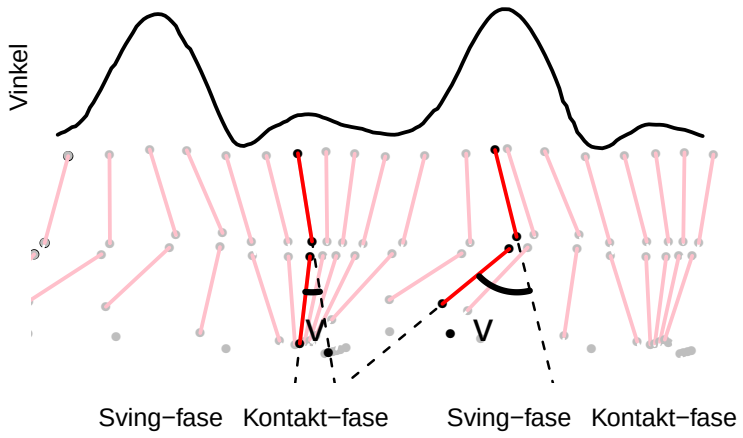
Lad X_i betegne målingen (=udbyttet) på den i-te jordlod. Vi antager, at X_i'erne er uafhængige og normalfordelte $\sim N(\xi_i, \sigma^2)$. Antag at $\xi = (\xi_i)_{i=1}^{120} \in A\beta$, hvor A er en $120 \times k$ matrix af fuld rang ($k < 120$). Betragt den statistiske model, hvor $\beta \in \mathbb{R}^k$ og $\sigma^2 > 0$ er ukendte parametre og $X = (X_i)_{i=1}^{120}$ er observationen.

3. Angiv formler for maksimaliseringsestimatoren (MLE) for β og σ^2 udtrykt ved A (som altså ikke er nærmere specificeret) og X. Angiv også fordelingen af maksimaliseringsestimatoren.

Opgave 4

I moderne ganglaboratorier kan man foretage målinger på forsøgspersoner, mens de går på et løbebånd. Vinklen mellem lårben og underben er særlig velegnet til at studere forskelle i gangmønstret fx. mellem grupper af individer som har været igennem forskellige genoptræningsforløb efter en knæskade.

Figuren nedenfor viser bøjningsvinklen mellem lårben og underben fra en måleserie for en forsøgsperson, der går med konstant hastighed. Vi interesserer os i denne opgave for den maksimale bøjning af knæet i løbet af et skridt (markeret med `v` på figuren).



Vi betragter data fra et forsøg, hvor 16 personer (givet ved faktoren `subj`) hver har fået foretaget to sæt målinger af den maksimale bøjningsvinkel `v`. Udstyret til måling af vinklen monteres hver gang kun på det ene ben, og der benyttes betegnelsen `ben=D` til at angive, at måleudstyret har været monteret på personens foretrukne springben. Niveauet `ben=N` angiver, at måleudstyret sidder på det ben, som personen normalt ikke bruger til afsæt ved spring.

Bemærk at hver gangcyklus består af to skridt (kaldet faser): `kontakt`-fasen hvor foden, på benet med måleudstyr, har berøring med jorden, og `sving`-fasen hvor foden og benet med måleudstyret løftes fremad. Hver måleserie omfatter en hel gangcyklus og giver således anledning til to maksimale bøjningsvinkler (angivet ved responsvariablen `v`), hvor faktoren `fase` med niveauerne `sving` og `kontakt` holder styr på, fra hvilken fase af gangcyklen målingen er taget.

Vi minder om, at der foretages to måleserier (for `ben = D` og `ben = N`) fra hver forsøgsperson, og at der for hver måleserie (dvs. `ben`) foretages to målinger af den maksimale bøjningsvinkel svarende til `kontakt`-fasen og `sving`-fasen. Der er således totalt 4 målinger for hver person (`subj`).

Data til opgaven er venligst stillet til rådighed af Tatiana Sato fra Federal University of São Carlos i Brasilien. Data er blevet gjort tilgængelige på vedlagte USB-nøgle. Data kan indlæses med kommandoen

```
knee <- read.table(file = "MatStatAug2019.txt", header = T)
```

og de første linjer i datafilen ser ud som følger

```
head(knee, 10)
```

```
##    subj ben    fase        v
## 1    ae   D   sving 41.70150
## 2    ae   D kontakt -4.22100
## 3    ae   N   sving 48.29400
## 4    ae   N kontakt  5.05350
## 5    al   D   sving 53.35200
## 6    al   D kontakt 13.26150
## 7    al   N   sving 43.59150
## 8    al   N kontakt 11.54455
## 9    as   D   sving 53.52750
## 10   as   D kontakt 10.05750
```

1. Opstil en passende varianskomponentmodel der kan benyttes til analyse af, hvordan den maksimale bøjningsvinkel `v` varierer for forskellige kombinationer af faktorerne `ben` og `fase`. Angiv variansmatricen for de fire målinger, der hører til personen givet ved `subj == ae`. Der kan være flere fornuftige modeller til løsning af opgaven.
2. Fit modellen fra delspørgsmål 1. i R og angiv variansestimaterne samt estimaterne for den forventede/gennemsnitlige maksimale vinkel for springbenet (`ben = D`) for hver af de to faser (`kontakt, sving`).
3. Undersøg om der er belæg for at hævde, at den maksimale bøjningsvinkel er forskellig for springben og ikke-springben. Husk at skrive en konklusion på din analyse.