Statistiske metoder

EM-algoritmen, Faktoranalyse og Bayesianske netværk

P7 - Efterår 2016

Gruppe G2-101b

Aalborg Universitet
Institut for Matematiske Fag

Institut for Matematiske Fag
Aalborg Universitet
http://www.math.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:
Statistiske metoder

Tema:
EM-algoritmen
Faktoranalyse
Bayesianske netværk

Projektperiode:
Efterårssemester 2016

Projektgruppe:
G2-101b

Deltagere:
Mikkel Findinge
Kristoffer Segerstrøm Mørk

Vejleder:
Søren Højsgaard

Oplagstal: 3

Sidetal: 57

Afleveringsdato:
21. December 2016

Abstract:

Dette projekt omhandler tre forskellige emner, EM-algoritmen, Faktoranalyse og Bayesianske netværk. Formålet er at præsentere den grundlæggende teori indenfor hver af de tre emner, hvorefter der gives et eksempel på, hvordan teorien anvendes i praksis. EM-algoritmen anvendes til at estimere parametrene i et givent datasæt `misdat.RData`, hvori der indgår nogle manglende observationer. Datasættet `misdat.RData` antages at følge en bivariat normalfordeling. Derudover bestemmes EM-algoritmen for en faktoranalyse model med henblik på at estimere parametrene i sådanne modeller. Der opstilles to forskellige faktoranalyse modeller for datasættet `math.csv` med henholdvis én og to faktorer. I disse modeller anvendes EM-algortimen til at estimere parametrene. Ydermere præsenteres den grundlæggende teori bag Bayesianske netværk, hvori der indgår relevante algoritmer til at håndtere samt estimere forskellige sandsynligheder i et Bayesiansk netværk. Variablene i datasættet `math.csv` inddeles i 3 niveauer, hvorpå teorien bag Bayesianske netværk anvendes til at udføre probabilistisk inferens.

Indhold

Forord

Aalborg Universitet, 21. December 2016

Vi henviser til ligninger med to tal i parentes. Derudover noteres figurer og tabeller med to tal uden parentes. Det første tal angiver kapitlet, og det andet tal angiver det nummer, som ligningen, figuren eller tabellen har i kapitlet. For sætninger, definitioner osv. angives referencerne med tre tal: det første for kapitel, det næste for afsnittet og det sidste for hvilket nummer i afsnittet, der er tale om. Kilder henvises til med forfatter(-ens/-nes) efternavn komma året for udgivelsen omkranset af kantede parenteser. Eventuelle sidetal angives til sidst i henvisningen, hvis der citeres direkte fra kilden. I litteraturlisten står kilderne i alfabetisk rækkefølge efter efternavn. Vi betragter som udgangspunkt altid vektorer som værende søjlevektorer, og de noteres med fed skrift. Ydermere siges en matrix at være negativ definit, hvis dens egenværdier er mindre end 0. Derudover siges en matrix at være negativ semidefinit, hvis dens egenværdier er mindre eller lig 0. Det tilsvarende gælder for positiv definit og semidefinit.

Mikkel Findinge
<mfindi13@student.aau.dk>

Kristoffer Segerstrøm Mørk
<kmark13@student.aau.dk>

1. Indledning

Statistik består af et væld af metoder til behandling af data. Disse data kan af flere grunde være ukomplette i form af manglende observationer. Der findes metoder til at estimere det komplette datasæt, således man ikke skal smide brugbart data væk, fordi en enkelt observation mangler. En af disse metoder er den såkaldte "Expectation Maximization"-algoritme (EM-algoritmen), som er det første af tre emner, der bliver behandlet i dette projekt.

Det andet emne, som behandles i dette projekt, er faktoranalyse (FA). Dette er en statistisk metode til at beskrive "højdimensionelt" data ud fra lavere dimensionelle latente variable. Projektet har til hensigt at estimere parametrene i faktoranalysen ved brug af førnævnte EM-algoritme.

Tredje og sidste emne behandlet i dette projekt er Bayesianske netværk. Dette er en statistisk metode til at lave probabilistisk inferens. Dette vil sige, at det er sandsynligheder, som estimeres. I teorien for Bayesianske netværk udarbejdes forskellige algoritmer, som anvendes til at udføre probabilistik inferens.

Der lægges vægt på teorien i de tre overordnede emner. Afslutningsvis for hvert emne laves dog et eksempel, hvor teorien anvendes på en praktisk problemstilling. Til dette anvendes programmet `R`.

2. EM-Algoritmen

Dette kapitel er baseret på [Dempster et al., 1977] og har til formål at introducere den grundlæggende teori bag EM-algoritmen samt anvendelse af denne. Først defineres EM-algoritmen, hvorefter konvergensen af denne præciseres. Dernæst introduceres eksponentielle familier samt EM-algoritmens anvendelse på disse. Afslutningsvis opstilles og anvendes EM-algoritmen på en bivariat normalfordeling. Bemærk, at der i dette kapitel henvises til resultater i [Dempster et al., 1977] med parenteser i titlen på sætninger, lemmaer og korollarer.

2.1 Teorien bag EM-Algoritmen

I følgende afsnit betragter vi to stokastiske vektorer $\mathbf{Y} \in \mathcal{Y}$ og $\mathbf{X} \in \mathcal{X}$, hvor $\mathcal{Y} \subseteq \mathbb{R}^n$ og $\mathcal{X} \subseteq \mathbb{R}^m$. Vi lader $\mathbf{y}$ være en realisering af $\mathbf{Y}$ og kalder den for *det observerede data*. Derudover lader vi $\mathbf{x}$ være en realisering af $\mathbf{X}$, som kun er observeret indirekte gennem $\mathbf{y}$. Dette betyder, at vi kan betragte $\mathbf{x}$ som værende en realisering fra $\mathcal{X}(\mathbf{y}) \subseteq \mathcal{X}$, hvor mængden $\mathcal{X}(\mathbf{y})$ afhænger af det observerede data $\mathbf{y}$. Fremover kalder vi $\mathbf{x}$ for *det fulde data*.
Vi betegner de parametriserede tæthedsfunktioner for $\mathbf{X}$ og $\mathbf{Y}$ som henholdsvis $f(\mathbf{x};\boldsymbol{\theta})$ og $g(\mathbf{y};\boldsymbol{\theta})$, hvor der om parameteren $\boldsymbol{\theta}$ gælder, at $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Relationen mellem $\mathbf{Y}$ og $\mathbf{X}$ betyder, at tæthedsfunktionen for $\mathbf{Y}$ kan bestemmes ud fra tæthedsfunktionen for $\mathbf{X}$ på følgende måde:

$$g(\mathbf{y};\boldsymbol{\theta}) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x};\boldsymbol{\theta})\, d\mathbf{x}_{\mathbf{y}}, \tag{2.1}$$

hvor $\mathbf{x}_{\mathbf{y}}$ består af de indgange i $\mathbf{x}$, som varierer over $\mathcal{X}(\mathbf{y})$. Den betingede tæthedsfunktion for $\mathbf{X} \mid \mathbf{Y}$ er dermed givet ved:

$$k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}) = \frac{f(\mathbf{x};\boldsymbol{\theta})}{g(\mathbf{y};\boldsymbol{\theta})}. \tag{2.2}$$

Dette medfører, at log-likelihoodfunktionen for $\mathbf{y}$ kan skrives som:

$$\ell_{obs}(\boldsymbol{\theta};\mathbf{y}) = \ell(\boldsymbol{\theta};\mathbf{x}) - \log(k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta})). \tag{2.3}$$

I det efterfølgende vil vi sommetider betragte to parametre $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Hvis det fremgår tydeligt ud af konteksten, at der kun er tale om en enkelt parameter, vil vi dog stadig anvende $\boldsymbol{\theta} \in \Theta$.

Definition 2.1.1
Lad $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Da er $Q\colon \Theta \times \Theta \to \mathbb{R}$ givet ved:

$$Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathrm{E}\left[\ell(\boldsymbol{\theta}_1;\mathbf{X}) \mid \mathbf{y};\boldsymbol{\theta}_2\right].$$

Derudover er funktionen $H\colon \Theta \times \Theta \to \mathbb{R}$ givet ved:

$$H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathrm{E}\left[\log(k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_1)) \mid \mathbf{y};\boldsymbol{\theta}_2\right].$$

Definition 2.1.1 og (2.3) giver, at

$$\ell_{obs}(\boldsymbol{\theta}_1;\mathbf{y}) = Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) - H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2), \tag{2.4}$$

hvor $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

I det efterfølgende definerer vi to iterative algoritmer, som kan anvendes til at maksimere $\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ med hensyn til $\boldsymbol{\theta}$. Den første algoritme kaldes for *Expectation–Maximization-algoritmen* og forkortes EM-algoritmen.

Definition 2.1.2
Lad $t = 0, 1, \ldots$ og $\boldsymbol{\theta} \in \Theta$. Iteration t i EM-algoritmen er givet ved trinene:

E: Bestem $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$, hvor $\boldsymbol{\theta}^{(t)}$ er givet.

M: Sæt $\boldsymbol{\theta}^{(t+1)}$ til at være det $\boldsymbol{\theta} \in \Theta$, som maksimerer $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$.

For den t'te iteration gælder ydermere, at

$$M(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)},$$

hvor M er en afbildning fra Θ til Θ.

E-trinet i EM-algoritmen går altså ud på at bestemme den betinget middelværdi af $\ell(\boldsymbol{\theta};\mathbf{x})$ givet det observerede data $\mathbf{y}$ og en fast parameterværdi $\boldsymbol{\theta}^{(t)}$. M-trinet går derefter ud på at bestemme det $\boldsymbol{\theta}$, som maksimerer den evaluerede betinget middelværdi fra E-trinet. Det fundne maksimum sættes til at være $\boldsymbol{\theta}^{(t+1)}$, hvorefter algoritmen kan itereres. Bemærk, at afbildningen $M\colon \Theta \to \Theta$ er et trin i EM-algoritmen.

EM-algoritmen er særdeles anvendelig i de tilfælde, hvor $\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ er besværlig at maksimere med hensyn til $\boldsymbol{\theta}$, men $\ell(\boldsymbol{\theta};\mathbf{x})$ er simpel at maksimere. Dette skyldes, at EM-algoritmen anvender $\ell(\boldsymbol{\theta};\mathbf{x})$ til at maksimere $\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$. Den anden algoritme, som vi definerer, er en generalisering af EM-algoritmen.

Definition 2.1.3
Lad $\boldsymbol{\theta} \in \Theta$. En iterativ algoritme med en afbildning $M\colon \Theta \to \Theta$ er en *generaliseret EM-algoritme* (GEM-algoritme), hvis

$$Q(M(\boldsymbol{\theta});\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta};\boldsymbol{\theta}),$$

for alle $\boldsymbol{\theta} \in \Theta$.

Ved EM-algoritmen vælges $\boldsymbol{\theta}^{(t+1)}$ som maksimum af $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$, hvorimod $\boldsymbol{\theta}^{(t+1)}$ under en GEM-algoritme blot skal opfylde, at $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$ ikke aftager.

Bemærk, at M-trinet i EM-algoritmen betyder, at følgende skal være opfyldt for alle $\boldsymbol{\theta} \in \Theta$:

$$Q\left(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\right) = Q\left(M(\boldsymbol{\theta}^{(t)});\boldsymbol{\theta}^{(t)}\right) \geq Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right).$$

Dermed gælder ovenstående ulighed også for $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$, hvorfor EM-algoritmen er en GEM-algoritme. Vi vil derfor fremover tage udgangspunkt i en GEM-algoritme med mindre, at der fremgår andet. I det efterfølgende afsnit vil vi udlede egenskaber for en GEM-algoritme, som begrunder anvendeligheden for en sådan algoritme.

2.2 Konvergens af en GEM-algoritme

I dette afsnit præciseres konvergensen af henholdsvis $\ell_{obs}(\boldsymbol{\theta}^{(t)};\mathbf{y})$ og $\boldsymbol{\theta}^{(t)}$ under anvendelse af en GEM-algortime samt betingelserne herfor. Afsnittet er baseret på [Dempster et al., 1977].

Lemma 2.2.1 (Lemma 1)
Lad parret $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta \times \Theta$, så er

$$H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) \leq H(\boldsymbol{\theta}_2;\boldsymbol{\theta}_2),$$

hvor der er lighed, hvis og kun hvis $k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1)$=$k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_2)$ næsten overalt.

Bevis
Vi viser, at $H(\boldsymbol{\theta}_2;\boldsymbol{\theta}_2) - H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) \geq 0$. Betragt

$$\begin{aligned}
H(\boldsymbol{\theta}_2;\boldsymbol{\theta}_2) - H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) &= \mathrm{E}\left[\log\left(k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_2)\right) \mid \mathbf{y};\boldsymbol{\theta}_2\right] \\
&\quad - \mathrm{E}\left[\log\left(k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_1)\right) \mid \mathbf{y};\boldsymbol{\theta}_2\right] \\
&= \mathrm{E}\left[\log\left(\frac{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_2)}{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_1)}\right) \,\middle|\, \mathbf{y};\boldsymbol{\theta}_2\right] \\
&= \mathrm{E}\left[-\log\left(\frac{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_1)}{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_2)}\right) \,\middle|\, \mathbf{y};\boldsymbol{\theta}_2\right] \\
&\geq -\log\left(\mathrm{E}\left[\frac{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_1)}{k(\mathbf{X} \mid \mathbf{y};\boldsymbol{\theta}_2)} \,\middle|\, \mathbf{y};\boldsymbol{\theta}_2\right]\right) \\
&= -\log\left(\int_{\mathcal{X}(\mathbf{y})} \frac{k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1)}{k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_2)} k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_2)\, d\mathbf{x}_{\mathbf{y}}\right) \\
&= -\log(1) = 0,
\end{aligned}$$

hvor vi har anvendt Jensens ulighed. Hvis og kun hvis relationen følger af ovenstående udregninger. ■

Sætning 2.2.2 (Theorem 1)
For enhver GEM-algoritme gælder der, at

$$\ell_{obs}(M(\boldsymbol{\theta});\mathbf{y}) \geq \ell_{obs}(\boldsymbol{\theta};\mathbf{y}) \quad \text{for alle } \boldsymbol{\theta} \in \Theta,$$

hvor ligheden holder, hvis og kun hvis følgende to betingelser er opfyldt:

1) $Q(M(\boldsymbol{\theta});\boldsymbol{\theta}) = Q(\boldsymbol{\theta};\boldsymbol{\theta})$.

2) $k(\mathbf{x} \mid \mathbf{y};M(\boldsymbol{\theta})) = k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta})$ næsten overalt.

Bevis
Betragt

$$\ell_{obs}\left(M(\boldsymbol{\theta});\mathbf{y}\right)-\ell_{obs}\left(\boldsymbol{\theta};\mathbf{y}\right)=\left(Q(M(\boldsymbol{\theta});\boldsymbol{\theta})-Q(\boldsymbol{\theta};\boldsymbol{\theta})\right)+\left(H(\boldsymbol{\theta};\boldsymbol{\theta})-H(M(\boldsymbol{\theta});\boldsymbol{\theta})\right). \quad (2.5)$$

Defintion 2.1.3 giver, at den første parentes i (2.5) er ikke-negativ. Derudover giver Lemma 2.2.1, at anden parentes i (2.5) er ikke-negativ. Dermed er uligheden i Sætning 2.2.2 bevist. Hvis og kun hvis relationen følger af Definition 2.1.3, Lemma 2.2.1 samt (2.5). ■

Sætning 2.2.2 medfører, at $\ell_{obs}(\boldsymbol{\theta}^{(t)};\mathbf{y})$ er voksende under anvendelse af en GEM-algoritme. Denne egenskab er særdeles vigtig, da den begrunder anvendeligheden af en GEM-algoritme. De to følgende korollarer er konsekvenser af Sætning 2.2.2.

Korollar 2.2.3 (Corollary 1)
Antag, at der for et $\boldsymbol{\theta}^*\in\Theta$ gælder, at $\ell_{obs}(\boldsymbol{\theta}^*;\mathbf{y})\geq\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ for alle $\boldsymbol{\theta}\in\Theta$. Så gælder der for enhver GEM-algoritme, at

- $\ell_{obs}(M(\boldsymbol{\theta}^*);\mathbf{y})=\ell_{obs}(\boldsymbol{\theta}^*;\mathbf{y})$.
- $Q(M(\boldsymbol{\theta}^*);\boldsymbol{\theta}^*)=Q(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$.
- $k(\mathbf{x}\mid\mathbf{y},M(\boldsymbol{\theta}^*))=k(\mathbf{x}\mid\mathbf{y},\boldsymbol{\theta}^*)$ næsten overalt.

Bevis
Sætning 2.2.2 giver, at $\ell_{obs}\left(M(\boldsymbol{\theta}^*);\mathbf{y}\right)\geq\ell_{obs}\left(\boldsymbol{\theta}^*;\mathbf{y}\right)$. Dette er dog i modstrid med antagelsen om, at $\ell_{obs}(\boldsymbol{\theta}^*;\mathbf{y})\geq\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ for alle $\boldsymbol{\theta}\in\Theta$ i Korollar 2.2.3, hvorfor $M(\boldsymbol{\theta}^*)=\boldsymbol{\theta}^*$. Af Sætning 2.2.2 følger de to sidste betingelser. ■

Korollar 2.2.4 (Corollary 2)
Hvis der for et vilkårligt $\boldsymbol{\theta}^*\in\Theta$ gælder, at $\ell_{obs}(\boldsymbol{\theta}^*;\mathbf{y})>\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ for alle $\boldsymbol{\theta}\in\Theta$, hvor $\boldsymbol{\theta}\neq\boldsymbol{\theta}^*$. Så gælder der for alle GEM-algoritmer, at $M(\boldsymbol{\theta}^*)=\boldsymbol{\theta}^*$.

Bevis
Beviset er baseret på lignende argumenter, som i beviset for Korollar 2.2.3. ■

Den efterfølgende sætning beskriver hvilke betingelser, der skal være opfyldt for, at $\boldsymbol{\theta}^{(t)}$ konvergerer under anvendelse af en GEM-algoritme. Det skal dog bemærkes, at sætningen er ændret i forhold til den oprindelige *Theorem 2* i [Dempster et al., 1977]. Dette skyldes, at beviset for *Theorem 2* er ugyldigt grundet fejlagtig brug af trekantsuligheden. Denne fejlagtige brug af trekanstuligheden kan korrigeres ved at ændre betingelse 2) i *Theorem 2*. Den efterfølgende sætning er den korrigerede version af *Theorem 2*.

Sætning 2.2.5 (Theorem 2*)
Lad $t=0,1,2,...$ og $\boldsymbol{\theta}^{(t)}$ være bestemt af en GEM-algoritme. Følgen af $\boldsymbol{\theta}^{(t)}$ konvergerer mod et $\boldsymbol{\theta}^*\in\overline{\Theta}$, hvis $\boldsymbol{\theta}^{(t)}$ opfylder følgende:

1) Følgen af $\ell_{obs}(\boldsymbol{\theta}^{(t)};\mathbf{y})$ er begrænset.
2) $Q(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)})-Q(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)})\geq c\|\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)}\|$ for en skalar $c>0$ og for alle t.

Bevis
Vi antager, at $\boldsymbol{\theta}^{(t)}$ opfylder de to betingelser i Sætning 2.2.5. Af betingelse 1) i Sætning 2.2.5 og Sætning 2.2.2 har vi, at følgen af $\ell_{obs}(\boldsymbol{\theta}^{(t)};\mathbf{y})$ konvergerer mod et $\ell^*_{obs} < \infty$. Konvergensen giver, at der for alle $\varepsilon > 0$ eksisterer et T sådan, at der for alle $t \geq T$ og alle $r \geq 1$ gælder:

$$\sum_{j=1}^{r}\left(\ell_{obs}\left(\boldsymbol{\theta}^{(t+j)};\mathbf{y}\right) - \ell_{obs}\left(\boldsymbol{\theta}^{(t+j-1)};\mathbf{y}\right)\right) = \ell_{obs}\left(\boldsymbol{\theta}^{(t+r)};\mathbf{y}\right) - \ell_{obs}\left(\boldsymbol{\theta}^{(t)};\mathbf{y}\right) < \varepsilon. \tag{2.6}$$

Lemma 2.2.1 og (2.5) giver, at der for $j \geq 1$ gælder:

$$0 \leq Q\left(\boldsymbol{\theta}^{(t+j)};\boldsymbol{\theta}^{(t+j-1)}\right) - Q\left(\boldsymbol{\theta}^{(t+j-1)};\boldsymbol{\theta}^{(t+j-1)}\right) \leq \ell_{obs}\left(\boldsymbol{\theta}^{(t+j)};\mathbf{y}\right) - \ell_{obs}\left(\boldsymbol{\theta}^{(t+j-1)};\mathbf{y}\right).$$

Hvormed (2.6) giver, at

$$\sum_{j=1}^{r}\left(Q\left(\boldsymbol{\theta}^{(t+j)};\boldsymbol{\theta}^{(t+j-1)}\right) - Q\left(\boldsymbol{\theta}^{(t+j-1)};\boldsymbol{\theta}^{(t+j-1)}\right)\right) < \varepsilon \tag{2.7}$$

for alle $t \geq T$ og alle $r \geq 1$, hvor hvert led i summen er ikke-negativ. Betingelse 2) i Sætning 2.2.5 samt (2.7) giver, at

$$c\|\boldsymbol{\theta}^{(t+r)} - \boldsymbol{\theta}^{(t)}\| \leq \sum_{j=1}^{r} c\|\boldsymbol{\theta}^{(t+j)} - \boldsymbol{\theta}^{(t+j-1)}\| < \varepsilon, \tag{2.8}$$

hvor vi har anvendt trekantsuligheden. Ved at lade $r \to \infty$ medfører (2.8), at følgen bestående af $\boldsymbol{\theta}^{(t)}$ konvergerer mod et $\boldsymbol{\theta}^* \in \overline{\Theta}$. ■

I det efterfølgende betragtes $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ som variable, og $\boldsymbol{\theta} \in \Theta$ betragtes som en vilkårlig fast værdi.

Lemma 2.2.6 (Lemma 2)
For alle $\boldsymbol{\theta} \in \Theta$ gælder følgende:

$$\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}) \,\middle|\, \mathbf{y};\boldsymbol{\theta}\right] = \frac{\partial}{\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta};\boldsymbol{\theta}) = \mathbf{0}. \tag{2.9}$$

Derudover gælder:

$$\mathrm{Var}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}) \,\middle|\, \mathbf{y};\boldsymbol{\theta}\right] = \frac{\partial^2}{\partial\boldsymbol{\theta}_2^T\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta};\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta};\boldsymbol{\theta}). \tag{2.10}$$

Bevis
Vi har, at der for alle $\boldsymbol{\theta}_1 \in \Theta$ gælder:

$$\int_{\mathcal{X}(\mathbf{y})} k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1)\, d\mathbf{x}_{\mathbf{y}} = 1. \tag{2.11}$$

Ved at differentiere (2.11) med hensyn til $\boldsymbol{\theta}_1$ får vi under passende regularitetsbetingelser følgende:

$$\mathbf{0} = \int_{\mathcal{X}(\mathbf{y})}\frac{\partial}{\partial\boldsymbol{\theta}_1}k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1)\, d\mathbf{x}_{\mathbf{y}} = \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta}_1)\, d\mathbf{x}_{\mathbf{y}}. \tag{2.12}$$

Dernæst betrager vi:

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\,\middle|\,\mathbf{y};\boldsymbol{\theta}_2\right] &= \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2)\,d\mathbf{x}_\mathbf{y}\\
&= \frac{\partial}{\partial\boldsymbol{\theta}_1}\int_{\mathcal{X}(\mathbf{y})}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2)\,d\mathbf{x}_\mathbf{y}\\
&= \frac{\partial}{\partial\boldsymbol{\theta}_1}\mathrm{E}\left[\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\mid\mathbf{y};\boldsymbol{\theta}_2\right]\\
&= \frac{\partial}{\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2),
\end{aligned}$$

hvor $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta$. Vælges $\boldsymbol{\theta}_2=\boldsymbol{\theta}_1$ får vi (2.9) ved hjælp af (2.12). Til at bevise (2.10) differentierer vi (2.12) med hensyn til $\boldsymbol{\theta}_1^T$, hvormed vi under passende regularitetsbetingelser får:

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial\boldsymbol{\theta}_1^T}\int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1)\,d\mathbf{x}_\mathbf{y}\\
&= \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1)\,d\mathbf{x}_\mathbf{y}\\
&+ \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1^T}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1)\,d\mathbf{x}_\mathbf{y}\\
&= \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1)\,d\mathbf{x}_\mathbf{y}+\mathrm{Var}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\,\middle|\,\mathbf{y};\boldsymbol{\theta}_1\right],
\end{aligned}\tag{2.13}$$

hvor vi har anvendt, at $\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\,\middle|\,\mathbf{y};\boldsymbol{\theta}_1\right]=\mathbf{0}$, hvilket følger af (2.12). Hvis (2.13) evalueres i $\boldsymbol{\theta}_1=\boldsymbol{\theta}\in\Theta$ fås:

$$\mathbf{0}=\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta};\boldsymbol{\theta})+\mathrm{Var}\left[\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}))\,\middle|\,\mathbf{y};\boldsymbol{\theta}\right].\tag{2.14}$$

Dernæst differentieres $H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)$ med hensyn til $\boldsymbol{\theta}_1$ og $\boldsymbol{\theta}_2^T$:

$$\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\theta}_2^T\partial\boldsymbol{\theta}_1}H(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) &= \int_{\mathcal{X}(y)}\frac{\partial^2}{\partial\boldsymbol{\theta}_2^T\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2)\,d\mathbf{x}_\mathbf{y}\\
&= \int_{\mathcal{X}(y)}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_2^T}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2)\right\}d\mathbf{x}_\mathbf{y}\\
&= \int_{\mathcal{X}(y)}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_1}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_1))\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}_2^T}\log(k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2))\right\}k(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}_2)\,d\mathbf{x}_\mathbf{y}.
\end{aligned}$$

Evalueres dette for $\boldsymbol{\theta}_1=\boldsymbol{\theta}_2$ følger den ene lighed i (2.10). Den anden lighed følger af (2.14). ■

Sætning 2.2.7 (Theorem 3)
Lad $t=0,1,2,\ldots$ og $\boldsymbol{\theta}^{(t)}$ være bestemt af en GEM-algoritme. Hvis

$$\frac{\partial}{\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\right)=\mathbf{0},\tag{2.15}$$

så eksisterer der for alle t et $\boldsymbol{\theta}_0^{(t+1)}$ på linjestykket mellem $\boldsymbol{\theta}^{(t)}$ og $\boldsymbol{\theta}^{(t+1)}$, sådan at

$$Q\left(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\right)-Q\left(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)}\right)=-\frac{1}{2}\left(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)}\right)^T\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}_0^{(t+1)};\boldsymbol{\theta}^{(t)}\right)\left(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)}\right). \tag{2.16}$$

Derudover gælder der, at hvis $\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}_0^{(t+1)};\boldsymbol{\theta}^{(t)}\right)$ er negativ definit og følgen af $\ell_{obs}\left(\boldsymbol{\theta}^{(t)};\mathbf{y}\right)$ er begrænset, så konvergerer følgen af $\boldsymbol{\theta}^{(t)}$ mod et $\boldsymbol{\theta}^*\in\overline{\Theta}$.

Bevis
Beviset følger af en første ordens Taylorudvikling af $Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}^{(t)})$ omkring $\boldsymbol{\theta}^{(t+1)}$ med restled:

$$\begin{aligned}Q\left(\boldsymbol{\theta}_1;\boldsymbol{\theta}^{(t)}\right)&=Q\left(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\right)+\frac{\partial}{\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\right)\left(\boldsymbol{\theta}_1-\boldsymbol{\theta}^{(t+1)}\right)\\&\quad+\frac{1}{2}\left(\boldsymbol{\theta}_1-\boldsymbol{\theta}^{(t+1)}\right)^T\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}_0^{(t+1)};\boldsymbol{\theta}^{(t)}\right)\left(\boldsymbol{\theta}_1-\boldsymbol{\theta}^{(t+1)}\right),\end{aligned}$$

hvor $\boldsymbol{\theta}_0^{(t+1)}$ ligger på linjen mellem $\boldsymbol{\theta}_1$ og $\boldsymbol{\theta}^{(t+1)}$. Ved at lade $\boldsymbol{\theta}_1=\boldsymbol{\theta}^{(t)}$ samt anvende antagelsen i (2.15) får vi:

$$Q(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)})=Q(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)})+\frac{1}{2}\left(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t+1)}\right)^T\frac{\partial^2}{\partial\boldsymbol{\theta}_1^T\partial\boldsymbol{\theta}_1}Q\left(\boldsymbol{\theta}_0^{(t+1)};\boldsymbol{\theta}^{(t)}\right)\left(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t+1)}\right).$$

Heraf følger (2.16). ■

Den sidste del angående konvergensen af $\boldsymbol{\theta}^{(t)}$ i Sætning 2.2.7 kan ikke bevises, da den er baseret på det fejlagtige bevis for *Theorem 2* i [Dempster et al., 1977]. Selvom Sætning 2.2.5 er en korrigeret version af *Theorem 2* retter den ikke op på ugyldigheden af beviset for konvergensen i Sætning 2.2.7.

2.2.1 Konvergensraten af GEM-algoritmen

I følgende afsnit bestemmes konvergensraten af en GEM-algoritme. Fra [McLachlan og Krishnan, 2008] har vi, at konvergensraten for en iterativ proces er givet ved:

$$r=\lim_{t\to\infty}\frac{\|\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|}.$$

Derudover gælder der, at en GEM-algoritme kan beskrives som en lineær iteration med rate matrix $\frac{\partial}{\partial\boldsymbol{\theta}_1}M(\boldsymbol{\theta}^*)$, hvis $\boldsymbol{\theta}^{(t)}$ er tilstrækkeligt tæt på $\boldsymbol{\theta}^*$. Dette følger af en første ordens Taylorudvikling af $M(\boldsymbol{\theta}_1)$ omkring $\boldsymbol{\theta}^*$:

$$M(\boldsymbol{\theta}_1)\approx M(\boldsymbol{\theta}^*)+\frac{\partial}{\partial\boldsymbol{\theta}_1}M(\boldsymbol{\theta}^*)(\boldsymbol{\theta}_1-\boldsymbol{\theta}^*).$$

Ved at evaluere dette i $\boldsymbol{\theta}_1=\boldsymbol{\theta}^{(t)}$ samt anvende, at $M(\boldsymbol{\theta}^{(t)})=\boldsymbol{\theta}^{(t+1)}$, og $M(\boldsymbol{\theta}^*)=\boldsymbol{\theta}^*$, får vi:

$$\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^*\approx\frac{\partial}{\partial\boldsymbol{\theta}_1}M(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*), \tag{2.17}$$

hvorfor en GEM-algoritme kan betragtes som en lineær iteration med rate matrix $\frac{\partial}{\partial \boldsymbol{\theta}_1} M(\boldsymbol{\theta}^*)$, når $\boldsymbol{\theta}^{(t)}$ er tilstrækkeligt tæt på $\boldsymbol{\theta}^*$. Den efterfølgende sætning kan anvendes til at bestemme rate matricen $\frac{\partial}{\partial \boldsymbol{\theta}_1} M(\boldsymbol{\theta}^*)$ mere specifikt i forhold til en vilkårlig GEM-algoritme.

Sætning 2.2.8 (Theorem 4)
Lad $t = 0, 1, 2, ...$ og $\boldsymbol{\theta}^{(t)}$ være bestemt af en GEM-algoritme. Hvis $\boldsymbol{\theta}^{(t)}$ opfylder følgende tre betingelser:

1) $\boldsymbol{\theta}^{(t)}$ konvergerer mod et $\boldsymbol{\theta}^* \in \overline{\Theta}$.

2) $\frac{\partial}{\partial \boldsymbol{\theta}_1} Q\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\right) = \mathbf{0}$.

3) $\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} Q\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\right)$ er negativ definit.

Så gælder der, at $\frac{\partial}{\partial \boldsymbol{\theta}_1} \ell_{obs}\left(\boldsymbol{\theta}^*; \mathbf{y}\right) = \mathbf{0}$, $\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} Q\left(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*\right)$ er negativ definit samt følgende:

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} M(\boldsymbol{\theta}^*) = \frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} H\left(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*\right) \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} Q\left(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*\right) \right\}^{-1}. \tag{2.18}$$

Beviset for Sætning 2.2.8 udelades, men det kan findes i [Dempster et al., 1977, s. 9]. Beviset for *Theorem 4* i Dempster et al. [1977] er baseret på Lemma 2.2.6. Bemærk, at vi har følgende relationer under tilstrækkelige regularitetsbetingelser:

$$\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) = \mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} \log\left(k(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\theta}^*)\right) \middle| \mathbf{y}; \boldsymbol{\theta}^* \right] = -\mathcal{I}_k(\boldsymbol{\theta}^*; \mathbf{y}) \tag{2.19}$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} Q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) = \mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1} \ell(\boldsymbol{\theta}^*; \mathbf{x}) \middle| \mathbf{y}; \boldsymbol{\theta}^* \right] = -\mathcal{I}_c(\boldsymbol{\theta}^*; \mathbf{y}), \tag{2.20}$$

hvor vi har anvendt Definition 2.1.1, og hvor $\mathcal{I}_c(\boldsymbol{\theta}; \mathbf{y})$ er det betingede Fisher informationskriterie for $\boldsymbol{\theta}$ baseret på $\mathbf{x}$ givet $\mathbf{y}$, og $\mathcal{I}_k(\boldsymbol{\theta}; \mathbf{y})$ kaldes for informationskriteriet for det uobserverede data. Ud fra (2.18), (2.19) samt (2.20) har vi dermed, at

$$\frac{\partial}{\partial \boldsymbol{\theta}_1} M(\boldsymbol{\theta}^*) = \mathcal{I}_k(\boldsymbol{\theta}^*; \mathbf{y}) \left(\mathcal{I}_c(\boldsymbol{\theta}^*; \mathbf{y})\right)^{-1}. \tag{2.21}$$

Den generelle teori bag en GEM-algoritme er nu blevet præsenteret. I det efterfølgende vil vi opstille EM-algoritmen for en eksponentiel familie.

2.3 Eksponentielle familier

Formålet med dette afsnit er at specificere EM-algoritmen for eksponentielle familier. Til dette introduceres først grundlæggende teori vedrørende eksponentielle familier, hvorefter EM-algoritmen opstilles. Afsnittet er baseret på [Dempster et al., 1977] og [McLachlan og Krishnan, 2008].

Tæthedsfunktionen for en eksponentiel familie er givet ved:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{a(\boldsymbol{\theta})} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x})\right) b(\mathbf{x}), \tag{2.22}$$

hvor $s(\mathbf{x})$ kaldes for den sufficiente stikprøvefunktion, og $\boldsymbol{\theta} \in \Theta$ er en kanonisk parameter. Log-likelihoodfunktionen for en eksponentiel familie er dermed

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}) - \log(a(\boldsymbol{\theta})). \tag{2.23}$$

Hvis vi lader $\mathbf{x}$ være det fulde data, og $\mathbf{y}$ være det observerede data, følger tæthedsfunktionen for $\mathbf{X} \mid \mathbf{Y}$ af (2.2):

$$k(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\theta}) = \frac{a(\boldsymbol{\theta})}{a(\boldsymbol{\theta})} \cdot \frac{\exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}))b(\mathbf{x})}{\exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{y}))b(\mathbf{y})} = \frac{1}{a(\boldsymbol{\theta} \mid \mathbf{y})} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x})\right) b(\mathbf{x}), \tag{2.24}$$

hvor vi har anvendt (2.1) samt, at

$$a(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}))b(\mathbf{x})\, d\mathbf{x}_{\mathbf{y}} = \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{y}))b(\mathbf{y}).$$

Hvis $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, så siges den eksponentielle familie at være kanonisk. Derudover siges den at være krum, hvis $\dim(\boldsymbol{\theta}) < \dim(\boldsymbol{\eta}(\boldsymbol{\theta}))$. Bemærk, at hvis en eksponentiel familie er krum, medfører det, at den er ikke-kanonisk. I det efterfølgende vil vi betragte kanoniske eksponentielle familier med henblik på at opstille EM-algoritmen for disse. Ved hjælp af Definition 2.1.1 kan vi finde et udtryk for $Q(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2)$ givet, at fordelingen af $\mathbf{X}$ tilhører de kanoniske eksponentielle familier:

$$\begin{aligned} Q(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) &= \mathrm{E}\left[\ell(\boldsymbol{\theta}_1; \mathbf{X}) \mid \mathbf{y}; \boldsymbol{\theta}_2\right] \\ &= \mathrm{E}\left[-\log(a(\boldsymbol{\theta}_1)) + \boldsymbol{\theta}_1^T s(\mathbf{X}) \mid \mathbf{y}; \boldsymbol{\theta}_2\right] \\ &= -\log(a(\boldsymbol{\theta}_1)) + \boldsymbol{\theta}_1^T \mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y}; \boldsymbol{\theta}_2\right]. \end{aligned} \tag{2.25}$$

Ud fra denne kan vi opstille EM-algoritmen for en kanonisk eksponentiel familie.

Sætning 2.3.1
EM-algoritmen for en kanonisk eksponentiel familie er givet ved:

E: Bestem $s^{(t)} = \mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}\right]$.

M: Bestem $\boldsymbol{\theta}^{(t+1)} \in \Theta$, som løsningen til $\mathrm{E}\left[s(\mathbf{X}); \boldsymbol{\theta}\right] = s^{(t)}$.

Bevis
E-trinet følger af Definition 2.1.2 og (2.25). Hernæst bevises M-trinet. Ud fra Definition 2.1.2 samt (2.25) opnås M-trinet ved at maksimere $-\log(a(\boldsymbol{\theta})) + \boldsymbol{\theta}^T s^{(t)}$ med hensyn til $\boldsymbol{\theta}$. Betragt

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\theta}} \left(-\log a(\boldsymbol{\theta}) + \boldsymbol{\theta}^T s^{(t)}\right) = -\frac{1}{a(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} a(\boldsymbol{\theta}) + s^{(t)}.$$

Ved hjælp af omskrivning og passende regularitetsbetingelser får vi, at

$$\begin{aligned} s^{(t)} &= \frac{1}{a(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} a(\boldsymbol{\theta}) = \frac{1}{a(\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{X}} \exp\left(\boldsymbol{\theta}^T s(\mathbf{x})\right) b(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\mathcal{X}} s(\mathbf{x}) \exp\left(\boldsymbol{\theta}^T s(\mathbf{x})\right) \frac{b(\mathbf{x})}{a(\boldsymbol{\theta})}\, d\mathbf{x} \\ &= \int_{\mathcal{X}} s(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta})\, d\mathbf{x} \\ &= \mathrm{E}\left[s(\mathbf{X}); \boldsymbol{\theta}\right], \end{aligned}$$

hvor vi har anvendt, at $f(\mathbf{x};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T s(\mathbf{x})\right)\frac{b(\mathbf{x})}{a(\boldsymbol{\theta})}$ og $a(\boldsymbol{\theta}) = \int_{\mathcal{X}} \exp(\boldsymbol{\theta}^T s(\mathbf{x}))b(\mathbf{x})\,d\mathbf{x}$. Dermed er M-trinet i EM-algoritmen for en kanonisk eksponentiel familie bevist. ■

For en generel eksponentiel familie er $Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2)$ givet ved:

$$Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = -\log(a(\boldsymbol{\theta}_1)) + \boldsymbol{\eta}(\boldsymbol{\theta}_1)^T \mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y};\boldsymbol{\theta}_2\right], \tag{2.26}$$

hvilket følger af tilsvarende udregninger som for (2.25). Ud fra denne kan vi opstille EM-algoritmen for en generel eksponentiel familie.

Sætning 2.3.2
EM-algoritmen for en generel eksponentiel familie er givet ved:

E: Bestem $s^{(t)} = \mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y};\boldsymbol{\theta}^{(t)}\right]$.

M: Bestem $\boldsymbol{\theta}^{(t+1)} \in \Theta$, som løsningen til $\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\}\mathrm{E}\left[s(\mathbf{X});\boldsymbol{\theta}\right] = \{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\}s^{(t)}$.

Bevis
E-trinet følger af (2.26). M-trinet for en generel eksponentiel familie er udledt ved hjælp af lignende udregninger som for en kanonisk eksponentiel familie, hvilke fremgår i beviset for Sætning 2.3.1. ■

Bemærk, at E-trinet for en generel eksponentiel familie er identisk med E-trinet for en kanonisk eksponentiel familie.

Vi vil nu bestemme log-likelihoodfunktionen for det observerede data $\mathbf{y}$. Ved at anvende (2.3), (2.22) og (2.24) får vi:

$$\ell_{obs}(\boldsymbol{\theta};\mathbf{y}) = \ell(\boldsymbol{\theta};\mathbf{x}) - \log(k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta})) = -\log(a(\boldsymbol{\theta})) + \log(a(\boldsymbol{\theta} \mid \mathbf{y})).$$

Bemærk, at $a(\boldsymbol{\theta}) = \int_{\mathcal{X}} \exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}))b(\mathbf{x})\,d\mathbf{x}$, da $f(\cdot;\boldsymbol{\theta})$ skal integrere til 1. Vi vil nu bestemme scorefunktionen:

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{obs}(\boldsymbol{\theta};\mathbf{y}) &= -\frac{\partial}{\partial\boldsymbol{\theta}}\log(a(\boldsymbol{\theta})) + \frac{\partial}{\partial\boldsymbol{\theta}}\log(a(\boldsymbol{\theta} \mid \mathbf{y})) \\
&= -\frac{1}{a(\boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}}a(\boldsymbol{\theta}) + \frac{1}{a(\boldsymbol{\theta} \mid \mathbf{y})}\frac{\partial}{\partial\boldsymbol{\theta}}a(\boldsymbol{\theta} \mid \mathbf{y}) \\
&= -\frac{1}{a(\boldsymbol{\theta})}\int_{\mathcal{X}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}s(\mathbf{x})\exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}))b(\mathbf{x})\,d\mathbf{x} \\
&\quad + \frac{1}{a(\boldsymbol{\theta} \mid \mathbf{y})}\int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}s(\mathbf{x})\exp(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}))b(\mathbf{x})\,d\mathbf{x}_{\mathbf{y}} \\
&= -\int_{\mathcal{X}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}s(\mathbf{x})f(\mathbf{x};\boldsymbol{\theta})\,d\mathbf{x} + \int_{\mathcal{X}(\mathbf{y})}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}s(\mathbf{x})k(\mathbf{x} \mid \mathbf{y};\boldsymbol{\theta})\,d\mathbf{x}_{\mathbf{y}} \\
&= -\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}\mathrm{E}\left[s(\mathbf{X});\boldsymbol{\theta}\right] + \left\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\right\}\mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y};\boldsymbol{\theta}\right].
\end{aligned}$$

Dette medfører, at MLE for $\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ er det $\boldsymbol{\theta}$, som opfylder $\{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\}\mathrm{E}\left[s(\mathbf{X});\boldsymbol{\theta}\right] = \{\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\}\mathrm{E}\left[s(\mathbf{X}) \mid \mathbf{y};\boldsymbol{\theta}\right]$. Bemærk, at dette betyder, at EM-algoritmen opnår MLE for $\ell_{obs}(\boldsymbol{\theta};\mathbf{y})$ givet, at startparameteren vælges tilstrækkeligt i forhold til MLE. I det følgende gives der et eksempel på, hvordan en bivariat normalfordeling kan omskrives til en eksponentiel familie.

Eksempel 2.3.3
Betragt en bivariat normalfordeling $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, hvor $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ og $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^T$. Bemærk, at $\sigma_{21} = \sigma_{12}$, da Σ er kovariansmatrix. Kovariansmatricen Σ antages at være invertibel, hvilket betyder, at $-1 < \rho < 1$, hvor $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ er korrelationen. Den inverse af Σ er givet ved:

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}.$$

Tæthedsfunktionen for $\mathbf{X}$ kan opskrives som en ikke-kanonisk eksponentiel familie:

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}) &= (2\pi)^{-1} |\Sigma|^{-1/2} \exp\left(\frac{-1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \\ &= b(\mathbf{x}) |\Sigma|^{-1/2} \exp\left(\frac{-1}{2} \left(\mathbf{x}^T \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \Sigma^{-1} \mathbf{x} \right) \right) \\ &= b(\mathbf{x}) |\Sigma|^{-1/2} \exp\left(\frac{-1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \right) \exp\left(\frac{-1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}^T \Sigma^{-1} \mathbf{x} \right) \\ &= \frac{b(\mathbf{x})}{a(\boldsymbol{\theta})} \exp\left(|\Sigma|^{-1} \left(\frac{-\sigma_{22}}{2} x_1^2 - \frac{\sigma_{11}}{2} x_2^2 + \sigma_{12} x_1 x_2 + (\mu_1 \sigma_{22} - \mu_2 \sigma_{12}) x_1 + (-\mu_1 \sigma_{12} + \mu_2 \sigma_{11}) x_2 \right) \right), \end{aligned}$$

hvor $a(\boldsymbol{\theta}) = |\Sigma|^{1/2} \exp(\frac{1}{2} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu})$, $b(\mathbf{x}) = (2\pi)^{-1}$. Vi lader:

$$s(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}, \quad \boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \mu_1 \sigma_{22} - \mu_2 \sigma_{12} \\ -\mu_1 \sigma_{12} + \mu_2 \sigma_{11} \\ -\frac{\sigma_{22}}{2} \\ -\frac{\sigma_{11}}{2} \\ \sigma_{12} \end{bmatrix},$$

hvormed $f(\mathbf{x}; \boldsymbol{\theta})$ kan opskrives som tæthedsfunktionen for en eksponentiel familie:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{b(\mathbf{x})}{a(\boldsymbol{\theta})} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}) \right). \qquad \square$$

I det efterfølgende afsnit anvendes den grundlæggende teori vedrørende EM-algoritmen og eksponentielle familier på et specifikt tilfælde af en eksponentiel familie.

2.4 Anvendelse af EM-algoritmen

I dette afsnit opstilles og anvendes EM-algoritmen på n stokastiske vektorer $\mathbf{X}_i = (X_{1\,i}, X_{2\,i})^T \sim N_2(\boldsymbol{\mu}, \Sigma)$, hvor $\boldsymbol{\mu}$ og Σ er givet som i Eksempel 2.3.3. Dette kan betragtes som en ny stokastisk vektor $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$. Da $\mathbf{X}_i$ og $\mathbf{X}_j$ er uafhængige for $i \neq j$, er tæthedsfunktionen for $\mathbf{X}$ givet ved:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) &= f_{\mathbf{X}_1}(\mathbf{x}_1; \boldsymbol{\theta}) \cdot \ldots \cdot f_{\mathbf{X}_n}(\mathbf{x}_n; \boldsymbol{\theta}) = \frac{(2\pi)^{-n}}{a(\boldsymbol{\theta})^n} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \sum_{i=1}^{n} s(\mathbf{x}_i) \right) \\ &= \frac{b(\mathbf{x})^n}{a(\boldsymbol{\theta})^n} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T s(\mathbf{x}) \right), \end{aligned} \tag{2.27}$$

hvor vi har anvendt Eksempel 2.3.3, $s(\mathbf{x}) = \sum_{i=1}^{n} s(\mathbf{x}_i)$ samt $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^T$. Den sufficiente stikprøvefunktion for $\mathbf{X}$ er dermed givet ved:

$$s(\mathbf{x}) = \sum_{i=1}^{n} s(\mathbf{x}_i) = \begin{bmatrix} \sum_{i=1}^{n} x_{1\,i} \\ \sum_{i=1}^{n} x_{2\,i} \\ \sum_{i=1}^{n} x_{1\,i}^2 \\ \sum_{i=1}^{n} x_{2\,i}^2 \\ \sum_{i=1}^{n} x_{1\,i} x_{2\,i} \end{bmatrix}. \tag{2.28}$$

Fremover noteres indgang j i $s(\mathbf{x})$ med s_j, hvor $j = 1, \ldots, 5$.

I det efterfølgende betragter vi en realisering af $\mathbf{X}$, hvilken noteres $\mathbf{x}$. Vi antager, at vi kun kender nogle vilkårlige indgange $x_{j\,i}$ i realiseringen $\mathbf{x}$, hvor $j = 1, 2$ og $i = 1, \ldots, n$. Disse kendte indgange i $\mathbf{x}$ samles i en vektor $\mathbf{x}_{obs}$. De ukendte indgange i $\mathbf{x}$ samles ligeledes i en vektor $\mathbf{x}_{mis}$. Uden tab af generalitet antager vi, at både $x_{1\,i}$ og $x_{2\,i}$ er observeret for de første n_1 observationer. Derudover antager vi, at for $i = n_1+1, \ldots, n_1+n_2$ er $x_{1\,i}$ observeret og $x_{2\,i}$ mangler. Det modsattes antages at gælde for $i = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3$. Yderligere antages $x_{1\,i}$ og $x_{2\,i}$ begge at mangle for $i = n_1 + n_2 + n_3 + 1, \ldots, n$, hvor $n_4 = n - n_1 - n_2 - n_3$. Da vi ikke kender det fulde data $\mathbf{x} = (\mathbf{x}_{obs}^T, \mathbf{x}_{mis}^T)^T$, kan vi ikke anvende maksimum likelihood estimation på det fulde data. Et alternativ kunne dog være at anvende maksimum likelihood estimation på det observerede data

$$\mathbf{x}_{obs} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_{n_1}^T, x_{1\,n_1+1}, \ldots, x_{1\,n_1+n_2}, x_{2\,n_1+n_2+1}, \ldots, x_{2\,n_1+n_2+n_3})^T.$$

Log-likelihoodfunktionen for $\mathbf{x}_{obs}$ er givet ved:

$$\begin{aligned} \ell_{obs}(\boldsymbol{\theta}; \mathbf{x}_{obs}) &= \sum_{i=1}^{n_1} \ell(\boldsymbol{\theta}; \mathbf{x}_i) + \sum_{i=n_1+1}^{n_1+n_2} \ell(\boldsymbol{\theta}; x_{1\,i}) + \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \ell(\boldsymbol{\theta}; x_{2\,i}) \\ &= -\frac{n_1}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n_1} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &\quad - \frac{n_2}{2} \log(\sigma_{11}) - \frac{1}{2\sigma_{11}} \sum_{i=n_1+1}^{n_1+n_2} (x_{1\,i} - \mu_1)^2 \\ &\quad - \frac{n_3}{2} \log(\sigma_{22}) - \frac{1}{2\sigma_{22}} \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} (x_{2\,i} - \mu_2)^2. \end{aligned}$$

Ud fra $\ell_{obs}(\boldsymbol{\theta}; \mathbf{x}_{obs})$ kan vi estimere $\boldsymbol{\theta}$ ved at anvende maksimum likelihood estimaton. En anden metode til at estimere $\boldsymbol{\theta}$ er EM-algoritmen, hvilken vi opstiller i det følgende. Det følger af (2.27), at der er tale om en ekponentiel familie, hvorfor vi kan anvende Sætning 2.3.2 til at opstille EM-algoritmen. E-trinet består i at bestemme $s^{(t)} = \mathrm{E}\left[s(\mathbf{X}) \,\middle|\, \mathbf{x}_{obs}; \boldsymbol{\theta}^{(t)} \right]$, hvor $\boldsymbol{\theta}^{(t)} = \left(\mu_1^{(t)}, \mu_2^{(t)}, \sigma_{11}^{(t)}, \sigma_{22}^{(t)}, \sigma_{12}^{(t)} \right)^T$. For at kunne bestemme $s^{(t)}$ er det nødvendigt at evaluere følgende betingede middelværdier, hvilket Lemma A.0.1 anvendes til:

- Bestem følgende for $i = n_1 + 1, \ldots, n_1 + n_2$:

$$\begin{aligned} \mathrm{E}\left[X_{2\,i} \,\middle|\, x_{1\,i}; \boldsymbol{\theta}^{(t)} \right] &= \mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}} (x_{1\,i} - \mu_1^{(t)}), \\ \mathrm{E}\left[X_{2\,i}^2 \,\middle|\, x_{1\,i}; \boldsymbol{\theta}^{(t)} \right] &= \mathrm{Var}\left[X_{2\,i} \,\middle|\, x_{1\,i}; \boldsymbol{\theta}^{(t)} \right] + \mathrm{E}\left[X_{2\,i} \,\middle|\, x_{1\,i}; \boldsymbol{\theta}^{(t)} \right]^2 \\ &= \sigma_{22}^{(t)} - \frac{\sigma_{12}^{(t)^2}}{\sigma_{11}^{(t)}} + \mathrm{E}\left[X_{2\,i} \,\middle|\, x_{1\,i}; \boldsymbol{\theta}^{(t)} \right]^2. \end{aligned}$$

- Bestem følgende for $i = n_1 + n_2 + 1, \ldots, n_1 + n_2 + n_3$:

$$\begin{aligned}
\mathrm{E}\left[X_{1\,i} \,\middle|\, x_{2\,i}; \boldsymbol{\theta}^{(t)}\right] &= \mu_1^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{22}^{(t)}}(x_{2\,i} - \mu_2^{(t)}), \\
\mathrm{E}\left[X_{1\,i}^2 \,\middle|\, x_{2\,i}; \boldsymbol{\theta}^{(t)}\right] &= \mathrm{Var}\left[X_{1\,i} \,\middle|\, x_{2\,i}; \boldsymbol{\theta}^{(t)}\right] + \mathrm{E}\left[X_{1\,i} \,\middle|\, x_{2\,i}; \boldsymbol{\theta}^{(t)}\right]^2 \\
&= \sigma_{11}^{(t)} - \frac{\sigma_{12}^{(t)^2}}{\sigma_{22}^{(t)}} + \mathrm{E}\left[X_{1\,i} \,\middle|\, x_{2\,i}; \boldsymbol{\theta}^{(t)}\right]^2.
\end{aligned}$$

- Bestem følgende for $i = n_1 + n_2 + n_3 + 1, \ldots, n$ og $j = 1, 2$:

$$\begin{aligned}
\mathrm{E}\left[X_{j\,i}; \boldsymbol{\theta}^{(t)}\right] &= \mu_j^{(t)}, \\
\mathrm{E}\left[X_{j\,i}^2; \boldsymbol{\theta}^{(t)}\right] &= \mathrm{Var}\left[X_{j\,i}; \boldsymbol{\theta}^{(t)}\right] + \mathrm{E}\left[X_{j\,i}; \boldsymbol{\theta}^{(t)}\right]^2 = \sigma_{jj}^{(t)} + \mu_j^{(t)^2}, \\
\mathrm{E}\left[X_{1\,i}X_{2\,i}; \boldsymbol{\theta}^{(t)}\right] &= \mathrm{Cov}\left[X_{1\,i}, X_{2\,i}; \boldsymbol{\theta}^{(t)}\right] + \mathrm{E}\left[X_{1\,i}; \boldsymbol{\theta}^{(t)}\right] \mathrm{E}\left[X_{2\,i}; \boldsymbol{\theta}^{(t)}\right] = \sigma_{12}^{(t)} + \mu_1^{(t)}\mu_2^{(t)}.
\end{aligned}$$

Ud fra de tre ovenstående punkter kan vi bestemme $s^{(t)}$. For at kunne opstille M-trinet er det nødvendigt at bestemme middelværdien af $s(\mathbf{X})$:

$$\mathrm{E}[s(\mathbf{X}); \boldsymbol{\theta}] = \begin{bmatrix} \sum_{i=1}^n \mathrm{E}[X_{1\,i}; \boldsymbol{\theta}] \\ \sum_{i=1}^n \mathrm{E}[X_{2\,i}; \boldsymbol{\theta}] \\ \sum_{i=1}^n \mathrm{E}[X_{1\,i}^2; \boldsymbol{\theta}] \\ \sum_{i=1}^n \mathrm{E}[X_{2\,i}^2; \boldsymbol{\theta}] \\ \sum_{i=1}^n \mathrm{E}[X_{1\,i}X_{2\,i}; \boldsymbol{\theta}] \end{bmatrix} = \begin{bmatrix} n\mu_1 \\ n\mu_2 \\ n(\sigma_{11} + \mu_1^2) \\ n(\sigma_{22} + \mu_2^2) \\ n(\sigma_{12} + \mu_1\mu_2) \end{bmatrix} = n \left(\boldsymbol{\theta} + \begin{bmatrix} 0 \\ 0 \\ \mu_1^2 \\ \mu_2^2 \\ \mu_1\mu_2 \end{bmatrix} \right), \tag{2.29}$$

hvor vi har anvendt (2.28) samt $\mathrm{E}[YZ] = \mathrm{Cov}[Y, Z] + \mathrm{E}[Y]\mathrm{E}[Z]$. Vi kan ved hjælp af Sætning 2.3.2 opstille M-trinet. Dette gøres ved at løse følgende ligning med hensyn til $\boldsymbol{\theta}$:

$$\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^T \right\} \mathrm{E}[s(\mathbf{X}); \boldsymbol{\theta}] = \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^T \right\} s^{(t)}, \tag{2.30}$$

hvor matricen $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^T$ er givet som i (B.1). Fra Appendiks B har vi, at $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^T$ er invertibel. Dermed er det nok at løse følgende ligning med hensyn til $\boldsymbol{\theta}$ fremfor 2.30:

$$\mathrm{E}[s(\mathbf{X}); \boldsymbol{\theta}] = s^{(t)}. \tag{2.31}$$

Ved at isolere $\boldsymbol{\theta}$ i (2.31) fås:

$$\boldsymbol{\theta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} s_1^{(t)} \\ \frac{1}{n} s_2^{(t)} \\ \frac{1}{n} s_3^{(t)} - \mu_1^2 \\ \frac{1}{n} s_4^{(t)} - \mu_2^2 \\ \frac{1}{n} s_5^{(t)} - \mu_1\mu_2 \end{bmatrix}, \tag{2.32}$$

hvor vi har anvendt (2.29) samt, at $s_i^{(t)}$ er indgang i i $s^{(t)}$.

Ud fra (2.32) kan vi opstille M-trinet:

- Bestem $\mu_1^{(t+1)} = \frac{s_1^{(t)}}{n}$.
- Bestem $\mu_2^{(t+1)} = \frac{s_2^{(t)}}{n}$.
- Bestem $\sigma_{11}^{(t+1)} = \frac{s_3^{(t)}}{n} - \mu_1^{(t+1)^2}$.
- Bestem $\sigma_{22}^{(t+1)} = \frac{s_4^{(t)}}{n} - \mu_2^{(t+1)^2}$.
- Bestem $\sigma_{12}^{(t+1)} = \frac{s_5^{(t)}}{n} - \mu_1^{(t+1)}\mu_2^{(t+1)}$.

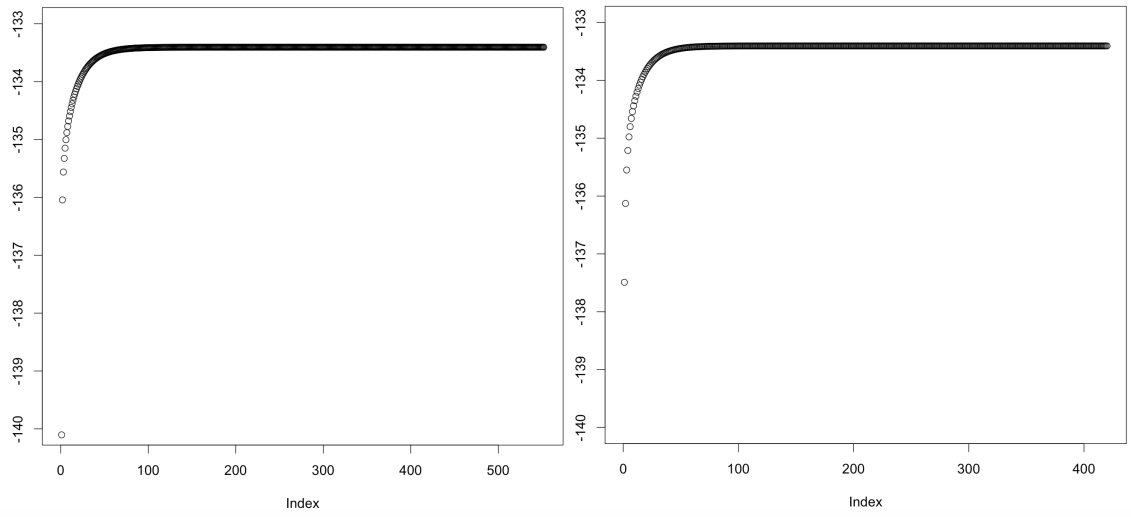
I Appendiks C fremgår `R`-koden for den EM-algoritme, som er blevet opstillet i dette afsnit.

2.4.1 Databehandling

I dette afsnit anvender vi `R`-koden i Appendiks C på datasættet `misdat.RData`. Dette datasæt består af 200 realiseringer $\mathbf{x}_i = (x_{1\,i}, x_{2\,i})^T$, som antages at følge en bivariat normalfordeling. Heraf er $x_{1\,i}$ og $x_{2\,i}$ begge givet for $i = 1, \ldots, 50$. Derudover er $x_{1\,i}$ kendt og $x_{2\,i}$ ukendt for $i = 51, \ldots, 100$, og det modsatte er gældende for de efterfølgende 50 realiseringer. Ydermere er $x_{1\,i}$ og $x_{2\,i}$ begge ukendte for de resterende 50 realiseringer. For datasættet `misdat.RData` giver `R`-koden i Appendiks C følgende parameterestimationer for $\boldsymbol{\mu}$ og Σ:

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \end{bmatrix} = \begin{bmatrix} 1,077 \\ -0,658 \end{bmatrix}, \quad \widehat{\Sigma} = \begin{bmatrix} \widehat{\sigma}_{11} & \widehat{\sigma}_{12} \\ \widehat{\sigma}_{21} & \widehat{\sigma}_{22} \end{bmatrix} = \begin{bmatrix} 3,161 & 2,193 \\ 2,193 & 1,74 \end{bmatrix}, \tag{2.33}$$

hvor den har anvendt 552 iterationer med en tolerance på $1e-10$. Hvis de sidste 50 realiseringer i `misdat.RData` udelades giver `R`-koden i Appendiks C de samme parameterestimationer som før, dog er antallet af iterationer kun 420. EM-algoritmen opnår altså samme parameterestimationer uanset, om der i datasættet inddrages realiseringer, hvor hverken $x_{1\,i}$ og $x_{2\,i}$ er kendte. Det viser sig dog, at hvis der inddrages realiseringer, hvor både $x_{1\,i}$ og $x_{2\,i}$ er ukendte, øges antallet af iterationer. Dette kan skyldes, at der i E-trinet bliver lagt en større vægt på de nuværende parameterestioner $\boldsymbol{\mu}^{(t)}$ og $\Sigma^{(t)}$, hvilket påvirker de efterfølgende estimationer i M-trinet. Denne påvirkning ses også i de to følgende figurer, hvor log-likelihoodfunktionerne er evalueret for hver iteration:



Figur 2.1:
Log-likelihoodfunktionerne for hver iteration.

Figuren til venstre i Figur 2.1 er log-likelihoodfunktionen for `misdat.RData` evalueret for hver iteration i EM-algoritmen. Tilsvarende er figuren til højre i Figur 2.1 de evaluerede værdier af log-likelihoodfunktionen for `misdat.RData` foruden de sidste 50 realiseringer. Bemærk, at den første evaluerede log-likelihoodfunktion ikke indgår i Figur 2.1. Dette skyldes, at den i begge tilfælde er -617,7743, hvilket blandt andet vil gøre plottene i Figur 2.1 svære at sammenligne.

3. Faktoranalyse

I dette kapitel præsenteres den grundlæggende teori bag faktoranalyse først. Dernæst opstilles EM-algoritmen for en faktoranalyse (FA) model med henblik på at bestemme parametrene i en sådan model. Kapitlet er baseret på [Seber, 2004] og [Petersen og Pedersen, 2012].

3.1 Grundlæggende teori

I dette afsnit betragtes en stokastisk vektor $\mathbf{X} = (X_1, \ldots, X_n)^T$ med udfaldsrum $\mathcal{X} \subseteq \mathbb{R}^n$. Middelværdien af $\mathbf{X}$ noteres $\boldsymbol{\mu}$, og kovariansmatricen for $\mathbf{X}$ noteres Σ.

Definition 3.1.1
Lad $\mathbf{f} \in \mathbb{R}^k$ være en stokastisk variable og Λ være en $n \times k$ matrix. En FA model for $\mathbf{X}$ er givet ved:

$$\mathbf{X} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\varepsilon}, \tag{3.1}$$

hvor $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \Psi)$ er et fejlled og $\Psi = \operatorname{diag}(\psi_1^2, \ldots, \psi_n^2)$. Indgangene i $\mathbf{f}$ kaldes for *faktorer*, og indgangene i Λ kaldes for faktorvægtene.

I det efterfølgende antages det, at $\mathbf{f}$ er normalfordelt med middelværdi $\mathbf{0}$, hvilket er i overensstemmelse med, at $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}$. Derudover antages $\mathbf{f}$ og $\boldsymbol{\varepsilon}$ at være ukorrelerede. Lad $\Sigma_{\mathbf{f}} = \mathrm{Var}[\mathbf{f}]$, da kan (3.1) omskrives til:

$$\mathbf{X} = \boldsymbol{\mu} + \left(\Lambda\Sigma_{\mathbf{f}}^{1/2}\right)\left(\Sigma_{\mathbf{f}}^{-1/2}\mathbf{f}\right) + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \widetilde{\Lambda}\widetilde{\mathbf{f}} + \boldsymbol{\varepsilon},$$

hvilket også er en FA model. Dette medfører, at

$$\mathrm{Var}\left[\widetilde{\mathbf{f}}\right] = \mathrm{Var}\left[\Sigma_{\mathbf{f}}^{-1/2}\mathbf{f}\right] = \Sigma_{\mathbf{f}}^{-1/2}\Sigma_{\mathbf{f}}\Sigma_{\mathbf{f}}^{-1/2} = I_k.$$

Det kan derfor fremover antages, at $\mathbf{f} \sim N_k(\mathbf{0}, I_k)$ uden tab af generalitet. Dermed har vi:

$$\Sigma = \mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\Lambda\mathbf{f}] + \mathrm{Var}[\boldsymbol{\varepsilon}] = \Lambda\mathrm{Var}[\mathbf{f}]\Lambda^T + \Psi = \Lambda\Lambda^T + \Psi, \tag{3.2}$$

hvor vi har anvendt, at $\mathbf{f}$ og $\boldsymbol{\varepsilon}$ er ukorrelerede. Altså er $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Lambda\Lambda^T + \Psi)$. Hver diagonalindgang i Ψ beskriver en såkaldt *entydighed* af den tilhørende variabel i $\mathbf{X}$. Årsagen til, at det kaldes entydigheden, er, at det er den del af $\mathrm{Var}[\mathbf{X}]$, som faktorvægtene i Λ ikke kan beskrive.

Lemma 3.1.2
For en FA model gælder følgende:

1) $\mathbf{X} \mid \mathbf{f} \sim N_n(\boldsymbol{\mu} + \Lambda\mathbf{f}, \Psi)$.

2) $\mathbf{f} \mid \mathbf{x} \sim N_k(\Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(\mathbf{x} - \boldsymbol{\mu}), (I_k + \Lambda^T\Psi^{-1}\Lambda)^{-1})$.

Bevis
Vi har, at

$$\begin{aligned}\mathrm{Cov}(\mathbf{X},\mathbf{f}) &= \mathrm{E}[(\mathbf{X}-\mathrm{E}[\mathbf{X}])(\mathbf{f}-\mathrm{E}[\mathbf{f}])^T] = \mathrm{E}[(\mathbf{X}-\boldsymbol{\mu})\mathbf{f}^T] \\ &= \mathrm{E}[(\Lambda\mathbf{f}+\boldsymbol{\varepsilon})\mathbf{f}^T] = \Lambda\mathrm{E}[\mathbf{f}\mathbf{f}^T] + \mathrm{E}[\boldsymbol{\varepsilon}\mathbf{f}^T] = \Lambda\mathrm{Var}[\mathbf{f}] = \Lambda,\end{aligned}$$

hvor vi har anvendt, at $\mathbf{f}$ og $\boldsymbol{\varepsilon}$ er ukorrelerede. Vi kan dermed opstille

$$\begin{pmatrix}\mathbf{f}\\ \mathbf{X}\end{pmatrix} \sim N_{k+n}\left(\begin{pmatrix}\mathbf{0}\\ \boldsymbol{\mu}\end{pmatrix}, \begin{bmatrix} I_k & \Lambda^T \\ \Lambda & \Lambda\Lambda^T+\Psi\end{bmatrix}\right).$$

Ved at anvende Lemma A.0.1 får vi, at

$$E[\mathbf{X}\mid\mathbf{f}] = \boldsymbol{\mu}+\Lambda\mathbf{f} \quad \text{og} \quad \mathrm{Var}[\mathbf{X}\mid\mathbf{f}] = \Lambda\Lambda^T+\Psi-\Lambda I_k\Lambda^T = \Psi.$$

Yderligere har vi, at $\mathrm{E}[\mathbf{f}\mid\mathbf{x}] = \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}(\mathbf{x}-\boldsymbol{\mu})$. Derudover giver Lemma A.0.1 følgende:

$$\mathrm{Var}[\mathbf{f}\mid\mathbf{x}] = I_k - \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}\Lambda = (I_k+\Lambda^T\Psi^{-1}\Lambda)^{-1}, \tag{3.3}$$

hvor vi har anvendt Woodbury's identitet:

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}. \qquad \blacksquare$$

Lemma 3.1.2 giver, at variablene i $\mathbf{X}$ er ukorrelerede givet, at faktorerne i $\mathbf{f}$ er kendte. Bemærk, at der for en ortogonal $k\times k$ matrix R gælder:

$$\Lambda\mathbf{f} = (\Lambda R^T)(R\mathbf{f}) = \widetilde{\Lambda}\widetilde{\mathbf{f}}, \tag{3.4}$$

hvor $R^TR = I_k$. Betragtes $\Lambda\mathbf{f} = \widetilde{\Lambda}\widetilde{\mathbf{f}}$ i (3.1) er fordelingen af $\mathbf{X}$ uændret, da

$$\mathrm{E}[\mathbf{X}] = \mathrm{E}[\boldsymbol{\mu}+\widetilde{\Lambda}\widetilde{\mathbf{f}}+\boldsymbol{\varepsilon}] = \boldsymbol{\mu}+\widetilde{\Lambda}R\mathrm{E}[\mathbf{f}] = \boldsymbol{\mu}$$

samt

$$\mathrm{Var}[\mathbf{X}] = \mathrm{Var}[\boldsymbol{\mu}+\widetilde{\Lambda}\widetilde{\mathbf{f}}+\boldsymbol{\varepsilon}] = \widetilde{\Lambda}R\mathrm{Var}[\mathbf{f}]R^T\widetilde{\Lambda}^T+\Psi = \widetilde{\Lambda}\widetilde{\Lambda}^T+\Psi = \Lambda\Lambda^T+\Psi,$$

hvor vi har anvendt (3.4). Dette medfører, at Λ og $\mathbf{f}$ i (3.1) ikke er entydige. Vi kan dermed altid rotere $\mathbf{f}$ og Λ. Dermed kan Λ og $\mathbf{f}$ blandt andet vælges sådan, at $\Lambda^T\Psi^{-1}\Lambda$ er en diagonalmatrix. Jævnfør Lemma 3.1.2 medfører dette, at faktorerne i $\mathbf{f}$ er ukorrelerede givet $\mathbf{X}$. I det kommende afsnit vil vi se på parameterestimation i en FA model.

3.2 Parameterestimation

I dette afsnit vil opstille en metode til at estimere $\boldsymbol{\theta} = (\boldsymbol{\mu},\Lambda,\Psi)$ i en FA model som givet i Definition 3.1.1. Fra Afsnit 3.1 har vi, at $\mathbf{X}\sim N_n(\boldsymbol{\mu},\Lambda\Lambda^T+\Psi)$, hvormed vi kan opskrive tæthedsfunktionen for $\mathbf{X}$:

$$f(\mathbf{x};\boldsymbol{\theta}) = (2\pi)^{-n/2}|\Lambda\Lambda^T+\Psi|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\Lambda\Lambda^T+\Psi)^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{3.5}$$

MLE kan anvendes til at estimere $\boldsymbol{\mu}$, hvorved $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Uden tab af generalitet kan vi fremover betragte $\mathbf{Y} = \mathbf{X}-\boldsymbol{\mu}\sim N_n(\mathbf{0},\Lambda\Lambda^T+\Psi)$ i stedet for $\mathbf{X}$. Bemærk, at $\mathbf{Y}\mid\mathbf{f}\sim N_n(\Lambda\mathbf{f},\Psi)$ jævnfør Lemma 3.1.2.

Vi kan anvende EM-algoritmen til at estimere Λ og Ψ. Dette gøres ved at betragte m realiseringer af $\mathbf{Y}$, hvilke noteres $\mathbf{y}_i$ med tilhørende $\mathbf{f}_i$ for $i = 1, \ldots, m$. Det observerede data består af $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^m$, og det manglende data består af $\mathbf{f} = \{\mathbf{f}_i\}_{i=1}^m$. For at kunne anvende EM-algoritmen er det nødvendigt at kende den fulde log-likelihoodfunktion for $\mathbf{y}, \mathbf{f}$:

$$\begin{aligned}
\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{f}) &= \sum_{i=1}^m \log(f(\mathbf{y}_i, \mathbf{f}_i; \boldsymbol{\theta})) = \sum_{i=1}^m \log(f(\mathbf{f}_i; \boldsymbol{\theta})) + \sum_{i=1}^m \log(f(\mathbf{y}_i \mid \mathbf{f}_i; \boldsymbol{\theta})) \\
&\propto \sum_{i=1}^m -\frac{1}{2}\mathbf{f}_i^T\mathbf{f}_i - \frac{m}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^m (\mathbf{y}_i - \Lambda\mathbf{f}_i)^T\Psi^{-1}(\mathbf{y}_i - \Lambda\mathbf{f}_i) \\
&\propto -\frac{m}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^m (\mathbf{y}_i - \Lambda\mathbf{f}_i)^T\Psi^{-1}(\mathbf{y}_i - \Lambda\mathbf{f}_i) \\
&= -\frac{m}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^m \operatorname{tr}\{(\mathbf{y}_i - \Lambda\mathbf{f}_i)^T\Psi^{-1}(\mathbf{y}_i - \Lambda\mathbf{f}_i)\} \\
&= -\frac{m}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^m \operatorname{tr}\{(\mathbf{y}_i - \Lambda\mathbf{f}_i)(\mathbf{y}_i - \Lambda\mathbf{f}_i)^T\Psi^{-1}\},
\end{aligned}$$

hvor vi har anvendt, at $\mathbf{f}_i \sim N_k(\mathbf{0}, I_k)$, $\mathbf{y}_i \mid \mathbf{f}_i \sim N_n(\Lambda\mathbf{f}_i, \Psi)$ og $\boldsymbol{\theta} = (\Lambda, \Psi)$. Hvis vi lader $S = \frac{1}{m}\sum_{i=1}^m \mathbf{y}_i\mathbf{y}_i^T + \Lambda\mathbf{f}_i\mathbf{f}_i^T\Lambda^T - 2\Lambda\mathbf{f}_i\mathbf{y}_i^T$, hvilken også er kendt som den empiriske kovarians, da kan den fulde log-likelihoodfunktion omskrives til:

$$\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{f}) \propto -\frac{m}{2}\log|\Psi| - \frac{m}{2}\operatorname{tr}\{S\Psi^{-1}\}.$$

Ud fra denne kan vi opstille $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$:

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) &= \mathrm{E}[\ell(\boldsymbol{\theta}; \mathbf{y}, \mathbf{f}) \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}] \\
&= \mathrm{E}\left[-\frac{m}{2}\log|\Psi| - \frac{m}{2}\operatorname{tr}\{S\Psi^{-1}\} \,\middle|\, \mathbf{y}; \boldsymbol{\theta}^{(t)}\right] \\
&= -\frac{m}{2}\log|\Psi| - \frac{m}{2}\operatorname{tr}\{\mathrm{E}[S \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}]\Psi^{-1}\},
\end{aligned}$$

hvor vi har anvendt, at trace er en lineær operator. Hermed kan vi opstille E-trinet:

E: Bestem $\mathrm{E}[S \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}] = \frac{1}{m}\sum_{i=1}^m \mathbf{y}_i\mathbf{y}_i^T + \Lambda\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\Lambda^T - 2\Lambda\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T$.

Lemma 3.1.2 kan anvendes til at bestemme ovenstående betingede middelværdier, hvor det kan bemærkes, at $\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}] = \mathrm{Var}[\mathbf{f}_i \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}] + \mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\mathrm{E}[\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]$. Til at bestemme M-trinet betragtes først:

$$\begin{aligned}
\frac{\partial}{\partial\Lambda}Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) &= \frac{\partial}{\partial\Lambda}\left\{-\frac{m}{2}\log|\Psi| - \frac{m}{2}\operatorname{tr}\{\mathrm{E}[S \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}]\Psi^{-1}\}\right\} \\
&= -\frac{1}{2}\sum_{i=1}^m \left\{\frac{\partial}{\partial\Lambda}\operatorname{tr}\{\Lambda\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\Lambda^T\Psi^{-1}\} - 2\frac{\partial}{\partial\Lambda}\operatorname{tr}\{\Lambda\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T\Psi^{-1}\}\right\} \\
&= -\frac{1}{2}\sum_{i=1}^m \left\{\Psi^{-1}\Lambda\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}] + \Psi^{-1}\Lambda\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}] - 2\Psi^{-1}\mathbf{y}_i\mathrm{E}[\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\right\} \\
&= -\Psi^{-1}\Lambda\sum_{i=1}^m \left\{\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\right\} + \Psi^{-1}\sum_{i=1}^m \left\{\mathbf{y}_i\mathrm{E}[\mathbf{f}_i^T \mid \mathbf{y}_i; \boldsymbol{\theta}^{(t)}]\right\},
\end{aligned}$$

hvor vi har anvendt, at $\frac{\partial}{\partial X}\{AXBX^TC\} = A^TC^TXB + CAXB$ og $\frac{\partial}{\partial X}\operatorname{tr}\{XA\} = A^T$. Ved at sætte $\frac{\partial}{\partial \Lambda}Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \mathbf{0}$ fås:

$$\Lambda^{(t+1)} = \left(\sum_{i=1}^{m}\mathbf{y}_i\mathrm{E}[\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\right)\left(\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\right)^{-1}.$$

Hernæst betragtes:

$$\begin{aligned}\frac{\partial}{\partial \Psi^{-1}}Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) &= \frac{\partial}{\partial \Psi^{-1}}\left\{-\frac{m}{2}\log|\Psi| - \frac{m}{2}\operatorname{tr}\{E[S \mid \mathbf{y};\boldsymbol{\theta}^{(t)}]\Psi^{-1}\}\right\}\\ &= \frac{m}{2}\Psi - \frac{m}{2}\mathrm{E}[S \mid \mathbf{y};\boldsymbol{\theta}^{(t)}],\end{aligned}$$

hvor vi har anvendt, at $\frac{\partial}{\partial \Psi^{-1}}\log|\Psi| = -\Psi$ og $\frac{\partial}{\partial X}\operatorname{tr}\{AX\} = A^T$ samt $S = S^T$. Ved at sætte $\frac{\partial}{\partial \Psi^{-1}}Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \mathbf{0}$ samt indsætte Λ^{t+1} fås:

$$\begin{aligned}\Psi^{(t+1)} &= \operatorname{diag}\{\mathrm{E}[S \mid \mathbf{y};\boldsymbol{\theta}^{(t)}]\}\\ &= \frac{1}{m}\operatorname{diag}\left\{\sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i^T + \Lambda^{(t+1)}\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\Lambda^{(t+1)^T} - 2\Lambda^{(t+1)}\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T\right\}\\ &= \frac{1}{m}\operatorname{diag}\left\{\sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i^T - \Lambda^{(t+1)}\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T\right\},\end{aligned}$$

hvor vi har anvendt, at $\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\Lambda^{(t+1)^T} = \sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T$. Dermed bliver M-trinet følgende:

M: Bestem $\Lambda^{(t+1)}$ og $\Psi^{(t+1)}$ ved

$$\begin{aligned}\Lambda^{(t+1)} &= \left(\sum_{i=1}^{m}\mathbf{y}_i\mathrm{E}[\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\right)\left(\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i\mathbf{f}_i^T \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\right)^{-1},\\ \Psi^{(t+1)} &= \frac{1}{m}\operatorname{diag}\left\{\sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i^T - \Lambda^{(t+1)}\sum_{i=1}^{m}\mathrm{E}[\mathbf{f}_i \mid \mathbf{y}_i;\boldsymbol{\theta}^{(t)}]\mathbf{y}_i^T\right\}.\end{aligned}$$

Bemærk, at hvis $\mathbf{f}_i$ betragtes som en fast værdi fremfor en stokastisk variabel, da er $\mathbf{y}_i \sim N_n(\Lambda\mathbf{f}_i, \Psi)$ jævnfør Lemma 3.1.2, og $\mathbf{f}_i$ kan estimeres ved hjælp af "weighted least squares":

$$\widehat{\mathbf{f}}_i = (\Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}\mathbf{y}_i.$$

Hvis parametrene erstattes med deres estimater fås:

$$\widehat{\mathbf{f}}_i = (\widehat{\Lambda}^T\widehat{\Psi}^{-1}\widehat{\Lambda})^{-1}\widehat{\Lambda}^T\widehat{\Psi}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}),$$

hvor $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$.

Der findes en metode til at ordne faktorerne i $\mathbf{f}_i$ efter den største variation i data $\mathbf{x}$. Denne metode kaldes for *Varimax rotationskriteriet* (Varimax) og er givet ved:

$$R_{Vmax} = \arg\max_R \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}(\Lambda R)_{ij}^4 - \sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}(\Lambda R)_{ij}^2\right)^2,$$

hvor R er en ortogonal matrix. Bemærk, da vi tidligere har antaget, at $\mathbf{f}_i \sim N_k(\mathbf{0}, I_k)$, er dette også gældende for $R_{Vmax}\mathbf{f}_i \sim N_k(\mathbf{0}, I_k)$. Dermed er faktorerne stadig ukorrelerede under anvendelse af Varimax.

I det følgende afsnit anvendes ovenstående teori vedrørende EM-algoritmen for en FA model på et givet datasæt.

3.3 Databehandling

I dette afsnit ønsker vi at implementere og anvende en R-kode, som anvender metoden beskrevet i Afsnit 3.2 til at estimere parametrene $\boldsymbol{\mu}$, Λ og Ψ i en FA model. Vi betragter datasættet `math.csv`, hvilket indeholder 88 observationer. En observation $\mathbf{x}_i$ består af en elevs score for hver af de fem variable *me*, *ve*, *al*, *an* og *st*, hvilke er forskellige matematiske fag. En score ligger mellem 0 og 100. Vi vil i det følgende opstille to FA modeller med henholdsvis én og to faktorer, i hvilke vi ønsker at estimere parametrene $\boldsymbol{\mu}$, Λ og Ψ. Middelværdien $\boldsymbol{\mu}$ estimeres ved $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Herefter anvendes $\widehat{\boldsymbol{\mu}}$ til at centrere data. Derudover standardiseres data for at få sammenlignelige resultater på tværs af variablene i forhold til Λ og Ψ. Denne skalering af data udføres ved brug af R-funktionen `scale`. R-koden til alt dette fremgår af Appendiks D.

Begyndelsesbetingelserne for Λ vælges ud fra egenværdierne af den empiriske kovariansmatrix S for $\mathbf{y}$, hvor $\mathbf{y}$ er det skalerede data. Hvis egenværdierne λ_i er sorteret således, at $\lambda_i \geq \lambda_{i+1}$, da vælges søjlerne i $\Lambda^{(0)}$ til at være $\sqrt{\lambda_j}e_j$ for $j = 1, \ldots, k$, hvor k er antallet af faktorer i FA modellen, og e_j er den tilhørende egenvektor til λ_j. Ved at vælge disse begyndelsesbetingelser for Λ sikres det, at faktorerne er ordnet efter den største variation i data jævnfør [Ruppert, 2011, s. 162]. Ud fra $\Lambda^{(0)}$ kan begyndelsesbetingelserne for Ψ vælges til at være $\Psi^{(0)} = \text{diag}(S - \Lambda^{(0)}\Lambda^{(0)^T})$.

Anvendes R-koden fra Appendiks D på `math.csv` til at estimere parametrene i en FA model med én faktor fås følgende estimater:

Variabel	Faktorvægt	Entydighed
me	-0,5955	0,634
ve	-0,6635	0,5483
al	-0,9121	0,1567
an	-0,768	0,3988
st	-0,7195	0,4709

Tabel 3.1: Estimaterne for en FA model med én faktor.

Ud fra Tabel 3.1 kan det bemærkes, at faktorvægten for *al* er relativt høj samtidig med, at entydigheden er markant lavere sammenlignet med entydigheden for de andre variable. Dette kan indikere, at faktoren beskriver størstedelen af *al*.

Anvendes R-koden ligeledes til at estimere parametrene i en FA model med to faktorer, får vi følgende estimater:

Variabel	Faktorvægt 1	Faktorvægt 2	Entydighed
me	-0,6728	0,45540	0,3296
ve	-0,6926	0,18649	0,4746
al	-0,8894	-0,1181	0,1841
an	-0,7611	-0,2398	0,3521
st	-0,7104	-0,2221	0,4348

Tabel 3.2: Estimaterne for en FA model med to faktorer.

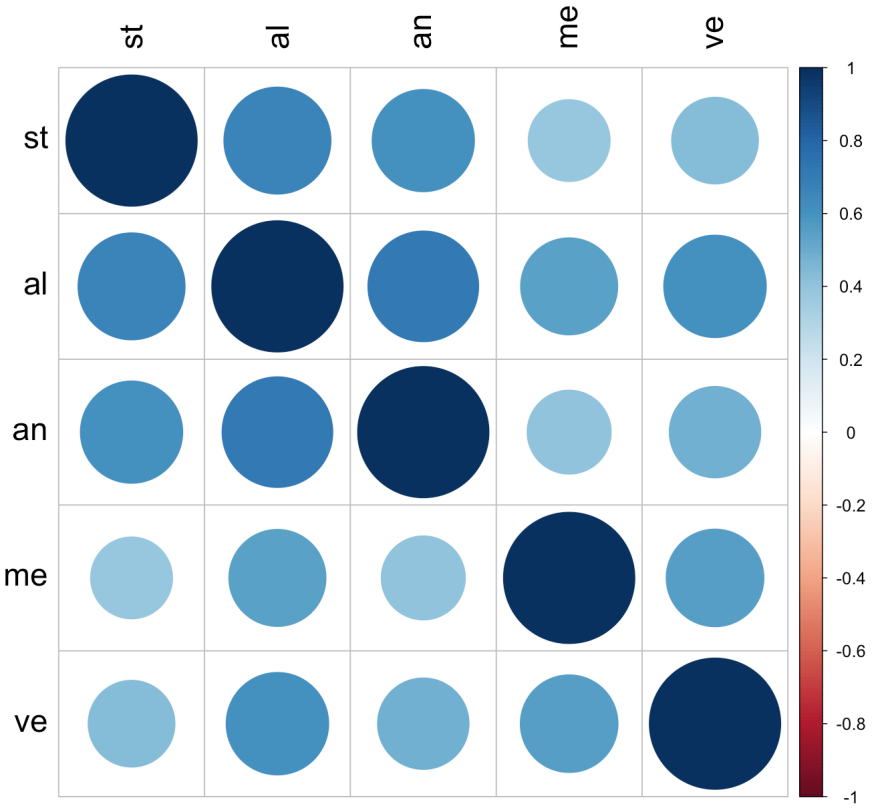
Af Tabel 3.2 fremgår det også, at *al* har en lav entydighed sammenlignet med entydigheden for de andre variable. Derudover kan det bemærkes, at entydigheden for de resterende variable er mindre i denne model sammenlignet med den forrige FA model med én faktor. Dermed beskriver denne FA model overordnet set datasættet `math.csv` bedre

end den forrige FA model. Benyttes Varimax på faktorvægtene fra Tabel 3.2 opnår vi følgende:

Variabel	Faktorvægt 1	Faktorvægt 2	Entydighed
me	-0,2526	0,7722	0,3296
ve	-0,4330	0,5718	0,4746
al	-0,7752	0,4517	0,1841
an	-0,7484	0,277	0,3521
st	-0,6975	0,2599	0,4348

Tabel 3.3: Varimax anvendt på faktorvægtene i FA modellen med to faktorer.

Tabel 3.3 indikerer, at den første faktor har en stor indvirkning på *al*, *an* og *st*, hvorimod den anden faktor har en stor indvirkning på *me* og *ve*.



Figur 3.1: Korrelationsplot.

Figur 3.1 antyder, at der er en høj korrelation mellem variablene *st*, *al* og *an*. Derudover antyder figuren, at der er en korrelation mellem *me* og *ve*. Alt dette peger i samme retning som konklusionen lavet ud fra faktorvægtene i Tabel 3.3.

4. Bayesianske netværk

I dette kapitel specificeres teorien vedrørende Bayesianske netværk på baggrund af den grundlæggende grafteori fra Appendiks E. Dette kapitel er baseret på [Lauritzen og Spiegelhalter, 1988].

Definition 4.0.1
Et Bayesiansk netværk er en DAG $G = (V, E)$, hvis knuder repræsenterer stokastiske variable X_v, hvor $v \in V$.

Bemærk, at en kant $w \to v$ i et Bayesiansk netværk indikerer en "direkte" afhængighed mellem X_w og X_v. Yderligere gælder der, at hvis $u \notin pa(v)$ og $w \in pa(v)$, da er X_v og X_u betinget uafhængige givet X_w. Dette noteres $X_v \perp\!\!\!\perp X_u \mid X_w$. Betinget uafhængighed er beskrevet i Appendiks F.

En stokastisk process $\mathbf{X}$ er et Bayesiansk netværk med hensyn til en orienteret graf $G = (V, E)$, hvis

$$P(V) = \prod_{v \in V} P(v \mid pa(v)), \tag{4.1}$$

hvor $v \in V$ er et indeks for en stokastisk variabel $X_v \in \mathbf{X}$. Dette betyder, at man kan nøjes med at betragte de betingede sandsynligheder $P(v \mid pa(v))$ fremfor den simultane $P(V)$, hvilken kan være relativt kompliceret at håndtere og beregne. Der findes andre måder at betragte $P(V)$ på, men før disse introduceres er det nødvendigt at præsentere en algoritme til at opstille et kliketræ ud fra en uorienteret graf.

Algoritme 4.0.2 (Kliketræ)
Lad $G = (V, E)$ være en uorienteret graf. Anvend Algoritme E.0.18 til at nummerere knuderne i grafen. Hvis nummereringen ikke er perfekt tilføjes der kanter i grafen sådan, at nummereringen bliver perfekt. Dette medfører, at grafen er trianguleret jævnfør Proposition E.0.11. Ud fra nummereringen dannes mængderne $K_j = A_j \cap \{j\}$ for $j = 1, \ldots, \#V$, hvor $A_j = nabo(j) \cap \{1, \ldots, j-1\}$. Herefter fjernes de mængder K_i, som opfylder, at $K_i \subset K_j$ for $i \neq j$. De resterende K_j'er ordnes efter deres maksimale nummereret knude, fra lavest til højest. Dette giver en ordnet mængde af kliker, $C_1, \ldots, C_p$. Til sidst dannes kliketræet Ψ_p med roden C_1 ved at forbinde kliken C_j, for $j = 2, \ldots, p$, med dens potentielle forælder C_i, hvor $i < j$ og i er størst mulig. □

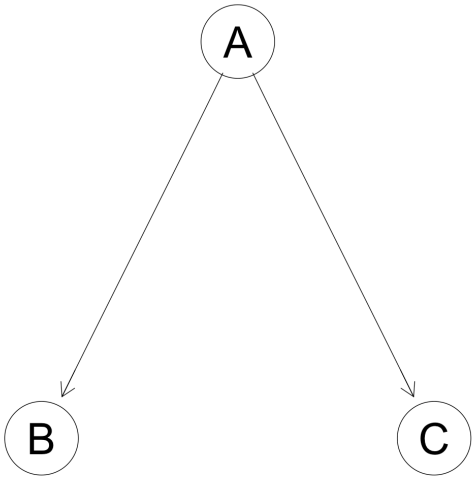
Bemærk, at et kliketræ ikke nødvendigvis er entydigt, da forskellige nummereringer kan lede til forskellige kliketræer.

Ud fra et kliketræ Ψ_p for $\mathbf{X}$ kan (4.1) omskrives til

$$P(V) = q_1(C_1) \cdot \ldots \cdot q_p(C_p), \tag{4.2}$$

hvor $C_i \in \Psi_p$. I det følgende gives der et eksempel på dette.

Eksempel 4.0.3
Lad $V = (A, B, C)$ samt $\mathbf{X} = (X_A, X_B, X_C)$ med tilhørende DAG:

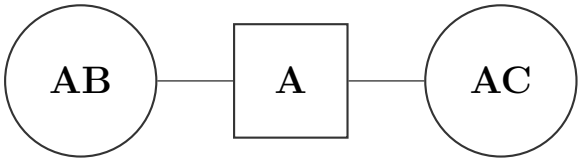


Figur 4.1: DAG, hvor A er forælder til både B og C.

Dette er et Bayesiansk netværk, da

$$P(V) = P(A)P(B \mid A)P(C \mid A). \tag{4.3}$$

Ved at moralisere DAG'en og dernæst anvende Algoritme 4.0.2 kan vi opstille følgende kliketræ Ψ_2, hvor $\{A, B\}$ er roden:



Figur 4.2: Kliketræ Ψ_2 for Figur 4.1.

Ud fra kliketræet Ψ_2 kan (4.3) omskrives til $P(V) = q_1(A, B)q_2(A, C)$. Hermed er $P(V)$ opskrevet ud fra funktioner af de enkelte kliker i Ψ_2. □

Vi har indtil nu beskrevet et Bayesiansk netværk ud fra de betingede sandsynligheder samt funktioner af klikerne i et tilhørende kliketræ. I det efterfølgende betragtes mængderne $S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1})$ og $R_i = C_i \backslash S_i$, hvor $i = 1, \ldots, p$ og C_i er en klike i et kliketræ. Bemærk, at $S_1 = \emptyset$ og $R_1 = C_1$. Givet et kliketræ Ψ_p, hvor $S_i \neq \emptyset$ for $i = 2, \ldots, p$, er det muligt at beskrive et Bayesiansk netværk ud fra *klikemarginalerne* $P(C_i)$ samt sandsynlighederne $P(S_i)$. Dette følger af, at der findes en metode til at omskrive (4.1) til:

$$P(V) = P(C_1) \prod_{i=2}^{p} \frac{P(C_i)}{P(S_i)}. \tag{4.4}$$

For at kunne opstille denne metode introduceres først to algoritmer.

Algoritme 4.0.4 (Collect evidence)
Lad Ψ_p være et kliketræ bestående af klikerne $C_1, \ldots, C_p$, hvor $p \geq 2$. Omskriv den simultane sandsynlighed for V fra (4.1) til:

$$P(V) = q_1(C_1) \cdot \ldots \cdot q_p(C_p). \tag{4.5}$$

Det efterfølgende trin anvendes fra $j = p$ til og med $j = 2$. Trin j:

Først opskrives den betingede sandsynlighed for R_j givet S_j:

$$P(R_j \mid S_j) = P(R_j \mid C_1, \ldots, C_{j-1}) = \frac{P(C_1, \ldots, C_j)}{P(C_1, \ldots, C_{j-1})} = \frac{q_j(C_j)}{\sum_{R_j} q_j(C_j)}$$
$$= \frac{q_j(C_j)}{\widetilde{q}_j(S_j)},$$

hvor vi har anvendt, at

$$P(C_1, \ldots, C_{j-1}) = \sum_{R_j} q_1(C_1) \cdot \ldots \cdot q_j(C_j) = q_1(C_1) \cdot \ldots \cdot \sum_{R_j} q_j(C_j).$$

Dernæst opdateres $q_j(C_j)$ til at være $P(R_j \mid S_j)$. Derudover opdateres $q_i(C_i)$ til at være $q_i(C_i)\widetilde{q}_j(S_j)$, hvor C_i er den forbundet forælder til C_j i Ψ_p. Med disse opdateringer er (4.5) stadig opfyldt. □

Bemærk, at for $j = 2$ i Algoritme 4.0.4 er $P(C_1) = \sum_{R_2} P(C_1, C_2) = q_1(C_1) \sum_{R_2} q_2(C_2) = q_1(C_1)\widetilde{q}_2(S_2)$, hvormed $q_1(C_1)$ opdateres til at være $P(C_1)$. Dermed giver Algoritme 4.0.4:

$$P(V) = P(C_1)P(R_2 \mid S_2) \cdot \ldots \cdot P(R_p \mid S_p).$$

Den næste algoritme bygger på Algoritme 4.0.4.

Algoritme 4.0.5 (Distribute evidence)
Lad Ψ_p være et kliketræ bestående af klikerne $C_1, \ldots, C_p$, hvor $p \geq 2$. Antag, at den simultane sandsynlighed for V er opskrevet som:

$$P(V) = q_1(C_1) \cdot \ldots \cdot q_p(C_p), \tag{4.6}$$

hvor $q_1(C_1) = P(C_1)$ og $q_i(C_i) = P(R_i \mid S_i)$ for $i = 2, \ldots, p$. Det efterfølgende trin anvendes fra $j = 2$ til og med $j = p$. Trin j:

Først opskrives den marginale sandsynlighed:

$$P(C_j) = P(R_j \mid S_j)P(S_j) = q_j(C_j) \sum_{C_i \setminus S_j} q_i(C_i),$$

hvor C_i er den forbundet forælder til C_j i Ψ_p. Dernæst opdateres $q_j(C_j)$ til at være $P(C_j)$. Derudover divideres der med $P(S_j) = \sum_{C_i \setminus S_j} q_i(C_i)$ i (4.6), hvormed (4.6) stadig er opfyldt. □

Algoritme 4.0.4 og 4.0.5 medfører:

$$P(V) = P(C_1) \prod_{i=2}^{p} \frac{P(C_i)}{P(S_i)}.$$

Algoritme 4.0.4 og 4.0.5 udgør altså en væsentlig del af metoden til at omskrive (4.1) til (4.4). Givet et Bayesiansk netværk og en tilhørende DAG består metoden af følgende tre trin:

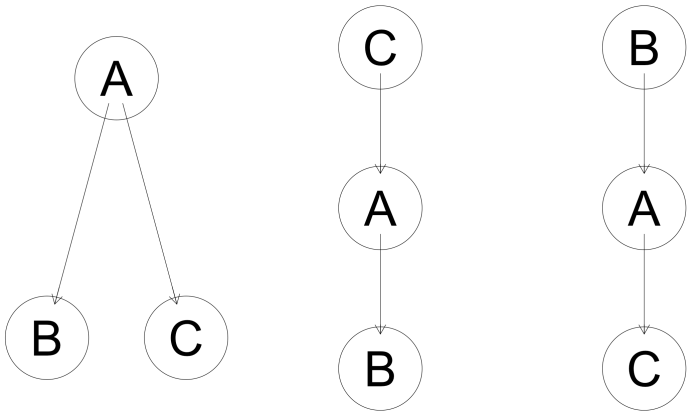
1) Moraliser DAG'en og anvend Algoritme 4.0.2 på denne.

2) Anvend Algoritme 4.0.4.

3) Anvend Algoritme 4.0.5.

Der er nu blevet specificeret flere forskellige måder at betragte et Bayesiansk netværk på. I det efterfølgende gives der et eksempel på, at den grafiske repræsentation af et Bayesiansk netværk ikke nødvendigvis er entydig.

Eksempel 4.0.6
Lad $V = \{A, B, C\}$ og $\mathbf{X}_V$ være en stokastisk proces, hvor $C \perp\!\!\!\perp B \mid A$. Der findes tre grafiske repræsentationer af $\mathbf{X}_V$, som opfylder $C \perp\!\!\!\perp B \mid A$:



Figur 4.3: De tre DAG'er, som opfylder, at C og B er betinget uafhængige givet A.

Alle tre DAG'er i Figur 4.3 medfører, at $\mathbf{X}_V$ er et Bayesiansk netværk, da den simultane sandsynlighed $P(V)$ kan skrives som (4.1). De tre DAG'er kan alle lede til det samme kliketræ, hvormed $P(V)$ i alle tre tilfælde kan opskrives ud fra de samme klikemarginaler.□

I praksis er den grafiske repræsentation ikke givet på forhånd. Måden, hvorpå den grafiske repræsentation findes, er at teste for uafhængighed samt betinget uafhængighed i data og ud fra dette opstille en DAG.

I følgende afsnit anvendes ovenstående teori på et givet datasæt.

4.1 Databehandling

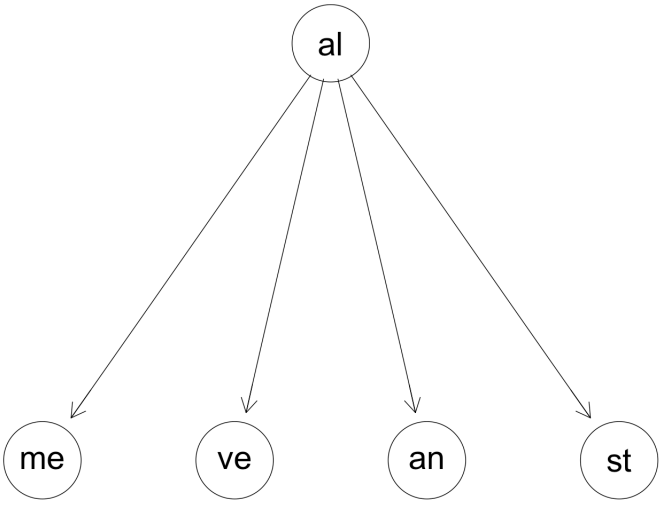
Dette afsnit anvender teorien om Bayesianske netværk på datasættet `math.csv`, hvilket, som sagt i Afsnit 3.3, består af 88 elevers score i fem forskellige fag. Hvert fag i datasættet er så vidt muligt blevet inddelt i tre lige store niveauer i forhold til scoren. De tre niveauer er: Lav (L), Mellem (M) og Høj (H). I dette afsnit anvendes `R`-koden i Appendiks G. For at kunne opstille en grafisk repræsentation af `math.csv` testes følgende betingede uafhængigheder i det tilpassede datasæt ved hjælp af `R`-funktionen `ciTest` fra `R`-pakken `gRim`:

- Test af $an \perp\!\!\!\perp st \mid al, me, ve$ giver en p-værdi på 0,075.
- Test af $an \perp\!\!\!\perp ve \mid al, me$ giver en p-værdi på 0,1252.
- Test af $ve \perp\!\!\!\perp me \mid al$ giver en p-værdi på 0,0587.
- Test af $st \perp\!\!\!\perp me \mid al, ve$ giver en p-værdi på 0,2197.
- Test af $an \perp\!\!\!\perp me \mid al$ giver en p-værdi på 0,1301.
- Test af $st \perp\!\!\!\perp ve \mid al$ giver en p-værdi på 0,7451.

Anvendes et signifikansniveau på 5%, kan vi ud fra forrige test af betingede uafhængigheder omskrive den simultane sandsynlighed for $V = (me, ve, al, an, st)$ til:

$$\begin{aligned} P(V) &= P(me, ve, al, an, st) \\ &= P(an \mid al, me, ve)P(st \mid al, me, ve)P(al, me, ve) \\ &= P(an \mid al, me)P(st \mid al, me, ve)P(al, me, ve) \\ &= P(an \mid al, me)P(st \mid al, me, ve)P(me \mid al)P(ve \mid al)P(al) \\ &= P(an \mid al, me)P(st \mid al, ve)P(me \mid al)P(ve \mid al)P(al) \\ &= P(an \mid al)P(st \mid al, ve)P(me \mid al)P(ve \mid al)P(al) \\ &= P(an \mid al)P(st \mid al)P(me \mid al)P(ve \mid al)P(al). \end{aligned} \tag{4.7}$$

Ud fra (4.7) kan vi opstille følgende DAG:



Figur 4.4: DAG for `math.csv`.

Sandsynlighederne i (4.7) bestemmes ved hjælp af R-koden i Appendiks G. Først betragtes $P(me \mid al)$:

me \ al	L	M	H
L	0,8571	0,1967	0,1
M	0	0,6721	0,35
H	0,1429	0,1312	0,55

Tabel 4.1: Sandsynligheden for $me \mid al$.

Hernæst betragtes sandsynligheden for $ve \mid al$:

ve \ al	L	M	H
L	0,4286	0,0819	0
M	0,5714	0,7049	0,35
H	0	0,2132	0,65

Tabel 4.2: Sandsynligheden for $ve \mid al$.

Sandsynligheden for $an \mid al$ betragtes:

an \ al	L	M	H
L	0,4286	0,2132	0
M	0,5714	0,3934	0,05
H	0	0,3934	0,95

Tabel 4.3: Sandsynligheden for $an \mid al$.

Ydermere betragtes $P(st \mid al)$:

st \ al	L	M	H
L	0,5714	0,4098	0
M	0,4286	0,4918	0,4
H	0	0,0984	0,6

Tabel 4.4: Sandsynligheden for $st \mid al$.

Afslutningsvis betragtes $P(al)$:

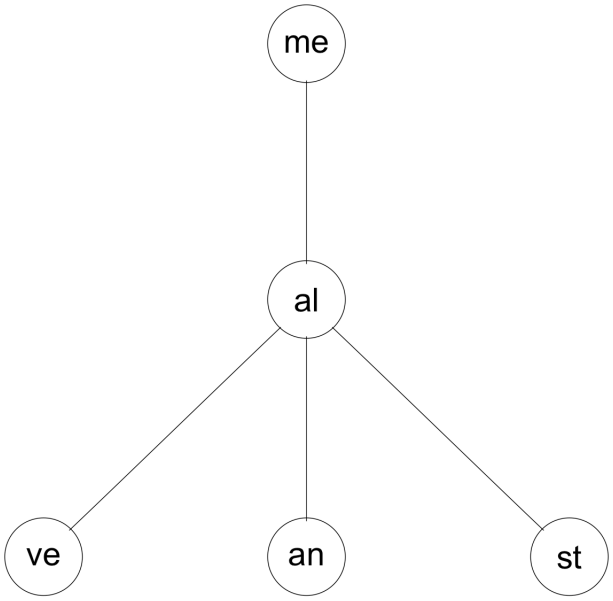
L	M	H
0,0795	0,6932	0,2273

Tabel 4.5: Sandsynligheden for al.

Ovenstående tabeller beskriver, hvilken indflydelse al har på de andre fag. Blandt andet ses det, at hvis en elev får en "Lav" score i al, da er sandsynligheden 0 for, at eleven får en "Høj" score i ve, an eller st. Dog er der en meget lille sandsynlighed for, at eleven får en "Høj" score i me givet, at al er "Lav". Derudover ses det, at hvis eleven har en "Høj" score i al, da er sandsynligheden 0 eller meget lille for, at eleven får en "Lav" score i

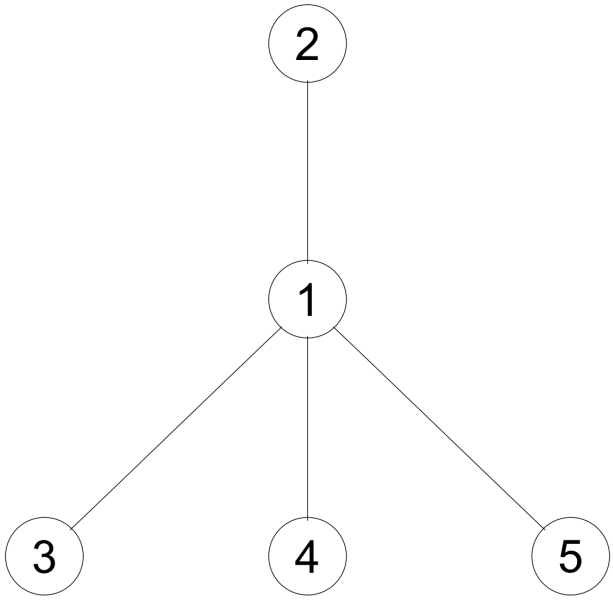
en af de andre fag. Det kan ydermere bemærkes, at en elev med sandsynlighed 0,6932 får en "Mellem" score i *al*.

Den moraliseret graf af DAG'en i Figur 4.4 bestemmes med henblik på at beregne klikemarginalerne:



Figur 4.5: Den moraliseret graf.

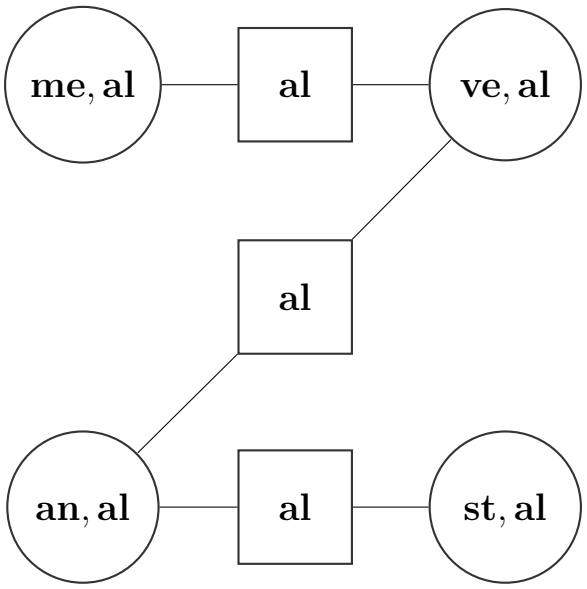
Den moraliseret graf i Figur 4.5 nummereres ved hjælp af Algoritme E.0.18 med henblik på at kunne opstille et kliketræ. Hvis knuden *al* i Figur 4.5 nummereres 1 kan følgende nummerering opnås:



Figur 4.6: Nummereringen af den moraliseret graf.

Nummereringen i Figur 4.6 er perfekt jævnfør Defintion E.0.10, og dermed er grafen i Figur 4.5 trianguleret jævnfør Proposition E.0.11.

Anvendes Algoritme 4.0.2 på Figur 4.6 fås følgende kliketræ, hvor kliken $\{me, al\}$ er roden:



Figur 4.7: Kliketræ.

Vi kan ud fra (4.7) samt Figur 4.7 omskrive $P(V)$ til følgende:

$$P(V) = q_1(me, al)q_2(ve, al)q_3(an, al)q_4(st, al). \tag{4.8}$$

Klikemarginalerne kan bestemmes ved at anvende Algoritme 4.0.4 og Algoritme 4.0.5 på (4.8). Dette er udført med R-koden i Appendiks G. Først betragtes $P(me, al)$:

me \ *al*	L	M	H
L	0,0682	0,1363	0,023
M	0	0,466	0,0795
H	0,011	0,091	0,125

Tabel 4.6: Sandsynligheden for me, al.

Hernæst betragtes sandsynligheden for ve, al:

ve \ *al*	L	M	H
L	0,0341	0,0568	0
M	0,0455	0,4887	0,0795
H	0	0,1477	0,1477

Tabel 4.7: Sandsynligheden for ve, al.

Sandsynligheden for an, al betragtes:

an \ al	L	M	H
L	0,0341	0,1477	0
M	0,0455	0,2727	0,0113
H	0	0,2727	0,216

Tabel 4.8: Sandsynligheden for an, al.

Ydermere betragtes $P(st, al)$:

st \ al	L	M	H
L	0,0455	0,2840	0
M	0,0341	0,3409	0,091
H	0	0,0682	0,1363

Tabel 4.9: Sandsynligheden for st, al.

Ud fra tabellerne ovenfor ses det blandt andet, at en elev med sandsynlighed 0,4887 får en "Mellem" score i både al og ve. Yderligere ses det, at hændelsen, $al = M$ og $an = M$, har samme sandsynlighed som hændelsen, $al = M$ og $an = H$.

A. Betingede normalfordelinger

Dette appendiks er baseret på [Seber, 2004].

Lemma A.0.1
Lad

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

hvor $\mathbf{X}_1 \in \mathbb{R}^n$ og $\mathbf{X}_2 \in \mathbb{R}^m$. Da er

$$\mathbf{X}_1 \mid \mathbf{x}_2 \sim N_n\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right).$$

B. Udregninger til M-trinet

Dette appendiks indeholder relevante udregninger til at kunne opstille M-trinet i Afsnit 2.4. Først bestemmes de afledte af $\boldsymbol{\eta}(\boldsymbol{\theta})$ med hensyn til $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$ og σ_{12}. Den afledte funktion $\frac{\partial}{\partial \mu_1}\boldsymbol{\eta}(\boldsymbol{\theta})$ bestemmes:

$$\frac{\partial}{\partial \mu_1}\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{bmatrix}\sigma_{22}\\ -\sigma_{12}\\ 0\\ 0\\ 0\end{bmatrix} = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}\begin{bmatrix}\sigma_{22}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\\ -\sigma_{12}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\\ 0\\ 0\\ 0\end{bmatrix}.$$

Den afledte funktion $\frac{\partial}{\partial \mu_2}\boldsymbol{\eta}(\boldsymbol{\theta})$ bestemmes:

$$\frac{\partial}{\partial \mu_2}\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{bmatrix}-\sigma_{12}\\ \sigma_{11}\\ 0\\ 0\\ 0\end{bmatrix} = \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}\begin{bmatrix}-\sigma_{12}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\\ \sigma_{11}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\\ 0\\ 0\\ 0\end{bmatrix}.$$

Den afledte funktion $\frac{\partial}{\partial \sigma_{11}}\boldsymbol{\eta}(\boldsymbol{\theta})$ bestemmes:

$$\begin{aligned}\frac{\partial}{\partial \sigma_{11}}\boldsymbol{\eta}(\boldsymbol{\theta}) &= \frac{-1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}\sigma_{22}\begin{bmatrix}\mu_1\sigma_{22} - \mu_2\sigma_{12}\\ -\mu_1\sigma_{12} + \mu_2\sigma_{11}\\ -\sigma_{22}/2\\ -\sigma_{11}/2\\ \sigma_{12}\end{bmatrix} + \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{bmatrix}0\\ \mu_2\\ 0\\ -1/2\\ 0\end{bmatrix}\\ &= \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}\begin{bmatrix}-\mu_1\sigma_{22}^2 + \mu_2\sigma_{12}\sigma_{22}\\ \mu_1\sigma_{12}\sigma_{22} - \mu_2\sigma_{22}\sigma_{11} + \mu_2(\sigma_{11}\sigma_{22} - \sigma_{12}^2)\\ \sigma_{22}^2/2\\ -(\sigma_{11}\sigma_{22} - \sigma_{12}^2)/2 + (\sigma_{11}\sigma_{22})/2\\ -\sigma_{22}\sigma_{12}\end{bmatrix}\\ &= \frac{1}{(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^2}\begin{bmatrix}-\mu_1\sigma_{22}^2 + \mu_2\sigma_{12}\sigma_{22}\\ \mu_1\sigma_{12}\sigma_{22} - \mu_2\sigma_{12}^2\\ \sigma_{22}^2/2\\ \sigma_{12}^2/2\\ -\sigma_{22}\sigma_{12}\end{bmatrix}.\end{aligned}$$

Den afledte funktion $\frac{\partial}{\partial\sigma_{22}}\boldsymbol{\eta}(\boldsymbol{\theta})$ bestemmes:

$$\begin{aligned}\frac{\partial}{\partial\sigma_{22}}\boldsymbol{\eta}(\boldsymbol{\theta}) &= \frac{-1}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\sigma_{11}\begin{bmatrix}\mu_1\sigma_{22}-\mu_2\sigma_{12}\\-\mu_1\sigma_{12}+\mu_2\sigma_{11}\\-\sigma_{22}/2\\-\sigma_{11}/2\\\sigma_{12}\end{bmatrix}+\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{bmatrix}\mu_1\\0\\-1/2\\0\\0\end{bmatrix}\\ &= \frac{1}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\begin{bmatrix}-\mu_1\sigma_{12}^2+\mu_2\sigma_{12}\sigma_{11}\\-\mu_2\sigma_{11}^2+\mu_1\sigma_{12}\sigma_{11}\\\sigma_{12}^2/2\\\sigma_{11}^2/2\\-\sigma_{11}\sigma_{12}\end{bmatrix}.\end{aligned}$$

Den afledte funktion $\frac{\partial}{\partial\sigma_{12}}\boldsymbol{\eta}(\boldsymbol{\theta})$ bestemmes:

$$\begin{aligned}\frac{\partial}{\partial\sigma_{12}}\boldsymbol{\eta}(\boldsymbol{\theta}) &= \frac{-1}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}(-2\sigma_{12})\begin{bmatrix}\mu_1\sigma_{22}-\mu_2\sigma_{12}\\-\mu_1\sigma_{12}+\mu_2\sigma_{11}\\-\sigma_{22}/2\\-\sigma_{11}/2\\\sigma_{12}\end{bmatrix}+\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}\begin{bmatrix}-\mu_2\\-\mu_1\\0\\0\\1\end{bmatrix}\\ &= \frac{1}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\begin{bmatrix}2\mu_1\sigma_{12}\sigma_{22}-2\mu_2\sigma_{12}^2-\mu_2(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\\-2\mu_1\sigma_{12}^2+2\mu_2\sigma_{12}\sigma_{11}-\mu_1(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\\-\sigma_{12}\sigma_{22}\\-\sigma_{12}\sigma_{11}\\2\sigma_{12}^2+(\sigma_{11}\sigma_{22}-\sigma_{12}^2)\end{bmatrix}\\ &= \frac{1}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}\begin{bmatrix}2\mu_1\sigma_{12}\sigma_{22}-\mu_2(\sigma_{11}\sigma_{22}+\sigma_{12}^2)\\2\mu_2\sigma_{12}\sigma_{11}-\mu_1(\sigma_{11}\sigma_{22}+\sigma_{12}^2)\\-\sigma_{12}\sigma_{22}\\-\sigma_{12}\sigma_{11}\\\sigma_{11}\sigma_{22}+\sigma_{12}^2\end{bmatrix}.\end{aligned}$$

Alle disse afledte funktioner kan samles i en matrix:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T = \begin{bmatrix}\frac{\partial}{\partial\mu_1}\boldsymbol{\eta}(\boldsymbol{\theta})^T\\\frac{\partial}{\partial\mu_2}\boldsymbol{\eta}(\boldsymbol{\theta})^T\\\frac{\partial}{\partial\sigma_{11}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\\\frac{\partial}{\partial\sigma_{22}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\\\frac{\partial}{\partial\sigma_{12}}\boldsymbol{\eta}(\boldsymbol{\theta})^T\end{bmatrix}. \tag{B.1}$$

Det kan ved hjælp af elementære række-operationer vises, at $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T$ er invertibel. Rækkeoperationerne er følgende, hvor række j i $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T$ noteres r_j:

1. Multiplicér r_1 og r_2 med $\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}$.
2. Læg $\frac{\sigma_{12}}{\sigma_{22}}r_1$ til r_2.
3. Multiplicér r_2 med $\frac{\sigma_{22}}{\sigma_{11}\sigma_{22}-\sigma_{12}^2}$.
4. Læg $\sigma_{12}r_2$ til r_1.

5. Multiplicér r_1 med $\frac{1}{\sigma_{22}}$.
6. De første to indgange i r_3, r_4 og r_5 kan erstattes med 0.
7. Læg $2\frac{\sigma_{22}}{\sigma_{12}}r_4$ til r_5 og $-\frac{\sigma_{22}^2}{\sigma_{12}^2}r_4$ til r_3.
8. Læg $\frac{\sigma_{22}}{\sigma_{12}}r_5$ til r_3.
9. Multiplicér r_3 med $\frac{2\sigma_{12}^2}{(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^2}$.
10. Fjerde indgang i r_4 og r_5 kan erstattes med 0.
11. Multiplicér r_5 med $\frac{1}{\sigma_{12}^2-\sigma_{11}\sigma_{22}}$.
12. Femte indgang i r_4 kan erstattes med 0.
13. Multiplicér r_4 med $\frac{2}{\sigma_{12}^2}$.
14. Byt rundt på r_3 og r_4.

Bemærk, at disse rækkeoperationer kræver en antagelse om, at der for korrelationen gælder $-1 < \rho < 1$, hvor $\rho = \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22})^{\frac{1}{2}}}$. Denne antagelse er i overenstemmelse med, at kovariansmatricen $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$ skal være invertibel, da:

$$0 \neq |\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho^2).$$

Det kan yderligere bemærkes, at hvis $\rho = 0$ er $\sigma_{12} = 0$, hvorfor man efter de første seks rækkeoperationer blot skal multiplicere r_3, r_4 og r_5 med tre passende konstanter for at opnå invertibiliteten af $\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})^T$.

C. R-kode til EM-algoritmen

```
#EM-algoritmen
#Funktion, som estimere parametrene i en bivariat
   normalfordeling vha. EM-algoritmen
bivnEM=function(x, tol){
  #x er en nx2 matrix bestaaende af realiseringer af en
     bivariat normalfordeling. tol er tolerancen

  #Antal realiseringer
  n=dim(x)[1]
  #Identifikation af observeret samt manglende data
  na.ind=!is.na(x)

  #Oversigt
  x1.obs=na.ind[,1]
  x2.obs=na.ind[,2]
  both.obs=x1.obs & x2.obs
  x1.obs.x2.mis=x1.obs & !x2.obs
  x1.mis.x2.obs=x2.obs & !x1.obs
  both.mis=!x1.obs & !x2.obs

  #Vektor til theta_t1
  theta_t1=numeric(5)

  #Begyndelsesbetingelserne er bestemt ud fra de
     realiseringer, hvor baade x1 og x2 er kendt
  theta_t1[1]=sum(x[both.obs,1])/length(x[both.obs,1])
  theta_t1[2]=sum(x[both.obs,2])/length(x[both.obs,2])
  theta_t1[3]=sum(x[both.obs,1]^2)/length(x[both.obs,1])-
     theta_t1[1]^2
  theta_t1[4]=sum(x[both.obs,2]^2)/length(x[both.obs,2])-
     theta_t1[2]^2
  theta_t1[5]=sum(x[both.obs,1]*x[both.obs,2])/length(x[both.
     obs,2])-theta_t1[1]*theta_t1[2]

  #Vektor til theta_t
  theta_t=numeric(5)

  #Vektor til den sufficiente stikproevefunktion
  s_t=numeric(5)

  #Vektor til log-likelihood
  ll=c()

  #Variabel til bestemmelse af antal iterationer
  k=0
```

```
40
41  #While-loekke, som udfoerer EM-algoritmen
42  while(all((abs(theta_t1-theta_t)) > tol)){
43    theta_t=theta_t1
44    #E-trin
45    s_t[1]=sum(x[x1.obs,1])+sum(theta_t[1]+theta_t[5]/theta_t
        [4]*(x[x1.mis.x2.obs,2]-theta_t[2]))+sum(theta_t[1]*
        both.mis)
46    s_t[2]=sum(x[x2.obs,2])+sum(theta_t[2]+theta_t[5]/theta_t
        [3]*(x[x1.obs.x2.mis,1]-theta_t[1]))+sum(theta_t[2]*
        both.mis)
47    s_t[3]=sum(x[x1.obs,1]^2)+sum(theta_t[3]-theta_t[5]^2/
        theta_t[4]+(theta_t[1]+theta_t[5]/theta_t[4]*(x[x1.mis
        .x2.obs,2]-theta_t[2]))^2)+sum((theta_t[3]+theta_t
        [1]^2)*both.mis)
48    s_t[4]=sum(x[x2.obs,2]^2)+sum(theta_t[4]-theta_t[5]^2/
        theta_t[3]+(theta_t[2]+theta_t[5]/theta_t[3]*(x[x1.obs
        .x2.mis,1]-theta_t[1]))^2)+sum((theta_t[4]+theta_t
        [2]^2)*both.mis);
49    s_t[5]=sum(x[both.obs,1]*x[both.obs,2])+sum((theta_t[1]+
        theta_t[5]/theta_t[4]*(x[x1.mis.x2.obs,2]-theta_t[2]))
        *x[x1.mis.x2.obs,2])+sum(x[x1.obs.x2.mis,1]*(theta_t
        [2]+theta_t[5]/theta_t[3]*(x[x1.obs.x2.mis,1]-theta_t
        [1])))+sum((theta_t[5]+theta_t[1]*theta_t[2])*both.mis
        )
50
51    #Log-likelihood
52    ll[k+1] = -(length(which(both.obs))/2)*log(theta_t[3]*
        theta_t[4]-theta_t[5]^2)-1/(2*(theta_t[3]*theta_t[4]-
        theta_t[5]^2))*sum((x[both.obs,1]-theta_t[1])^2*theta_
        t[4]+(x[both.obs,2]-theta_t[2])^2*theta_t[3]-2*theta_t
        [5]*(x[both.obs,1]-theta_t[1])*(x[both.obs,2]-theta_t
        [2]))-
53      (length(which(x1.obs.x2.mis))/2)*log(theta_t[3])-1/(2*
          theta_t[3])*sum((x[x1.obs.x2.mis,1]-theta_t[1])^2)-
54      (length(which(x1.mis.x2.obs))/2)*log(theta_t[4])-1/(2*
          theta_t[4])*sum((x[x1.mis.x2.obs,2]-theta_t[2])^2)
55
56    #M-trin
57    theta_t1[1]=s_t[1]/n
58    theta_t1[2]=s_t[2]/n
59    theta_t1[3]=s_t[3]/n-theta_t1[1]^2
60    theta_t1[4]=s_t[4]/n-theta_t1[2]^2
61    theta_t1[5]=s_t[5]/n-theta_t1[1]*theta_t1[2]
62
63    k=k+1
64  }
65
66  #Den sidste log-likelihood
```

```
67    ll[k+1] = (-(length(which(both.obs))/2)*log(theta_t1[3]*
          theta_t1[4]-theta_t1[5]^2)-1/(2*(theta_t1[3]*theta_t1
          [4]-theta_t1[5]^2))*sum((x[both.obs,1]-theta_t1[1])^2*
          theta_t1[4]+(x[both.obs,2]-theta_t1[2])^2*theta_t1[3]-2*
          theta_t1[5]*(x[both.obs,1]-theta_t1[1])*(x[both.obs,2]-
          theta_t1[2]))
68                -(length(which(x1.obs.x2.mis))/2)*log(theta_t1
                     [3])-1/(2*theta_t1[3])*sum((x[x1.obs.x2.mis
                     ,1]-theta_t1[1])^2)
69                -(length(which(x1.mis.x2.obs))/2)*log(theta_t1
                     [4])-1/(2*theta_t1[4])*sum((x[x1.mis.x2.obs
                     ,2]-theta_t1[2])^2))
70
71    #De estimerede parametre returneres
72    return(list(mean=c(theta_t1[1],theta_t1[2]), cov=matrix(c(
          theta_t1[3], theta_t1[5], theta_t1[5], theta_t1[4]),
          nrow=2), LogLike=ll, iterationer=k))
73  }
74
75  #Anvendelse af ovenstaaende funktion
76  load("~/Desktop/misdat.RData")
77  tol=1.e-10 #Tolerance
78  X=misdat #Datasaet
79
80  EM1=bivnEM(X,tol)
81  EM1$mean; EM1$cov; EM1$iterationer
82  plot(EM1$LogLike[-1],ylim=c(-140,-133))
83
84  #Datasaet uden realiseringer, hvor baade x1 og x2 mangler
85  X1=matrix(c(X[1:150,1],X[1:150,2]),ncol=2)
86
87  EM2=bivnEM(X1,tol)
88  EM2$mean; EM2$cov; EM2$iterationer
89  plot(EM2$LogLike[-1],ylim=c(-140,-133))
```

D. R-kode til Faktor analyse

```r
#Faktor analyse
library(corrplot)
#Funktion, som estimere parametrene i en FA model vha. EM-algoritmen
faEM=function(y,nfac,tol=1e-4,stand=F){
  #Data centreres(og standardiseres, hvis stand==T)
  x=as.matrix(scale(y,scale=stand))

  #Begyndelsesparametrene bestemmes
  C=cov(x)
  eig_val=eigen(C)$values[1:nfac]
  eig_vec=eigen(C)$vectors[,1:nfac]
  Lambda_t1=eig_vec%*%diag(sqrt(eig_val),nfac)
  Psi_t1=diag(diag(C-Lambda_t1%*%t(Lambda_t1)))

  #Matricer til Lambda_t og Psi_t
  Lambda_t=matrix(0,ncol(x),nfac)
  Psi_t=matrix(0,ncol(x),ncol(x))

  #Vektorer til E-trinet
  mean_f=matrix(0,nfac,nrow(x))
  mean_ff=matrix(0,nfac,nfac)

  #Til M-trinet
  xMeanf=matrix(0,ncol(x),nfac)
  xx=t(x)%*%x

  #Vektor til log-likelihood
  ll=c()

  #Variabel til bestemmelse af antal iterationer
  k=0

  #EM-algoritmen
  while(all(abs(Lambda_t1-Lambda_t) > tol) || all(abs(Psi_t1-Psi_t) > tol)){
    mean_ff[,]=0
    Lambda_t=Lambda_t1
    Psi_t=Psi_t1
    var_f=solve(diag(1,nfac)+t(Lambda_t)%*%solve(Psi_t,Lambda_t))
    inv_var_x=solve(Lambda_t %*% t(Lambda_t)+Psi_t)

    #Log-likelihood
```

```
42      ll[k+1] = (nrow(x)/2)*log(det(inv_var_x))-(1/2)*sum(diag(
            xx %*%inv_var_x))
43
44
45      #E-trin
46      for(i in 1:nrow(x)){
47        mean_f[,i]=t(Lambda_t)%*%inv_var_x%*%x[i,]
48        mean_ff=mean_ff+var_f+mean_f[,i]%*%t(mean_f[,i])
49      }
50
51      #M-trin
52      xMeanf=xx%*%inv_var_x%*%Lambda_t
53
54      Lambda_t1=xMeanf%*%solve(mean_ff)
55      Psi_t1=diag(diag(xx-Lambda_t1%*%t(xMeanf))/nrow(x))
56
57      k=k+1
58    }
59
60    #Den sidste log-likelihood
61    inv_var_x=solve(Lambda_t1 %*% t(Lambda_t1)+Psi_t1)
62    ll[k+1]=(nrow(x)/2)*log(det(inv_var_x))-(1/2)*sum(diag(t(x)
          %*% x %*%inv_var_x))
63
64    #De estimerede parametre returneres
65    return(list(Lambda=Lambda_t1,Psi=Psi_t1,LogLike=ll,
          iterationer=k))
66  }
67
68  #FA model med 1 faktor
69  X=read.csv("~/Desktop/math.csv") #Datasaet
70  stand=T #Data standardiseres
71  tol=1.e-6 #Tolerance
72  nfac=1 #Antal faktorer
73
74  FF1=faEM(X,nfac,tol,stand=stand)
75
76  FF1$Lambda; FF1$Psi; FF1$iterationer
77
78  #FA model med 2 faktorer
79  nfac=2 #Antal faktorer
80
81  FF2=faEM(X,nfac,tol,stand=stand)
82
83  FF2$Lambda; FF2$Psi; FF2$iterationer
84
85  varimax(FF2$Lambda)$loadings[,1:2]
86  corrplot(cor(X), order = "hclust", tl.col='black', tl.cex
        =1.5)
```

E. Grundlæggende grafteori

I dette appendiks præsenteres en række relevante begreber og resultater inden for grafteori, hvilke er baseret på [Lauritzen og Spiegelhalter, 1988].

Definition E.0.1
En (uorienteret) graf G består af et par (V, E), hvor V er en mængde af elementer kaldet *knuder*, og E er en mængde, hvis elementer er par af knuder i V. Et element i E kaldes for en *kant*. En uorienteret graf noteres $G = (V, E)$.

En kant i en uorienteret graf $G = (V, E)$ mellem $v, w \in V$ noteres $\{v, w\} \in E$.

Definition E.0.2
En orienteret graf $G = (V, E)$ er en graf, hvis kanter er orienteret mellem par af knuder i V.

Bemærk, at notationen $w \to v$ betyder, at der er en orienteret kant fra w til v, hvor $w, v \in V$. Herudover noteres kanten $w \to v$ ved $(w, v) \in E$.

Definition E.0.3
Lad $G = (V, E)$ være en uorienteret graf. To knuder $v, w \in V$ siges at være naboer, hvis $\{v, w\} \in E$. Derudover er $nabo(v)$ en mængde bestående af naboer til v.

Definition E.0.4
En sti i en graf $G = (V, E)$ er en følge af knuder $\{v_1, \ldots, v_n\}$, hvor $v_j \in V$ og $\{v_i, v_{i+1}\} \in E$ for $j = 1, \ldots, n$ og $i = 1, \ldots, n-1$. En sti, $\{v_1, \ldots, v_k, v_1\}$, som højst følger en kant i E én gang, kaldes for *en k-cykel*.

Definition E.0.5
En graf $G = (V, E)$ siges at være *acyklisk*, hvis der ikke eksisterer nogen cykel i G.

En acyklisk uorienteret graf G kaldes for et *træ*. En acyklisk orienteret graf G kaldes for en *DAG* (directed acyclic graph).

Definition E.0.6
Lad $G = (V, E)$ være en orienteret graf. *Forældrene* til en knude $v \in V$ er en mængde bestående af de knuder $w \in V$, hvor $w \to v$. Forældrene til en knude $v \in V$ noteres $pa(v)$. Derudover siges v at være barn af w, hvis $w \to v$.

Definition E.0.7
For en graf $G = (V, E)$ siges en delmængde $D \subseteq V$ at være *fuldstændig*, hvis der er kanter mellem alle par af knuder i D. Derudover siges en fuldstændig delmængde $D \subseteq V$ at være *en klike*, hvis D er maksimal.

Definition E.0.8
Moraliseringen af en DAG $G = (V, E)$ er en uorienteret graf $M = (V, \widetilde{E})$, hvor $\widetilde{E}$ består af kanterne i E samt kanterne mellem par af knuder i $pa(v)$ for alle $v \in V$.

Definition E.0.9
En uorienteret graf G siges at være *trianguleret*, hvis der ikke eksisterer en k*-cykel* uden en korde for $k \geq 4$.

Definition E.0.10
Lad $G = (V, E)$ være en uorienteret graf. En nummerering af G består i at tildele alle knuder $v \in V$ et unikt tal $i \in \{1, \ldots, n\}$, hvor $n = \#V$. Derudover siges en nummerering at være perfekt, hvis der for alle $i \in \{1, \ldots, n\}$ gælder, at

$$A_i = nabo(i) \cap \{1, \ldots, i-1\}$$

er fuldstændig.

Proposition E.0.11
En uorienteret graf $G = (V, E)$ er trianguleret, hvis og kun hvis der kan dannes en perfekt nummerering.

Definition E.0.12
En *hypergraf* Γ_p består af et par (V, H), hvor V er en mængde af knuder, og H er en mængde bestående af p delmængder af V. Et element i H kaldes for en *hyperkant*.

Definition E.0.13
En hypergraf $\Gamma_p = (V, H)$ siges at være acyklisk, hvis den opfylder følgende:

1) Der eksisterer ikke et $C_i \in H$, sådan at $C_i \subseteq C_j \in H$ for $i, j = \{1, \ldots, p\}$ og $i \neq j$.
2) Elementerne i H kan ordnes $(C_1, \ldots, C_p)$ sådan, at der for alle $2 \leq i \leq p$ gælder:

$$C_i \cap (C_1 \cup \ldots \cup C_{i-1}) \subseteq C_j,$$

for et $j < i$.

Betingelse 2 i Definition E.0.13 kaldes for *running intersection property*.

Definition E.0.14
Lad $\Gamma_p = (V, H)$ være en acyklisk hypergraf. Mængden $S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1})$, hvor $C_j \in H$, kaldes for en seperator. Ydermere kaldes mængden $R_i = C_i \backslash S_i$ for et residual.

Bemærk, at $S_1 = \emptyset$ og $R_1 = C_1$.

Definition E.0.15
Lad $\Gamma_p = (V, H)$ være en acyklisk hypergraf. En hyperkant C_i siges at være en potentiel forældre til C_j, hvis $S_j \subseteq C_i$ for $i \in \{1, \ldots, p\}$ og $i < j$.

Proposition E.0.16
En hypergraf $\Gamma_p = (V, H)$ er acyklisk, hvis og kun hvis $\{C_1, \ldots, C_p\}$ kan betragtes som klikerne i en trianguleret graf, hvor $C_i \in H$.

Bemærk, at Proposition E.0.16 giver, at man ud fra en trianguleret graf G med p kliker $C_1, \ldots, C_p$ kan opstille en acyklisk hypergraf $\Gamma_p = (V, H)$, hvor V er knuderne i G, og $H = \{C_1, \ldots, C_p\}$.

Bemærkning E.0.17
Lad $\Gamma_p = (V, H)$ være en acyklisk hypergraf, hvor elementerne $C_1, \ldots, C_p \in H$ er ordnet sådan, at Γ_p opfylder *running intersection property*. Da kan man opstille et træ $\Psi_p(H, E)$, hvor E er bestemt sådan, at en knude C_i er forbundet med præcis en af dens potentielle forældre for $i = \{2, \ldots, p\}$. Træet Ψ_p kaldes for et kliketræ. □

Algoritme E.0.18 (Maximal Cardinality Search)
Lad $G = (V, E)$ være en uorienteret graf. Nummerer en tilfældig knude 1. Nummerer herefter de resterende knuder fortløbende sådan, at den næste knude er den med det største antal af nummererede naboer. Hvis flere knuder har det største antal af nummererede naboer, så vælges der en tilfældig knude blandt disse. □

F. Betinget uafhængighed

I dette appendiks præsenteres relevante begreber og resultater inden for betinget uafhængighed, hvilke er baseret på [Dawid, 1979].

Definition F.0.1
Lad $\mathbf{X}, \mathbf{Y}$ og $\mathbf{Z}$ være stokastiske vektorer med simultan tæthedsfunktion $f_{\mathbf{XYZ}}$. Da siges $\mathbf{X}$ og $\mathbf{Y}$ at være betinget uafhængige givet $\mathbf{Z}$, hvis og kun hvis

$$f_{\mathbf{XY}|\mathbf{Z}}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x} \mid \mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z}),$$

for alle værdier af $\mathbf{x}$ og $\mathbf{y}$ samt alle værdier af $\mathbf{z}$, hvor $f(\mathbf{z}) > 0$. Dette noteres $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$.

Proposition F.0.2
Lad $\mathbf{X}, \mathbf{Y}$ og $\mathbf{Z}$ være stokastiske vektorer med simultan tæthedsfunktion $f_{\mathbf{XYZ}}$. Da er $\mathbf{X}$ og $\mathbf{Y}$ betinget uafhængige givet $\mathbf{Z}$, hvis og kun hvis, at der eksisterer to ikke-negative funktioner h og g sådan, at

$$f_{\mathbf{XYZ}}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = h(\mathbf{x}, \mathbf{z}) g(\mathbf{y}, \mathbf{z}),$$

for alle værdier af $\mathbf{x}$ og $\mathbf{y}$, samt alle værdier af $\mathbf{z}$, hvor $f(\mathbf{z}) > 0$. Dette noteres $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$.

Proposition F.0.3
Lad $\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}$ og $\mathbf{Z}$ være stokastiske vektorer. Da medfører $(\mathbf{X}_1, \mathbf{X}_2) \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, at $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ og $\mathbf{X}_2 \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$.

G. R-kode til Bayesianske netværk

```
#Bayesianske netvaerk
library(gRim)
library(Rgraphviz)
X=read.csv("~/Desktop/math.csv") #Datasaet

#Hver variabel i data inddeles i 3 niveauer
V=lapply(X,cut,breaks=3,labels=c("L","M","H"))
Y=matrix(0,nrow(X),ncol(X))
for(i in 1:ncol(X)){
  Y[,i]=V[[i]]
}
colnames(Y)=names(X)
Y[Y==1]="L"
Y[Y==2]="M"
Y[Y==3]="H"

#Frekvenstabel over data
ftab=as.table(ftable(xtabs(~me+ve+al+an+st,data=Y)))

#Test af betinget uafhaengighed
ciTest(ftab,set=c("an","st","al","me","ve"))
ciTest(ftab,set=c("an","ve","al","me"))
ciTest(ftab,set=c("ve","me","al"))
ciTest(ftab,set=c("st","me","al","ve"))
ciTest(ftab,set=c("an","me","al"))
ciTest(ftab,set=c("st","ve","al"))

#DAG
DAG=dag(~me|al + ve|al + an|al + st|al)
plot(DAG)

#Moraliseret graf
MG=ug(~me|al + ve|al + an|al + st|al)
plot(MG)

#Nummereret graf
NG=ug(~2|1 + 3|1 + 4|1 + 5|1)
plot(NG)

#Fordelingerne
al=armarg(ftab,~al); al
me_al=armarg(ftab,~me + al); me_al
ve_al=armarg(ftab,~ve + al); ve_al
an_al=armarg(ftab,~an + al); an_al
st_al=armarg(ftab,~st + al); st_al
```

```
46
47 #Normaliserede fordelinger
48 p.al=arnormalize(al, "first"); p.al
49 p.me_al=arnormalize(me_al, "first"); p.me_al
50 p.ve_al=arnormalize(ve_al, "first"); p.ve_al
51 p.an_al=arnormalize(an_al, "first"); p.an_al
52 p.st_al=arnormalize(st_al, "first"); p.st_al
53
54 #P(V) inddeles i 4 funktioner
55 q1.me.al=arprod(p.al,p.me_al); q1.me.al
56 q2.ve.al=p.ve_al; q2.ve.al
57 q3.an.al=p.an_al; q3.an.al
58 q4.st.al=p.st_al; q4.st.al
59
60 #Collect evidence
61 q4.al=armarg(q4.st.al,"al"); q4.al
62 q4.st.al=ardiv(q4.st.al,q4.al); q4.st.al
63 q3.an.al=armult(q3.an.al, q4.al); q3.an.al
64
65 q3.al=armarg(q3.an.al,"al"); q3.al
66 q3.an.al=ardiv(q3.an.al,q3.al); q3.an.al
67 q2.ve.al=armult(q2.ve.al, q3.al); q2.ve.al
68
69 q2.al=armarg(q2.ve.al,"al"); q2.al
70 q2.ve.al=ardiv(q2.ve.al,q2.al); q2.ve.al
71 q1.me.al=armult(q1.me.al, q2.al); q1.me.al
72
73 #Distribute evidence
74 q1.al=armarg(q1.me.al,"al"); q1.al
75 q2.ve.al=armult(q2.ve.al,q1.al); q2.ve.al
76
77 q2.al=armarg(q2.ve.al,"al"); q2.al
78 q3.an.al=armult(q3.an.al,q2.al); q3.an.al
79
80 q3.al=armarg(q3.an.al,"al"); q3.al
81 q4.st.al=armult(q4.st.al,q3.al); q4.st.al
82
83 #Klikemarginalerne
84 p.me.al=q1.me.al; p.me.al
85 p.ve.al=q2.ve.al; p.ve.al
86 p.an.al=q3.an.al; p.an.al
87 p.st.al=q4.st.al; p.st.al
```

Litteratur

A. P. Dawid, *Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society*, 41(1), 1–31, 1979.

A. P. Dempster, N. M. Laird og D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM algortihm. Journal of the Royal Statistical Society*, 39(1), 1–38, 1977.

S. L. Lauritzen og D. J. Spiegelhalter, *Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2), 157–224, 1988.

Geoffrey J. McLachlan og Thriyambakam Krishnan, *The EM Algorithm and Extensions*, John Wiley, 2008, 2. udgave.

K. B. Petersen og M. S. Pedersen, *The Matrix Cookbook*, 2012. URL `http://www2.imm.dtu.dk/pubdb/p.php?3274`, version 20121115.

David Ruppert, *Statistics and Data Analysis for Financial Engineering*, Springer, 2011.

G. A. F. Seber, *Multivariate observations*, John Wiley, 2004.