75554
MULTIVARIABEL 12.12.95

LØSNING

Oppgave 1

a) $\underset{\sim}{\Sigma}$ kjent:

$$U^2 = n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu_0})' \underset{\sim}{\Sigma}^{-1} (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu_0}) \sim \chi^2_p \text{ under } H_0$$

Forkast hvis $U^2 > \chi^2_{p,\alpha}$

$\underset{\sim}{\Sigma}$ ukjent:

$$T^2 = n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu_0})' \underset{\sim}{S}^{-1} (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu_0}) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

Forkast hvis $T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$

Konklusjon $p=5$, $\underset{\sim}{\Sigma}$ ukjent:
Kritisk verdi er $\frac{49.5}{45} \cdot F_{5,45,0.05} = 13.17$
$= 2.42$

Observert verdi: $50 \cdot 0.167 = 8.35$ dvs ikke forkast

Hvis $\underset{\sim}{\Sigma}$ er kjent er kritisk verdi $\chi^2_{5,0.05} = 11.07$

dvs forkast heller ikke da.

b. Bonferroni:

Intervall for μ_i: $\bar{x}_i \pm t_{49}\left(\frac{0.05}{4}\right)\sqrt{\frac{s_{ii}}{50}}$

dvs. $\bar{x}_i \pm 2.81\sqrt{\frac{s_{ii}}{50}}$

For μ_1: $13.38 \pm 2.81\sqrt{\frac{13.73}{50}}$

13.38 ± 1.24

dvs $(12.14, 14.62)$

T^2-metoden: $\bar{x}_i \pm \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p}(0.05)}\sqrt{\frac{s_{ii}}{50}}$

$\bar{x}_i \pm \sqrt{\frac{5.49}{45}2.42}\cdot\sqrt{\frac{s_{ii}}{50}}$

dvs $\bar{x}_i \pm 3.68\sqrt{\frac{s_{ii}}{50}}$

dvs. større intervaller. (For μ_1: (13.38 ± 1.98) dvs $(11.48, 15.28)$

Konfidensellipse for (μ_1, μ_2) kan lages ved å ta ut den øverste venstre 2x2 matrisen i $\underset{\sim}{S}$, dvs.

$$\underset{\sim}{S}^* = \begin{pmatrix} 13.73 & 8.69 \\ 8.69 & 8.95 \end{pmatrix}$$

Konfidensellipsen vil da ha sentrum: $\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 13.38 \\ 14.09 \end{pmatrix}$ og aksen langs egenvektorene for $\underset{\sim}{S}^*$. (Se boka kap. 5).

c) Andel av total sampelvarians forklart ved $\hat{y}_1$ er

$$\frac{180.70}{180.70+7.24+4.07+2.13+0.39} = \underline{0.929} \qquad (92.9\%)$$

Korrelasjoner:

Formelsamlingen ("5.8 av 10") gir formlene for populasjonsvariantene av korrelasjonene. Her skal vi ha sampel-versjonene:

$$\frac{\sqrt{\hat{\lambda}_i}\ \hat{e}_{ij}}{\sqrt{s_{jj}}}$$

Telleren er her elementene i første egenvektor, mens nevneren er diagonal-elementene i $\underset{\sim}{S}$

Dermed: $i=1$: $\frac{3.30}{\sqrt{13.73}} = 0.891$

$i=2$: 0.809

$i=3$: 0.885

$i=4$: 0.939

$i=5$: 0.994

Alle korrelasjonene er høye. Korrelasjonen med x_5 er dominerende siden x_5 har den suverent største sampelvariansen. Siden x_5 har en mye høyere varians enn de andre x'e kunne korrelasjonsmatrisen R muligens [illegible]

d) $\underset{\sim}{\Sigma} = E((\underset{\sim}{x}-\underset{\sim}{\mu})(\underset{\sim}{x}-\underset{\sim}{\mu})')$

$= E\left[(\underset{\sim}{l}F+\underset{\sim}{\varepsilon})(\underset{\sim}{l}F+\underset{\sim}{\varepsilon})'\right]$

$= \underset{\sim}{l}\,\underset{\sim}{l}' E(F^2) + E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}')$

$= \underset{\sim}{l}\,\underset{\sim}{l}' + \underset{\sim}{\Psi}$

Principal komponent-metoden:

$$\tilde{\underset{\sim}{l}} = \begin{pmatrix} 3.30 \\ 2.42 \\ 4.59 \\ 2.03 \\ 11.78 \end{pmatrix}, \quad \tilde{\psi}_i = -(\tilde{\underset{\sim}{l}}\,\tilde{\underset{\sim}{l}}')_{ii} + s_{ii}$$

dvs

$$\tilde{\underset{\sim}{\Psi}} = \begin{pmatrix} 2.84 & & & & 0 \\ & 3.09 & & & \\ & & 5.80 & & \\ 0 & & & 0.55 & \\ & & & & 1.54 \end{pmatrix}$$

siden

$$\tilde{\underset{\sim}{l}}\,\tilde{\underset{\sim}{l}}' = \begin{pmatrix} 10.89 & 7.99 & 15.15 & 6.20 & 38.87 \\ ? & 5.86 & ? & ? & ? \\ ? & ? & 21.07 & ? & ? \\ ? & ? & ? & 4.12 & ? \\ ? & ? & ? & ? & 138.77 \end{pmatrix}$$

Residualmatrisen:

$$\underset{\sim}{S} - \tilde{\underset{\sim}{l}}\,\tilde{\underset{\sim}{l}}' - \tilde{\underset{\sim}{\Psi}} = \begin{pmatrix} 0 & 0.70 & 0.50 & 0.33 & -1.22 \\ & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix}$$

Oppgave 2

a) $Cov(\underset{\sim}{X}, Y) \underset{def}{=} E\left[(\underset{\sim}{X}-\underset{\sim}{\mu})(Y-\eta)\right]$

$= E\left[(\underset{\sim}{l}F+\underset{\sim}{\varepsilon})(qF+\gamma)\right]$

$= q\underset{\sim}{l}\,E(F^2) = q\underset{\sim}{l}$

Dermed for formelsamling + innsetting:

$\hat{Y}_{LS} = \eta + \underset{\sim}{\sigma}'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})$

$= \eta + q\underset{\sim}{l}'(\underset{\sim}{l}\underset{\sim}{l}'+\underset{\sim}{\Psi})^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})$

Generell definisjon av $\hat{Y}_{LS}$:

Den er den lineære funksjon av $\underset{\sim}{X}$ (dvs "$\hat{Y} = a + b\underset{\sim}{X}$") som minimerer $E((\hat{Y}-Y)^2)$

Svaret er $\hat{Y} = \eta + \underset{\sim}{\sigma}'\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{X}-\underset{\sim}{\mu})$

som ifølge formel på "s. 6 av 10" er den betingede forventning for Y gitt $\underset{\sim}{X}$ ved multinormalitet.

b) Ren innsetting fra formelsamling (med $\underset{\sim}{w}_1 = \underset{\sim}{\delta} = q\,\underset{\sim}{l}$) gir

$$\hat{y}_{PLS} = \eta + q\,\underset{\sim}{l}'\underset{\sim}{l}\left(\underset{\sim}{l}'\left(\underset{\sim}{l}\underset{\sim}{l}' + \underset{\sim}{\Psi}\right)\underset{\sim}{l}\right)^{-1}\underset{\sim}{l}'(\underset{\sim}{x}-$$

Hovedidé: Finn lineærkombinasjonen $F_i = \underset{\sim}{w}_i'\underset{\sim}{x}$ som har størst mulig kovariansen med y (og som er ukorrelert, og med gitt lengden på $\underset{\sim}{w}_i$)

c) Vi har

$$\underset{\sim}{l}'\left(\underset{\sim}{l}\underset{\sim}{l}' + \underset{\sim}{\Psi}\right)^{-1}$$

$$= \underset{\sim}{l}'\left(\underset{\sim}{\Psi}^{-1} - a\,\underset{\sim}{\Psi}^{-1}\underset{\sim}{l}\underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}\right)$$

$$= \underset{\sim}{l}'\left(\underset{\sim}{I} - a\,\underset{\sim}{\Psi}^{-1}\underset{\sim}{l}\underset{\sim}{l}'\right)\underset{\sim}{\Psi}^{-1}$$

$$= \left(\underset{\sim}{l}' - a\,\underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}\underset{\sim}{l}\underset{\sim}{l}'\right)\Psi^{-1}$$

$$= \left(1 - a\,\underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}\underset{\sim}{l}\right)\underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}$$

$$= \frac{1}{1 + \underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}\underset{\sim}{l}}\,\underset{\sim}{l}'\underset{\sim}{\Psi}^{-1}$$

der $$c = \frac{q}{1 + \underset{\sim}{l}' \underset{\sim}{\Psi}^{-1} \underset{\sim}{l}}$$

Videre

$$d = q \, \underset{\sim}{l}' \underset{\sim}{l} \, (\underbrace{\underset{\sim}{l}' (\underset{\sim}{l} \underset{\sim}{l}' + \underset{\sim}{\Psi}) \underset{\sim}{l}}_{\text{et tall!}})^{-1}$$

Vi ser dermed at $\hat{F}_{LS}$ skaleres observasjonene $x_i - \mu_i$ med Ψ_i-ene (de spesifikke variansene). Legger altså størst vekt på de variable som har minst spesifikke varianser (og dermed er forklart mest av den felles faktoren F).

Bortsett fra denne skaleringen, er det den samme lineærkombinasjon med l_i-ene som går igjen.

$\underset{\sim}{\Psi} = \sigma^2 \underset{\sim}{I}$:

Da er

$$c = \frac{q}{1 + \frac{1}{\sigma^2} \underset{\sim}{l}' \underset{\sim}{l}} = \frac{q \sigma^2}{\sigma^2 + \underset{\sim}{l}' \underset{\sim}{l}}$$

$$d = \eta \underset{\sim}{l}' \underset{\sim}{l} \left(\underset{\sim}{l}' \left(\underset{\sim}{l} \underset{\sim}{l}' + \sigma^2 \underset{\sim}{I} \right) \underset{\sim}{l} \right)^{-1}$$

$$= \eta \underset{\sim}{l}' \underset{\sim}{l} \left[\underset{\sim}{l}' \underset{\sim}{l} \left(\underset{\sim}{l}' \underset{\sim}{l} + \sigma^2 \right) \right]^{-1}$$

$$= \frac{\eta}{\sigma^2 + \underset{\sim}{l}' \underset{\sim}{l}}$$

Siden $\psi_i = \sigma^2$ for hver i, ser vi at

$$\hat{\underset{\sim}{\gamma}}_{LS} = \hat{\underset{\sim}{\gamma}}_{PLS}$$

(d) $\begin{pmatrix} \underset{\sim}{x} \\ y \end{pmatrix}$ danner nå tilsammen den 5-dimensjonale faktormodellen fra Oppgave 1. Naturlig estimater er da

$$\hat{\underset{\sim}{l}} = \begin{pmatrix} 3.30 \\ 2.42 \\ 4.59 \\ 2.03 \end{pmatrix} \qquad \hat{\underset{\sim}{\mu}} = \begin{pmatrix} 13.38 \\ 14.09 \\ 119.45 \\ 10.00 \end{pmatrix} \qquad \hat{\underset{\sim}{\psi}} = \begin{pmatrix} 2.84 & & & \\ & 3.09 & & \\ & & 5.80 & \\ & & & \end{pmatrix}$$

$$\hat{\eta} = 11.78 \qquad \hat{\eta} = 39.19$$

Prediksjonene blir da

$$\hat{Y}_{LS} = 39.19 + 0.6598 \cdot \left[\frac{330}{2.84}(10-13.38) + \frac{2.42}{3.09}(10-14.09) + \frac{4.59}{5.80}(115-119.45) + \frac{2.03}{0.55}(10-10) \right] = \underline{32.16}$$

$$\hat{Y}_{PLS} = 39.19 + 0.2557 [330(10-13.38) + 2.42(10-14.09) + 4.59(115-119.45) + 2.03 \cdot (10-10)] = \underline{28.58}$$

Oppgave 3

a) Avstander (bruker $d(\underset{\sim}{x}, \underset{\sim}{y}) = \sqrt{\Sigma (x_i - y_i)^2}$)

	a	b	c	d
a	0			
b	6.40	0		
c	6.71	2.83	0	
d	4.12	3.16	5.10	0

Single linkage:

1) bc danner cluster ved 2.83

Ny avst. matrise:

	bc	a	d
bc	0		
a	6.40	0	
d	3.16	4.12	0

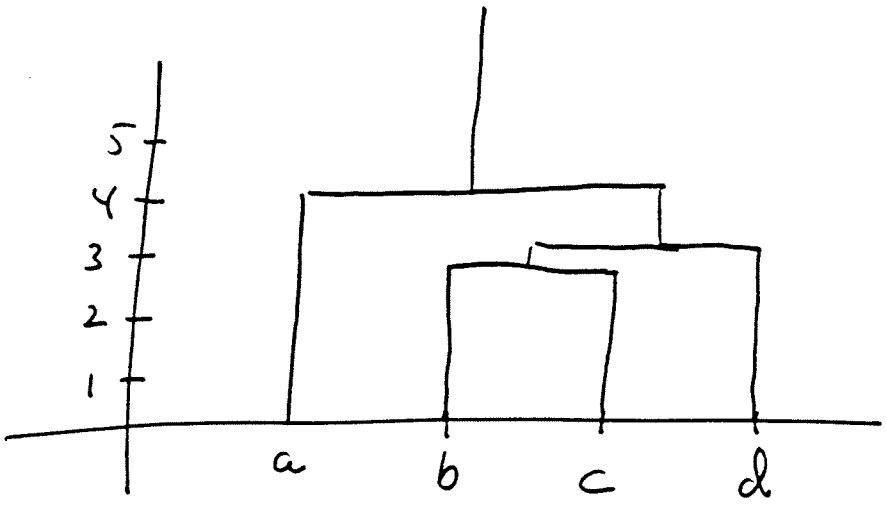
2) bcd danner cluste ved 3.16

Ny avst. matrise

	bcd	a
bcd	0	
a	4.12	0

3) bcd og a danner cluste ved 4.12

Dendrogram single linkage:



(single linkage)

Average linkage:

1) bc danner cluster ved 2.83

Ny avst. matrise:

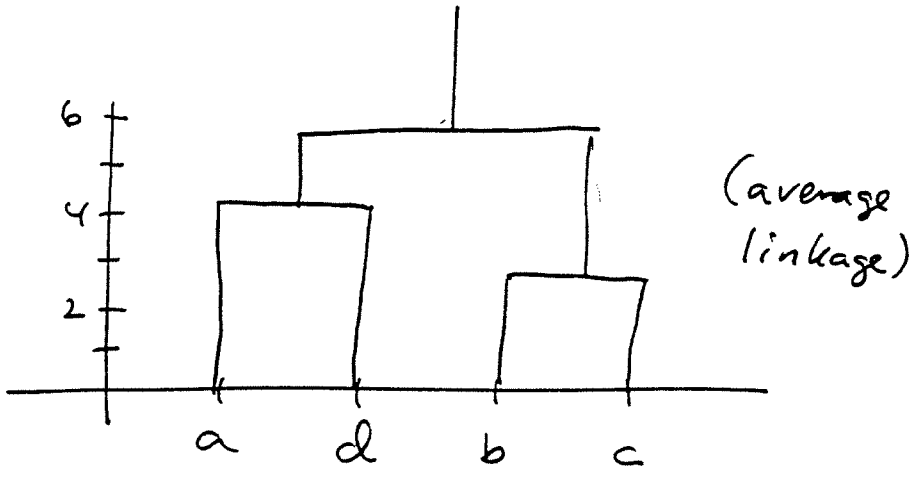
	bc	a	d
bc	0		
a	6.56	0	
d	4.63	4.12	0

2) acd danner nytt cluster ved 4.12

Ny avstandsmatrise:

	bc	ad
bc	0	
ad	5.35	0

Dendrogram



Dersom to clustre ønskes, ser vi av dendrogrammet for single at de to clustrene er {a} og {bcd} (mellom 3.16 og 4.12 består dendrogrammet av to clustre).

Tilsvarende for average: Clusterne er {ad}, {bc} (avst. 4.12 – 5.35)

K-means: Anta vi starter med clustene

$\{a\}$ Senter $\begin{pmatrix}5\\4\end{pmatrix}$

$\{bcd\}$ Senter $\begin{pmatrix}4/3\\0\end{pmatrix}$

Da vil a, b, c, d alle ligge nærmest senteret i det cluster de selv ligger.

Altså vil K-means med de initielle clustene som ovenfor lede til den samme oppdelingen. (Feks. Avst. fra b til $\{a\}$ er $\sqrt{(1-5)^2+(-1-4)^2} = \sqrt{41}$, mens avst. fra b til $\{bcd\}$ er $\sqrt{(1-4/3)^2+(-1)^2} = \sqrt{\frac{10}{9}} < \sqrt{41}$.)

Anta så vi starter med

$\{ad\}$ Senter $\begin{pmatrix}9/2\\2\end{pmatrix}$

$\{bc\}$ Senter $\begin{pmatrix}0\\0\end{pmatrix}$

Også da vil a, b, c, d alle ligge nærmest senteret i det cluster de selv ligger.